Assessing and reducing uncertainty in large-eddy simulation for microscale atmospheric dispersion

Eliott Lumet

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue le 12/01/2024 par :

Eliott Lumet

Évaluation et réduction des incertitudes pour la simulation numérique de la dispersion atmosphérique à micro-échelle

---

JURY

Bertrand CARISSIMO	CEREA, ENPC	Rapporteur
Ronan VICQUELIN	EM2C, CentraleSupélec	Rapporteur
Clémentine PRIEUR	Université Grenoble Alpes	Examinatrice
Lionel SOULHAC	LMFA, INSA Lyon	Examinateur
Ronan FABLET	Lab-STICC, IMT Atlantique	Examinateur
Céline MARI	LAERO, Toulouse	Présidente du jury
Mélanie ROCHOUX	CECI, Toulouse	Directrice de thèse

École doctorale et spécialité :
SDU2E : Océan, Atmosphère, Climat

Unité de recherche :
CECI : Climat, Environnement, Couplages et Incertitudes, CNRS-CERFACS

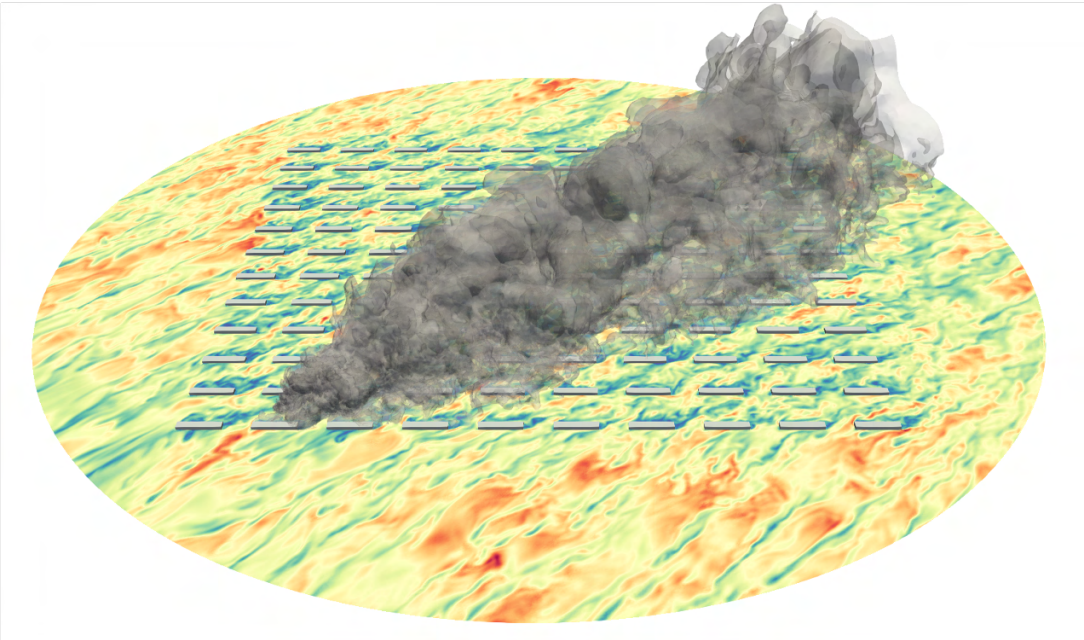
Directeurs de thèse :
Mélanie ROCHOUX, *Simon LACROIX* et *Thomas JARAVEL*

Rapporteurs :
Bertrand CARISSIMO et *Ronan VICQUELIN*

Évaluation et réduction des incertitudes pour la simulation numérique de la dispersion atmosphérique à micro-échelle

Eliott Lumet
(eliott.lumet@gmail.com)

28 Février 2024



Simulation aux grandes échelles de tourbillons de l'écoulement atmosphérique et de la dispersion d'un polluant dans une canopée urbaine simplifiée.

L'homme libre se dicte sa propre loi, inspirée par sa raison intacte, s'interdit le moindre poison. Il agit sur ses semblables par une puissance indestructible : l'exemple.

Sébastien Faure

Titre : Évaluation et réduction des incertitudes pour la simulation numérique de la dispersion atmosphérique à micro-échelle.

Mots-clés : Assimilation de données – Dispersion atmosphérique – Météorologie micro-échelle – Simulation aux grandes échelles de tourbillons – Quantification d'incertitudes – Réduction de modèle – Variabilité interne.

Résumé de vulgarisation : La qualité de l'air est fortement dégradée lors d'évènements comme les accidents industriels au cours desquels des gaz et des particules néfastes sont libérées dans l'atmosphère et transportées sous l'effet du vent. En milieu urbain, les bâtiments ont un effet de blocage sur l'écoulement ce qui peut entraîner des pics de pollution et donc des risques à court terme pour la santé et l'environnement. Localiser ces pics nécessite de recourir à des modèles résolvant les équations fondamentales de la physique des écoulements et leurs interactions avec le milieu bâti. Malgré leur complexité, ces modèles présentent des incertitudes notamment liées aux conditions atmosphériques. Cette thèse vise à construire et valider un système de modélisation permettant d'estimer ces incertitudes et d'identifier les scénarios possibles de dispersion, en s'appuyant sur des outils venant de l'apprentissage statistique et en informant le modèle à partir d'observations in-situ.

Title: Assessing and reducing uncertainty in large-eddy simulation for microscale atmospheric dispersion.

Keywords: Data assimilation – Atmospheric dispersion – Microscale meteorology – Large-eddy simulation – Uncertainty quantification – Model reduction – Internal variability.

Plain language summary: Air quality is severely degraded during events such as industrial accidents. Harmful gases and particles are released into the atmosphere and carried by the wind. In built environments, these pollutants can lead to local pollution peaks due to buildings blocking the flow, resulting in short-term health and environmental risks. Locating these peaks requires the use of models solving the fundamental equations of fluid dynamics and their interactions with the built environment. Despite their complexity, these models are subject to uncertainties that are partly linked to atmospheric conditions. The aim of this thesis is to build and validate a modeling system able of estimating these uncertainties and identifying possible dispersion scenarios. This is achieved by using tools derived from statistical learning and by informing the model with in-situ observations.

Remerciements

Parfois, pour savoir si une décision est la bonne, il n'y a d'autres choix que de se lancer. Ma thèse achevée, je sais maintenant que c'était la voie à emprunter car jamais je ne me suis autant épanoui qu'au cours de ces trois dernières années. Et pour cela j'aimerais remercier du fond du cœur tous les acteurs de cette épopée, grâce à eux je sais qu'elle ne fait que commencer.

Avant de choisir un sujet, j'ai choisi un laboratoire, le CERFACS, car toutes les thématiques de recherche qui m'avaient intéressé jusqu'alors s'y retrouvaient comme une évidence : la propulsion, la combustion, le calcul, les incertitudes, le climat et l'environnement. Grâce à un projet de thèse hautement pluridisciplinaire et à la liberté qu'on m'a offerte pour le réaliser, j'ai eu la chance de pouvoir bénéficier de cette synergie incroyable qui existe au CERFACS, à travers l'aide de nombreux experts d'horizons très différents. Merci à Olivier, Omar, Florent, Quentin, Alexis et Jérôme D. pour leur précieuse aide avec AVBP et l'intérêt qu'ils ont porté à mon application peu orthodoxe. Merci à Antoine, Gabriel et Luís de COOP pour les gains substantiels d'efficacité de mes codes. Merci aussi à Mayheul, Olivier, Paul, Selime et Antony d'ALGO pour toutes les discussions méthodes et assimilation de données que nous avons pu partager, et pour votre curiosité à l'égar de mes travaux. Je souhaiterais également remercier Nicolas, Patrick, Fabrice, Gérard, Yohann, Fred et Isabelle de l'équipe CSG pour leur support sans faille, leur bonne humeur et tout ce qu'ils m'ont appris, ainsi que Michèle, Chantal, Nathalie, Brigitte, Isabelle et Nadège qui m'ont permis de réaliser ma thèse dans les meilleures conditions possibles. J'espère avoir pu rendre un peu de ce que le Cerfacs m'a apporté à travers mon investissement dans le groupe carbone dont j'aimerais également saluer l'investissement de chacun des membres.

Au sein de ce riche environnement, j'ai eu la chance de rejoindre l'équipe GLOBC où j'ai apprécié l'atmosphère de travail à la fois performante, saine et agréable, et où la grande diversité des thématiques – en forte adéquation avec les enjeux de notre époque – m'a profondément stimulé. Pendant trois ans, je me suis levé chaque matin avec le sourire aux lèvres en sachant que j'allais passer une bonne journée aux côtés des collègues de l'équipe, et pour cela j'aimerais sincèrement tous les remercier. Un grand merci à Laurent et Julien pour m'avoir fait grandir en tant que chercheur. Malgré le fossé d'échelles qui sépare nos applications, leurs recherches auront inspiré une des contributions les plus fructueuses de ma thèse, illustrant ainsi l'importance de la curiosité et le potentiel de synergie de notre équipe. Merci à Olivier de m'avoir donné cette opportunité de donner des cours à la prépa. Merci à Laure pour sa prévenance et son grand dévouement. Merci aussi à mes fidèles camarades Gabriel, Éric, Marie-Pierre et Philippe. Merci à l'ensemble des

chercheurs et ingénieurs de l'équipe dont Emanuele, Émilia, Sophie R., Sophie V., Boris, Rym, Christian et Margot, avec qui j'ai apprécié pouvoir discuter dans la bienveillance et sans distinction de statut. Je remercie également tous les doctorants et doctorantes que j'ai côtoyées : Svenya, Siham, Mohammad, Saloua, Victoria, et Aurélien qui m'ont montré la voie; Théo D. et Émilio avec qui on s'est serré les coudes; William, Susanne, Suzanne, Romain, François, S, Camille et Théo G. qui ont apporté un vent nouveau dans l'équipe et que j'espère avoir pu aider à mon tour. J'aimerais aussi remercier plus spécialement Bastien pour notre fructueuse collaboration, les puissants outils qu'il a apporté à ma thèse, ainsi que son regard avisé de statisiticien. Merci aussi à mon fidèle voisin de bureau, Mohamed, d'avoir supporter mon caractère maussade et distant ainsi que le vacarme de mon clavier. Merci enfin à toutes les personnes avec qui j'ai eu la chance de partager un bout de leur passage dans l'équipe : Lara, Audrey, Abel, Sulian, Thanh Huy, Maxime, Rémy, Malak et Anne, vous avez ensolleillé chacune de mes pauses déjeuner et j'espère ne pas vous avoir trop bassiné avec mes histoire de course, de voyage à vélo et de tableurs.

Je n'oublie pas mon second laboratoire, le LAAS, et je remercie toutes les personnes de l'équipe RIS qui ont contribué à élargir ma vision de ce qu'est la recherche. Un grand merci à Rafael pour son grand investissement au début de ma thèse et à Lou pour l'intérêt porté à mes travaux.

J'aimerais également remercier Tim Nagel du CNRM pour son aide sur le cas MUST, ainsi que Robert Schoetter et Lionel Soulhac pour leur précieuse participation à mes comités de suivi de thèse. Je tiens aussi à remercier mes rapporteurs Bertrand Carissimo et Ronan Vicquelin d'avoir accepté de relire mon volumineux manuscrit puis de m'avoir formulé des critiques enrichissantes et constructives. Finalement, je souhaiterais remercier l'ensemble des membres de mon jury de thèse pour l'intérêt porté à mes travaux et la discussion passionnante que nous avons pu partager au cours de ma soutenance.

Car on ne se construit pas seul, je remercie aussi toute ma famille et plus particulièrement mes parents. Merci à ma mère, ma soeur et mon cousin d'avoir rendu mes angoisses un peu moins lourdes à porter tout en devant supporter mon irascibilité. Un grand merci aussi à tous mes camarades d'athlé du SATUC, Shadoks et jardiniers, qui forment pour moi une belle et grande famille ici à Toulouse. C'est aussi l'occasion de remonter le temps et de saluer tous les enseignants et camarades qui m'ont fait grandir tout au long de ma longue scolarité. Depuis l'école Louis Buton où Benoît, Linda et Joël m'ont transmis les qualités qui ont été les plus utiles à la réalisation de cette thèse : la curiosité, l'esprit critique, l'autonomie, l'auto-discipline, jusqu'à l'École Centrale où j'ai compris qu'il fallait sortir du chemin tracé pour trouver se propre voie.

Finalement, j'aimerais conclure ces remerciements par les plus importants: ceux adressés à mes encadrants de thèse.

Thomas, bien que tu te fasses souvent discret, j'ai très vite compris qu'il fallait considérer tes brillantes idées avec le plus grand intérêt et je ne compte plus le nombre d'entre elles qui ont illuminé mes travaux. J'ai particulièrement apprécié travailler de pair avec toi sur le code et les simulations, et je reste toujours ébahi par ton efficacité hors-pair et ton engagement. J'aimerais aussi te remercier pour la pédagogie dont tu as pu faire preuve pour combler mes lacunes en mécanique des fluides, m'évitant ainsi beaucoup de

déconvenues.

Simon, je souhaiterais te remercier pour ta bonne humeur, ta curiosité et la grande dévotion dont tu as fait preuve malgré l'éloignement de nos thématiques de recherche. Je salue ton ingéniosité à trouver toutes ces questions prétenduement naïves qui m'ont été d'une grande aide pour prendre du recul, remettre en question, et, finalement, maîtriser mon sujet. Enfin, tu as toujours vu en moi un futur chercheur et tout fait pour m'aider à la devenir, cette confiance c'est un roc auquel s'accrocher lors des tempêtes de questionnements que l'on doit surmonter au cours d'une thèse.

Merci *Mélanie* de m'avoir transmis ton appétence et un peu de ton expertise en assimilation de données et en méthodes ensemblistes. Tes grandes qualités d'anticipation et d'organisation m'ont permis de traverser cette thèse sereinement, y compris le sprint-marathon final qu'est la rédaction. Je te remercie du fond de cœur d'avoir toujours cherché à m'offrir les meilleures opportunités afin de pouvoir rayonner et réaliser la meilleure thèse possible. Tes copieuses corrections et ta grande disponibilité, témoignent de toute la considération que tu portes à tes étudiants et nous aident à donner le meilleur de nous-même. Enfin, je voudrais louer ta grande qualité d'écoute et te remercier pour tous ces moments d'échanges et de complicité que nous avons partagés.

Au-delà de votre expertise scientifique de très haut niveau, vous m'avez tous trois donné un magnifique exemple de ce que devait être l'encadrement: guider et non pas diriger. En effet, vous m'avez offert une grande autonomie dans mon travail, et m'avez apporté toutes les clés et les conseils nécessaires à ma réussite. Vous avez ouvert ma curiosité à de nouveaux horizons tout en m'encourageant à explorer mes propres pistes de recherche. Vous m'avez aussi démontré au jour le jour l'importance de la bienveillance, une qualité dont je mésestimais jusqu'alors la vertue. À vos côtés, j'ai travaillé d'égal à égal et j'ai compris tout ce qui faisait la beauté du métier de chercheur et j'espère, un jour, être capable de reproduire votre exemple. Enfin, merci de m'avoir donné cette opportunité de poursuivre encore un peu mon chemin à vos côtés avec ce post-doctorat qui sera, j'en suis convaincu, le meilleur des tremplins vers de nouvelles aventures scientifiques.

Merci encore et bonne lecture,
Eliott, le 28/02/2024.

Résumé

La dispersion des polluants à micro-échelle est une problématique fondamentale dans l'évaluation de la qualité de l'air, avec des implications importantes pour la santé publique. À cette échelle, de l'ordre de la centaine de mètres, l'objectif est de caractériser finement les niveaux de concentration en polluants, ce qui est particulièrement pertinent dans les environnements urbains du fait de leur grande hétérogénéité. La conception de modèles précis de dispersion à micro-échelle est d'une importance capitale pour prévoir l'exposition à la pollution atmosphérique et évaluer les risques associés, par exemple en cas d'accidents industriels. Toutefois, il s'agit d'une tâche difficile car la structure et la trajectoire des panaches de polluants sont fortement influencées par l'écoulement atmosphérique, qui est intrinsèquement multi-échelle et turbulent, et qui interagit de façon complexe avec l'environnement bâti.

La mécanique des fluides numérique (CFD) s'est imposée comme un outil puissant pour résoudre ce problème en fournissant des simulations d'écoulement et de dispersion qui prennent explicitement en compte les bâtiments. Cependant, les modèles CFD sont très couteux, ce qui entrave leur utilisation dans des contextes opérationnels et en situation d'urgence. Par ailleurs, leur précision reste limitée en raison des fortes incertitudes en jeu, en particulier celles liées au forçage atmosphérique à grande échelle et à la variabilité interne de la couche limite atmosphérique. Pour les applications d'évaluation des risques, il est essentiel de contrôler et de quantifier ces incertitudes, mais cela est rendu difficile par le coût des modèles CFD.

Pour répondre à ces deux problèmes, nous mettons en œuvre et validons un système de modélisation par simulation des grandes échelles de tourbillons (LES), qui inclut un modèle réduit et un algorithme d'assimilation de données basé sur un filtre de Kalman d'ensemble. Cette thèse établit une preuve de concept de la capacité du système à améliorer les prévisions LES de concentration de polluant dans le cas d'un essai de la campagne expérimentale MUST. En particulier, nous démontrons que les mesures locales de concentration peuvent être utilisées pour réduire les incertitudes paramétriques météorologiques et corriger les biais dans les conditions aux limites du modèle. L'utilisation d'un modèle réduit permet de générer des prévisions d'ensemble qui tiennent compte avec précision des fortes non-linéarités du modèle LES, en quelques dizaines de secondes seulement.

Une attention particulière est accordée à l'incertitude associée à la variabilité interne de la couche limite atmosphérique, que nous proposons de quantifier à l'aide d'une approche bootstrap. Nous démontrons que la variabilité interne impacte fortement les prévisions de dispersion à micro-échelle des modèles LES mais aussi les observations sur

le terrain, et qu'elle ne doit donc pas être ignorée lors de l'évaluation des modèles de dispersion. Nous allons ensuite plus loin en tenant compte de cette variabilité interne dans la construction du modèle réduit et du système d'assimilation de données. En particulier, nous montrons que l'analyse de la variabilité interne permet de choisir le nombre de modes du modèle réduit afin d'éviter d'introduire du bruit dans ses prévisions. Enfin, nous prenons en compte la variabilité interne à micro-échelle dans le processus d'assimilation de données afin de le rendre plus robuste et réaliste. Ces développements sont faits de manière élégante et sans générer de lourdeur d'implémentation ou de calcul.

D'un point de vue plus général, cette thèse montre quelques pistes pour adopter une approche de modélisation probabiliste de processus atmosphériques complexes basée sur la modélisation LES, qui est aujourd'hui reconnue comme une référence mais qui reste soumise à des incertitudes, dont certaines sont intrinsèquement irréductibles.

Abstract

Microscale pollutant dispersion is a critical aspect of air quality assessment with significant implications for the environment and public health. At this scale, of the order of a hundred meters, the aim is to characterize pollutant concentration levels in detail, which is particularly relevant in urban environments due to their great heterogeneity. Designing accurate microscale dispersion models is of paramount importance for predicting air pollution exposure and assessing risks, for example in case of industrial accidents. However, this is a challenging task, as the structure and trajectory of pollutant plumes are strongly influenced by atmospheric flow, which is inherently multi-scale and turbulent and interacts in complex ways with the built environment.

Computational Fluid Dynamics (CFD) has emerged as a powerful tool to address this issue by providing obstacle-resolving flow and dispersion predictions. However, CFD models are very costly, which hinders their use in operational and emergency contexts. In addition, their accuracy remains limited because of the significant uncertainties involved, in particular those arising from a lack of knowledge about the large-scale atmospheric forcing and from the internal variability of the atmospheric boundary layer. For risk assessment applications, controlling and quantifying these uncertainties is essential, but made difficult by the cost of CFD models.

To address these dual issues, we design and validate a large-eddy simulation (LES) modeling system that includes a reduced-order model and a data assimilation algorithm based on an ensemble Kalman filter. This thesis provides a proof-of-concept of the system's ability to improve LES pollutant concentration field predictions in a neutral trial of the MUST field experiment. In particular, we demonstrate that local pollutant concentration measurements can be used to reduce meteorological parametric uncertainties and correct bias in the model boundary conditions. The use of a reduced-order model enables generating ensemble predictions that accurately account for the strong nonlinearities of the LES model, in just a few tens of seconds.

Particular attention is paid to the uncertainty associated with the internal variability of the atmospheric boundary layer. We adapt a bootstrap approach to quantify its effect on microscale dispersion and demonstrate that internal variability significantly affects not only LES model predictions but also field observations. By propagating the associated uncertainties to the standard statistical metrics used for air quality model evaluation, we show that the resulting variability in the validation metrics is significant and cannot be ignored when evaluating LES model accuracy. We then go a step further by accounting for this internal variability in the construction of the reduced-order model and of the data assimilation system. In particular, we show that the analysis of internal

variability is of great interest to make an informed choice on the number of reduced-basis modes to avoid introducing noise into the reduced-order model. Finally, we take into account the microscale internal variability in the data assimilation process, making it much more robust and realistic. These additions to the data assimilation framework are made elegantly and without generating implementation or computational heaviness.

From a broader point of view, this thesis shows some ways to adopt a probabilistic modeling approach for complex atmospheric phenomena based on LES, which are nowadays recognized as references, but remain subject to uncertainties, some of which are inherently irreducible.

Contents

Remerciements **v**

Résumé **ix**

Abstract **xi**

Contents **xiii**

Introduction générale **1**

General introduction **5**

I Scientific context and approach **9**
I.1 Air pollutant dispersion in urban environment 11
I.2 The limits of computational fluid dynamics dispersion models 26
I.3 Thesis objectives and approach 32

II Microscale atmospheric dispersion large-eddy simulation **45**
II.1 Introduction 47
II.2 Large-eddy simulation dispersion modeling 50
II.3 The Mock Urban Setting Test field campaign 57
II.4 Large-eddy simulation model of the MUST trial 2681829 61
II.5 Preliminary verification: simulation of a free-field case 70
II.6 Summary 74

III Robust model validation under uncertainty **75**
III.1 Introduction 77
III.2 Internal microscale variability quantification 80
III.3 Model validation methodology 89
III.4 LES model validation and microscale internal variability quantification . 92
III.5 LES model sensitivity to the main sources of uncertainty 108
III.6 Conclusion 115

IV Reduced-order modeling based on LES statistical predictions **117**
IV.1 Introduction 119
IV.2 Reduced-order modeling approach 122

IV.3 Reduced-order model validation methodology . . . 133
IV.4 LES ensemble generation . . . 137
IV.5 Setting up the POD–GPRs model . . . 144
IV.6 Validation of the POD–GPRs model . . . 155
IV.7 Conclusion . . . 167

V Data assimilation for wind condition estimation 169
V.1 Introduction . . . 171
V.2 Data assimilation theoretical framework . . . 178
V.3 Application to the MUST trial 2681829 . . . 184
V.4 Validation and calibration of the data assimilation system . . . 193
V.5 Assimilation of the real field measurements . . . 200
V.6 Conclusion . . . 207

Conclusion and perspectives 209

Conclusion et perspectives 219

A Additional sensitivity tests of the LES model 231
A.1 LES model mesh convergence . . . 231
A.2 LES model sensitivity to computational domain height . . . 233
A.3 Impact of adding turbulence injection on LES predictions . . . 235

B Reduced-order model additional applications 239
B.1 Prediction of other fields . . . 239
B.2 Application to global sensitivity analysis . . . 242
B.3 Towards reduced-order modeling based on a mixture of experts . . . 247

C Carbon footprint estimation 251

Bibliography 261

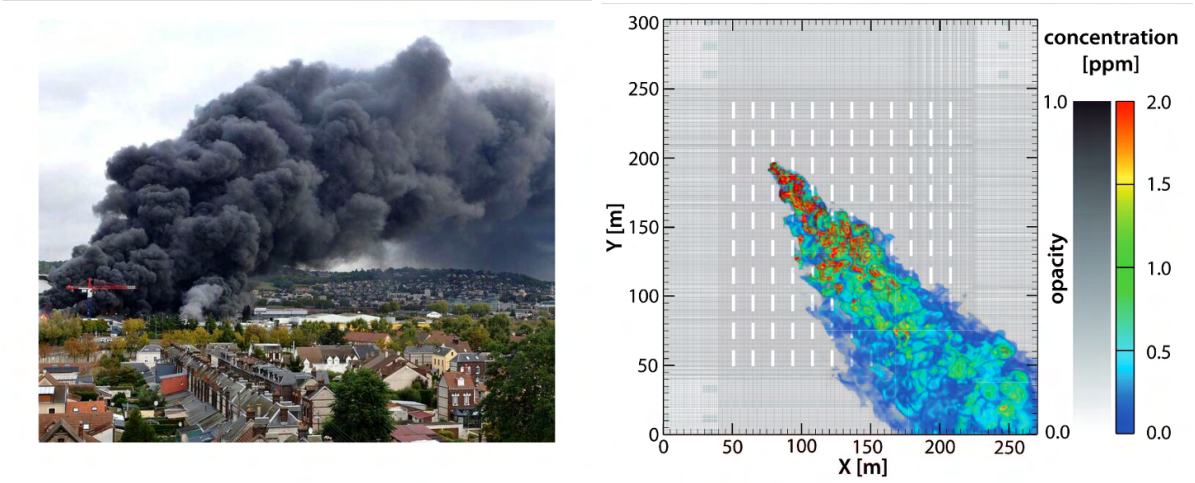
Introduction générale

Contexte et enjeux

La dispersion atmosphérique fait référence à la manière dont divers polluants, tels que des gaz, des particules et des composés chimiques, sont transportés et dispersés dans l'atmosphère terrestre. Un polluant est défini comme une substance qui contamine l'environnement et qui a des effets néfastes sur les écosystèmes et la santé humaine. Les polluants atmosphériques proviennent soit de phénomènes naturels tels que les éruptions volcaniques et les incendies de forêt, soit d'émissions anthropiques quotidiennes liées aux activités industrielles, à l'agriculture et aux transports. Les polluants peuvent également être dispersés dans l'atmosphère lors d'accidents industriels. Les accidents nucléaires de Tchernobyl et Fukushima en 1986 et 2011, l'explosion de l'usine AZF en 2001 à Toulouse et l'incendie de l'usine Lubrizol en 2019 à Rouen en sont la preuve (Fig. 1a). Depuis les années 2000, l'inquiétude croissante face au terrorisme a également donné un nouvel élan à l'étude de la dispersion des polluants dans les zones urbaines densément peuplées (Allwine et al. 2002, 2004; Fox and Storwold 2011; Fox et al. 2022).

La dégradation de la qualité de l'air associée à la pollution atmosphérique a des impacts significatifs sur la santé humaine (EEA 2020). À long terme, la contamination des environnements extérieurs et intérieurs est à l'origine de maladies respiratoires et cardiovasculaires, de cancers, d'allergies et d'asthme (Manisalidis et al. 2020). La pollution atmosphérique est ainsi responsable de millions de décès chaque année, ce qui en fait la troisième cause de mortalité dans le monde (Ritchie and Roser 2017). L'exposition à de fortes concentrations de polluants peut également avoir des effets très graves sur la santé à court terme, en particulier dans le cas d'accidents industriels. Par exemple, lors de la catastrophe de Bhopal en Inde en 1984, une fuite d'isocyanate de méthyle, un pesticide hautement toxique, a causé la mort d'environ 16 000 personnes (Eckerman 2005). Il est important de noter que les populations les plus pauvres sont souvent les plus exposées (Ritchie and Roser 2017).

La pollution atmosphérique a également des effets importants sur l'environnement et le changement climatique. La contamination de l'atmosphère peut se propager à l'eau, au sol et à la végétation, ce qui nuit à l'agriculture et menace la biodiversité (Rai et al. 2011). En outre, certains polluants atmosphériques, tels que le dioxyde de carbone et le méthane, sont des gaz à effet de serre responsables du changement climatique qui a des conséquences considérables et potentiellement catastrophiques pour la société humaine. Le changement climatique contribue à son tour à la détérioration de la qualité de l'air en augmentant l'ozone troposphérique, les allergènes en suspension dans l'air et le nombre



(a) **(b)**

Figure 1: *Exemples de dispersion de polluants en milieu urbain. (a) Photo de l'incendie de l'usine Lubrizol® en 2019 à Rouen, France. Crédit photo : Jean Pierre Mauger. (b) Concentration instantanée prédite par une simulation des grandes échelles de tourbillons (LES) de König (2014) reproduisant un essai de la campagne de terrain MUST (Biltoft 2001) dans le désert de l'Utah en 2001. Au cours de cette campagne, du propène a été relâché au milieu d'un ensemble de conteneurs imitant une canopée urbaine simplifiée.*

de feux de forêt (Reidmiller et al. 2017).

Le suivi de la concentration des polluants est un problème multi-échelle allant du champ proche (à quelques mètres de la source) au champ lointain (à des échelles allant de quelques centaines de mètres jusqu'à l'échelle globale). Dans cette thèse, nous examinons la question de la dispersion atmosphérique en champ proche, en mettant l'accent sur la dispersion de polluants gazeux en environnement urbain. Ce sont des environnements à haut risque en raison de leur forte densité de population et parce que la topographie urbaine ralentit la dispersion atmosphérique et induit une forte variabilité spatio-temporelle de la concentration des polluants, ce qui rend difficile la localisation des pics de pollution. La nature intrinsèquement multi-échelle des écoulements urbains, dirigés par les conditions météorologiques à grande échelle et influencés localement par l'environnement bâti, représente un défi important lorsqu'il s'agit de comprendre et de prévoir la dispersion de polluants en ville. Ainsi, le développement de modèles de dispersion à micro-échelle capables de prédire la distribution spatiale des polluants est d'une importance capitale pour l'évaluation des risques sanitaires pour les habitants. Ces modèles peuvent également être utilisés de manière préventive pour définir des réseaux de surveillance de la qualité de l'air et pour aider les urbanistes et les architectes à concevoir l'espace urbain de manière à minimiser l'exposition aux polluants. Dans le contexte spécifique d'un rejet accidentel de polluants, les modèles à micro-échelle sont nécessaires pour aider la sécurité civile à prendre des mesures d'urgence efficaces et, par la suite, pour déterminer les cartes d'exposition.

Positionnement de la thèse

Pour étudier les processus atmosphériques à micro-échelle, tels que la dispersion, en particulier en environnement urbain, la communauté scientifique s'accorde de plus en plus sur la nécessité d'utiliser des modèles de mécanique des fluides numérique (CFD) (Holmes and Morawska 2006; Blocken et al. 2013). Ces modèles haute fidélité résolvent les équations complexes de Navier-Stokes à l'aide de différentes approches : par exemple en les moyennant (RANS), ou bien en les filtrant (LES). La grande force des modèles CFD est leur capacité à représenter les détails des processus complexes et multi-échelle qui régissent les écoulements et la dispersion atmosphériques. En particulier, ces modèles permettent de prendre en compte de manière explicite les interactions complexes entre l'environnement bâti et la dispersion des polluants.

Néanmoins, il convient de noter que les modèles CFD sont très coûteux et requièrent donc des moyens de calcul avancés. De plus, il est reconnu que la précision des modèles CFD est fortement limitée par leurs incertitudes, et doit être encore améliorée (Blocken 2014; Dauxois et al. 2021). Ces incertitudes découlent en partie des hypothèses de modélisation ou d'un manque de connaissance sur les conditions météorologiques. Cette double limitation en coût et précision empêche la généralisation de l'utilisation des modèles CFD pour l'étude de la dispersion atmosphérique à micro-échelle.

Parmi les différentes sources d'incertitude affectant les modèles CFD, nous nous intéressons particulièrement aux incertitudes sur les conditions météorologiques qui affectent fortement les prédictions de dispersion atmosphérique. Une partie de ces incertitudes sont dues à un manque de connaissance de l'état de l'atmosphère à un moment donné, par exemple en raison de capacités d'observation limitées. Une autre partie de ces incertitudes sont liées à la variabilité interne de la couche limite atmosphérique (CLA). En effet, lorsque l'on considère des moyennes sur des horizons temporels finis, cette variabilité induit une incertitude irréductible tant sur les prévisions des modèles que sur les observations. Des études signalent que la variabilité interne est l'une des principales raisons des écarts entre les expériences sur le terrain et les simulations CFD ou les expériences en soufflerie et expriment la nécessité d'aller au-delà de la comparaison déterministe entre les modèles et les observations (Schatzmann and Leitl 2011; Antonioni et al. 2012; García-Sanchez et al. 2018; Dauxois et al. 2021). L'estimation de la variabilité interne constituerait donc une avancée méthodologique majeure pour la validation robuste des modèles CFD atmosphériques lorsque les données sont acquises sur des périodes limitées, mais aussi pour l'analyse de sensibilité des modèles et les comparaisons multi-modèle.

Actuellement, il y a une prise de conscience croissante du potentiel de la modélisation orientée données pour surmonter les limites des modèles CFD atmosphériques à micro-échelle. En particulier, les techniques de réduction de modèle basées sur l'apprentissage statistique fournissent une solution pour réduire le coût des modèles CFD, sans compromettre la précision (Margheri and Sagaut 2016; Nony et al. 2023a; Pasquier et al. 2023). Parallèlement, l'assimilation de données, un domaine de recherche initié en météorologie et en océanographie, commence à être appliquée à la micro-échelle pour réduire l'incertitude sur les conditions aux limites météorologiques des modèles CFD à l'aide de données d'observation, ce qui permet d'améliorer considérablement la précision des

prévisions (Mons et al. 2017; Sousa et al. 2018; Sousa and Gorlé 2019; Defforge et al. 2021).

Cette thèse vise à explorer les horizons prometteurs offerts par les techniques de réduction de modèle et d'assimilation de données dans la quête d'une modélisation plus efficace et plus précise de la dispersion atmosphérique à micro-échelle basée sur la CFD. La question spécifique de la prise en compte de l'incertitude irréductible liée à la variabilité interne de la CLA dans ces techniques est au cœur du travail présenté dans cette thèse.

General introduction

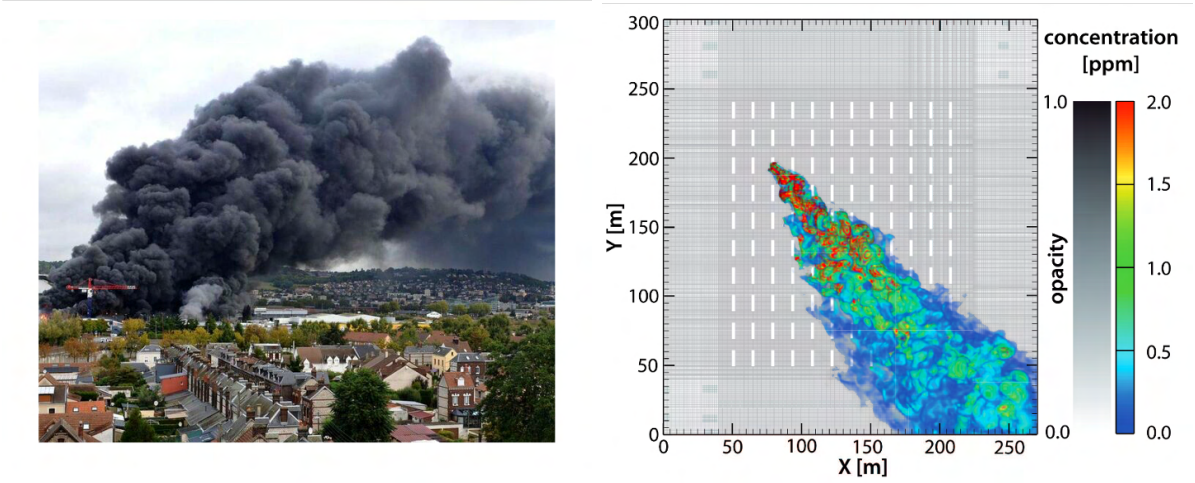
Context and issues

Atmospheric pollutant dispersion refers to the way various pollutants, such as gases, particulate matter, and chemical compounds, are transported and dispersed throughout the Earth's atmosphere. A pollutant is defined as a substance that contaminates the environment, causing adverse effects on ecosystems and human health. Air pollutants either originate from natural hazards such as volcanic eruptions and wildfires, or daily anthropogenic emissions related to industrial activities, agriculture, and transportation. Pollutants can also be dispersed into the atmosphere during industrial accidents. This was evidenced by the Chernobyl and Fukushima nuclear accidents in 1986 and 2011, the AZF factory explosion in 2001 in Toulouse, and the Lubrizol factory fire in 2019 in Rouen (Fig. 2a). Since the 2000s, growing concern about terrorism has also given new impetus to the study of pollutant dispersion in densely populated urban areas (Allwine et al. 2002, 2004; Fox and Storwold 2011; Fox et al. 2022).

Air pollutants can degrade air quality and have significant impacts on human health (EEA 2020). In the long term, the contamination of both outdoor and indoor environments is the cause of respiratory and cardiovascular diseases, cancers, allergies and asthma (Manisalidis et al. 2020) and is attributed to millions of deaths each year, making it the third largest cause of death worldwide (Ritchie and Roser 2017). Exposure to high concentrations of pollutants can also have very serious short-term health effects, particularly in the case of industrial accidents. For example, during the Bhopal disaster in India in 1984, a leak of methyl isocyanate, a highly toxic pesticide, caused the fatalities of approximately 16,000 people (Eckerman 2005). It is important to note that the poorest populations are often the most exposed (Ritchie and Roser 2017).

Air pollution also has far-reaching effects on both the environment and climate change. Contamination of the atmosphere can spread to water, soil and vegetation, adversely affecting agriculture and threatening biodiversity (Rai et al. 2011). Moreover, some air pollutants, such as carbon dioxide and methane, are greenhouse gases responsible for climate change leading to global warming and climate instability. Climate change in turn contributes to worsening air quality by increasing ground-level ozone, airborne allergens, and the number of wildfires (Reidmiller et al. 2017).

Tracking the pollutant concentration is a multi-scale problem ranging from the near field (within a few meters away from the source) to the far field (at scales ranging from a few hundred meters up to the global scale). In this thesis, we examine the question of atmospheric dispersion in the near field, with a particular focus on the dispersion of



(a) **(b)**

Figure 2: *Examples of pollutant dispersion in urban environments. (a) Picture of the Lubrizol® factory fire in 2019 in Rouen, France. Photo credit: Jean Pierre Mauger. (b) Instantaneous concentration predicted by a large-eddy simulation from König (2014). It reproduces one trial of the MUST field campaign (Biltoft 2001) in the Utah desert in 2001, in which propylene was released within an array of shipping containers that mimic a simplified urban canopy.*

gaseous pollutants in urban environments. These are high-risk environments because of their high population density and because the urban topography slows down atmospheric dispersion and induces a strong spatio-temporal variability of pollutant concentration, making it difficult to locate pollution peaks. The inherently multiscale nature of urban flows, governed by large-scale weather patterns down to microscale turbulence generated by the complex built environment, represents a substantial challenge when it comes to understanding and predicting pollutant dispersion in cities. Thus, the development of microscale dispersion models able to predict the spatial distribution of pollutants is of paramount importance in assessing health risks for the inhabitants. These models can also be used in a preventive manner to define air quality monitoring networks and to help urban planners and architects design urban space in such a way as to minimize pollutant exposure. In the specific context of accidental pollutant release, microscale models are required to assist civil security in taking effective emergency response measures and, subsequently, to determine exposure maps.

Thesis positioning

To study microscale atmospheric processes such as dispersion, especially in urban environments, there is a growing consensus in research for the use of Computational Fluid Dynamics (CFD) models (Holmes and Morawska 2006; Blocken et al. 2013). These high-fidelity models include various approaches, such as Reynolds-Averaged Navier-Stokes (RANS) and Large-Eddy Simulation (LES), all of which solve the complex Navier-Stokes equations. The greatest strength of CFD models is their ability to represent fine details

of the intricate and multiscale processes that govern atmospheric flows and dispersion. In particular, these models make it possible to explicitly account for the complex interactions between the built environment and pollutant dispersion.

Nonetheless, the application of CFD models is not without challenges. Two main limitations are the substantial computational costs associated with running these simulations and their various uncertainties, emerging for example from modeling assumptions, or a lack of knowledge about the meteorological conditions. To this date, it is recognized that the accuracy of CFD models is severely limited by their uncertainties, and needs to be further improved (Blocken 2014; Dauxois et al. 2021). These two limitations hinder the widespread application of CFD-based microscale atmospheric dispersion studies.

Among the different sources of uncertainty affecting CFD models, we are particularly interested in the uncertainty of the meteorological conditions that directly affect dispersion processes. Part of this uncertainty is due to a lack of knowledge about the state of the atmosphere at a given time, for example, because of limited observation capabilities. Another part of this uncertainty is linked to the internal variability of the turbulent atmospheric boundary layer (ABL). When considering averages over finite time horizons, this variability induces an irreducible uncertainty on both model predictions and observations. Studies report internal variability as one of the main reasons for the discrepancies between field-scale experiments and CFD simulations or wind-tunnel experiments and express the need to go beyond deterministic point-wise model/observations comparison (Schatzmann and Leitl 2011; Antonioni et al. 2012; García-Sanchez et al. 2018; Dauxois et al. 2021). Estimating internal variability would therefore be a major methodological advance for the robust validation of atmospheric CFD models when data are acquired over limited periods, but also for model sensitivity analysis and multi-model comparisons.

Recently, there has been a growing awareness of the potential of data-driven techniques to overcome the limitations of microscale atmospheric CFD models and thus improve our understanding of atmospheric microscale processes. In particular, data-driven model reduction techniques provide a solution to mitigate the cost of CFD models, without compromising on accuracy (Margheri and Sagaut 2016; Nony et al. 2023a; Pasquier et al. 2023). Meanwhile, data assimilation, a field of research initiated in meteorology and oceanography, is beginning to be applied at the microscale to reduce the uncertainty on the meteorological boundary conditions of CFD models using observational data, resulting in significant improvements in prediction accuracy (Mons et al. 2017; Sousa et al. 2018; Sousa and Gorlé 2019; Defforge et al. 2021).

This thesis aims to explore the promising horizons offered by model reduction and data assimilation techniques in the quest for more efficient and accurate microscale atmospheric dispersion modeling based on CFD. The specific question of how to take into account the irreducible uncertainty related to the internal variability of the ABL in these techniques is at the heart of the work presented in this thesis.

Chapter I

Scientific context and approach

The general aim of this chapter is to present and position the work of this thesis in the current state of scientific knowledge.

To begin with, we present the general physics of microscale pollutant dispersion with a focus on urban environments, highlighting its complex interactions with the atmospheric boundary layer and the challenges facing the modeler. An overview of modeling techniques and the main experiments used for model validation is provided. A particular focus is given to Computational Fluid Dynamics (CFD) models which explicitly solve the wind flow and pollutant dispersion between buildings.

In the second section, we take a closer look at the limitations of CFD models in terms of computational cost and accuracy. We propose a detailed classification of the source of errors involved, separating them between aleatory and epistemic uncertainties. This review of the scientific literature shows that the internal variability of the atmospheric boundary layer plays a major role in the lack of accuracy of CFD models and has not yet been fully addressed.

Recent works have shown that model reduction and data assimilation techniques could be particularly suited to address this dual problem of computational cost and uncertainty. We introduce the reader to these techniques before reviewing applications with the potential for improving microscale dispersion CFD modeling. In light of the scientific context presented throughout the chapter, we finally introduce our reduced-cost data assimilation system to quantify and reduce uncertainty related to large-scale atmospheric forcing while also accounting for microscale internal variability.

The last part of this chapter presents the main steps to build this system which defines the structure of the manuscript.

Chapter outline

I.1 Air pollutant dispersion in urban environment 11
I.1.1 A multi-scale problem 11
I.1.2 Atmospheric flow and dispersion at microscale 15
I.1.2.1 Governing equations 15
I.1.2.2 Atmospheric turbulence and impact on dispersion . . 17
I.1.2.3 Atmospheric stability and impact on dispersion 19
I.1.2.4 Similarity theory under neutral stratification conditions 19
I.1.3 A brief overview of dispersion modeling approaches 22
I.1.4 Experiments of dispersion in urban environment 24
I.2 The limits of computational fluid dynamics dispersion models 26
I.2.1 The computational burden of CFD models 26
I.2.2 Uncertainties in microscale atmospheric CFD models 27
I.2.2.1 Epistemic uncertainties 28
I.2.2.2 Aleatory uncertainty arising from the internal variability of the ABL 30
I.3 Thesis objectives and approach 32
I.3.1 Accelerating predictions using reduced-order model 32
I.3.2 Data assimilation: harnessing observations to improve predictions 35
I.3.3 Design of an efficient data assimilation system 39
I.3.4 Implementation strategy 41
I.3.5 Structure of the manuscript 43

I.1 Air pollutant dispersion in urban environment

In this section, we introduce the main physical phenomena involved in atmospheric flows and pollutant dispersion in urban environment. First, we define the different scales at stakes in Sect. I.1.1, before introducing the governing equations and challenges arising from the turbulence and stability of the atmosphere in Sect. I.1.2. Then, we give an overview of the existing dispersion modeling techniques (Sect. I.1.3) and of the experiments commonly used as validation basis (Sect. I.1.4).

I.1.1 A multi-scale problem

Atmospheric dispersion involves a wide variety of scales, from the transport of aerosols around the globe down to the emission of pollutants by vehicles at street level. The typical scales for studying dispersion, according to Seinfeld and Pandis (1998), are defined in Table I.1. Britter and Hanna (2003) further develop this classification in the specific case of dispersion in urban environments, defining four scales: regional, city, neighborhood and street. The city scale covers the complete area of a city and approximately corresponds to the urban scale defined in Table I.1. The regional scale includes a larger zone around the urban area to account for the interaction with local meteorology. Neighborhood and street scales both correspond to the microscale defined in Table I.1; the former typically extends up to one or two kilometers, while the latter is limited to one street or few buildings.

Table I.1: *The typical spatial scales of atmospheric pollutant transport models, from Seinfeld and Pandis (1998).*

Scale	Typical length scale	Typical study domain
Global	5°	65 000 × 65 000 × 20 km
Synoptic (continental)	80 km	3 000 × 3 000 × 20 km
Regional	20 km	1 000 × 1 000 × 10 km
Urban	4 km	100 × 100 × 5 km
Microscale	5 m	200 × 200 × 100 m

Each of these scales requires a different modeling approach. For instance, at the regional scale, a macroscopic viewpoint is adopted as the effect of buildings on atmospheric flow and dispersion is evaluated statistically. In this thesis, we focus on microscale dispersion which requires explicitly representing the interactions between the built environment and atmospheric flow to accurately represent pollutant concentration levels. It is however not possible to ignore larger atmospheric scales as all these scales are deeply interconnected due to the mixing processes occurring in the atmosphere. The uncertainty resulting from this coupling between scales is at the core of the thesis problematic. In the following, we introduce in more detail the structure of the atmospheric boundary layer.

Figure I.1 shows the typical scales of motion involved in the atmosphere above urban environments, as introduced in the pioneering works of Oke (1987) on urban climatology.

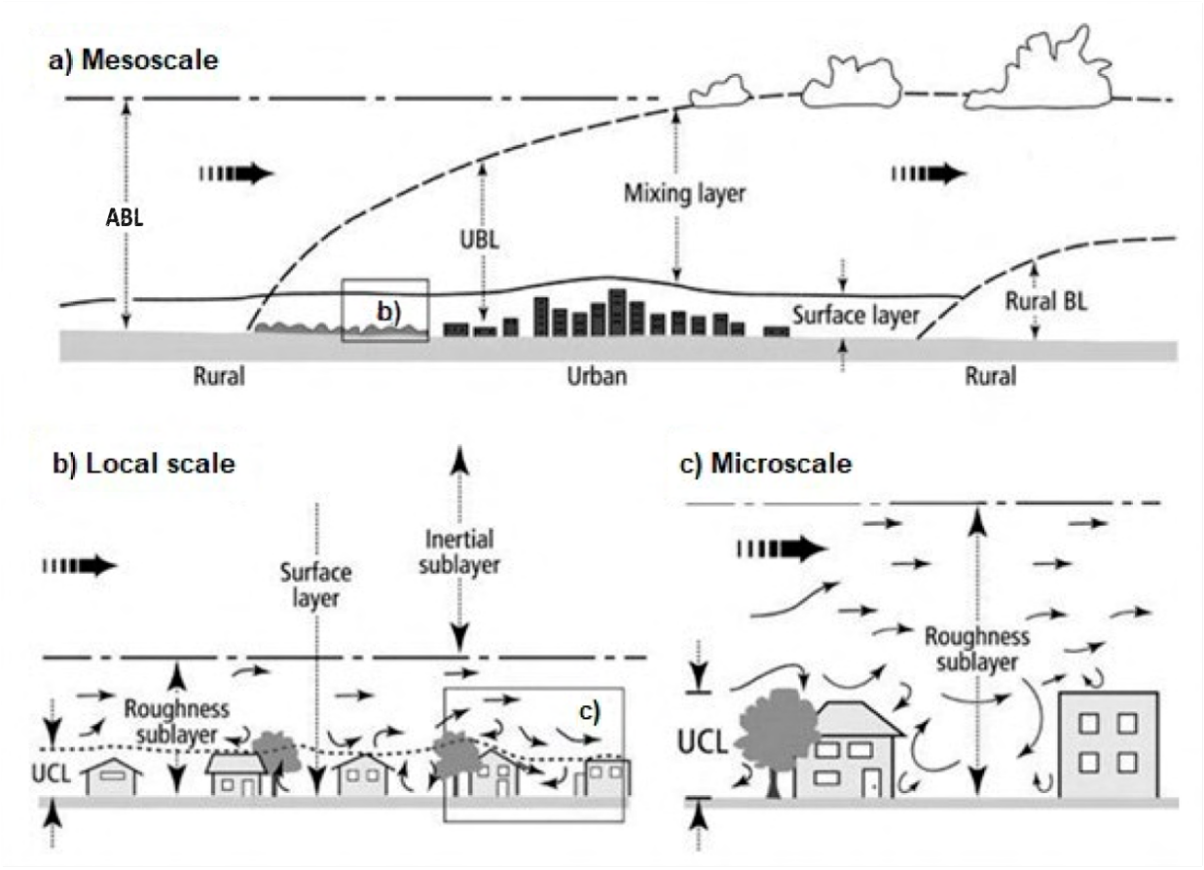


Figure I.1: *Conceptual representation of the urban atmosphere and different layers and flow scales at stakes (from Bailey et al. (1997), modified after Oke (1987)). The situation represented corresponds to an unstable daytime urban boundary layer. ABL, UBL, and UCL respectively stand for atmospheric boundary, urban boundary, and urban canopy layers.*

The urban boundary layer (UBL) is considered a special case of the atmospheric boundary layer (ABL), which corresponds to the part of the atmosphere influenced by the ground surface. It is characterized by well-developed mixing due to frictional drag induced by surface elements, and thermal convection caused by the temperature difference between air and surface.

The UBL can be divided into sublayers according to their mean and turbulent properties (Fig. I.1). A first distinction can be made between the surfacer layer, which is in direct contact with the surface and where Coriolis forces can be neglected, and the outer layer which is mostly determined by advection (Fisher et al. 2005) and features spatially homogeneous physical properties (Oke 1987). At smaller scales (Fig. I.1b), the surface layer is itself divided between:

- the roughness sublayer where the flow is directly driven by the urban canopy and therefore essentially heterogeneous and three-dimensional (Raupach et al. 1991),
- and the inertial sub-layer above, which is horizontally homogeneous on average and where the vertical flows of momentum and heat can be considered constant.

The height of the ABL, also known as layer thickness or mixing depth, varies spatially with ground topography, but also over time due to changes in the meteorological conditions at the synoptic scale and daily solar cycle. In the daytime, the ground surface is heated by the Sun, causing strong convection and increasing the height of the ABL to a few kilometers. Then, during the night, the ground surface becomes colder than the atmosphere, which reduces mixing and thus the height of the ABL to a few dozen meters. In this case, the outer layer and inertial sublayer shrink and can even disappear. A more detailed explanation of the effect of thermal stratification on atmospheric flow is given in Sect. I.1.2.3.

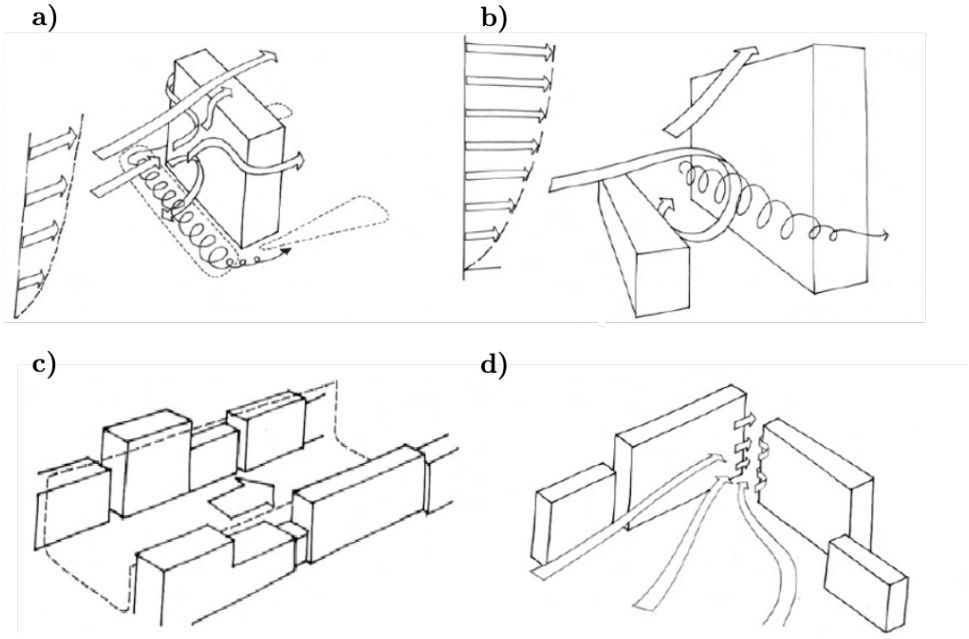


Figure I.2: *Schematic examples of different interaction regimes between buildings and atmospheric flow at microscale from Bouyer (2017). a) Wind flow around one isolated obstacle. b) Interaction between two obstacles of different heights. c) Channeling effect between aligned buildings. d) Local acceleration of the flow due to building layout.*

In this thesis, we are mainly interested in giving a detailed description of the microscale dispersion phenomena occurring in the urban canopy layer (UCL), the lowest part of the surface layer. At this scale which extends up to several times the roof height (Fig. I.1c), flow, heat transfer and pollutant dispersion are dominated by the local characteristics of the surroundings (buildings, vegetation, street furniture, etc). For example, Figure I.2a shows a schematic view of the effect of one isolated building on the flow: it is slowed down and deflected in each direction before separating at the corners of the obstacle. At ground level a standing vortex settles, and a recirculation builds up in the wake of the building. Flow and dispersion around isolated buildings have been studied extensively in the literature (Li and Meroney 1983a,b; Meroney et al. 1999; Tominaga and Stathopoulos 2010). Multi-obstacle configurations have also been widely studied (Hanna et al. 2002; Vardoulakis et al. 2003; Blocken and Carmeliet 2004) as they lead to specific interaction

effects such as vortex formation in street canyon (Fig I.2b), channeling effect (Fig I.2c), or local bottleneck effect (Fig I.2d).

As a result, air pollutant dispersion at the microscale is directly determined by the local configuration of the urban canopy. For example, when the wind is perpendicular to a street, the resulting recirculation zone (Fig. I.2b) can trap air pollutants and become a hot spot of pollutant concentrations (Fig. I.3a). This is why the street canyon configuration is a canonical microscale dispersion case study (Pavageau and Schatzmann 1999; Vardoulakis et al. 2003). As shown in Figs. I.3a, b, pollutant removal from the street notably depends on the ratio of the distance between buildings to their typical height. In addition to building layout, the plume shape is also strongly affected by the location of the source (Figs. I.3c, d). Because of the highly heterogeneous nature of the real urban environment, these complex microscale processes can combine in virtually unlimited ways within the urban canopy layer, thus motivating the development of obstacle-resolving models to predict and understand flows for any urban geometry.

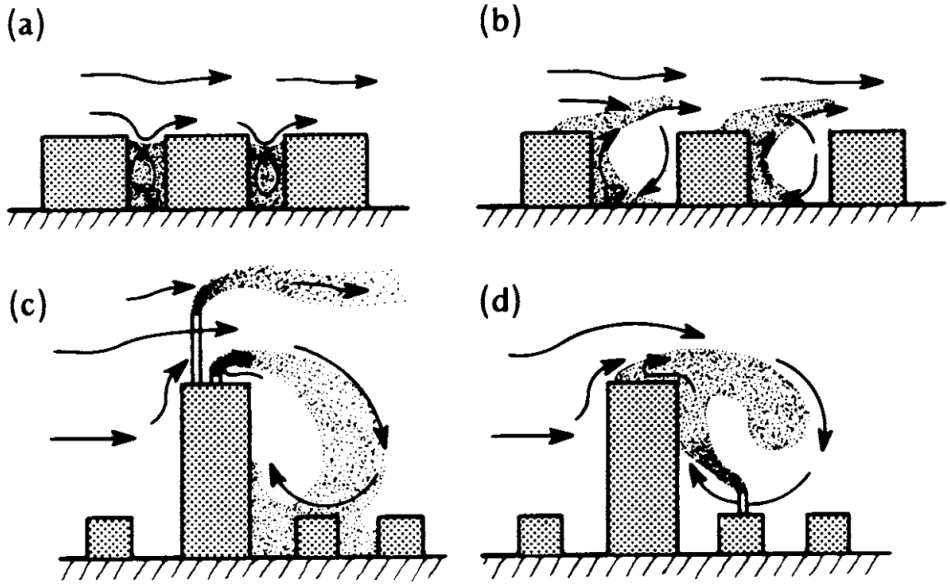


Figure I.3: *Influence of building on air pollutant dispersion for different geometries and source location, from Oke (1987).*

Finally, we emphasize that the scales introduced by Oke (1987) are strongly interconnected. On the one hand, the microscale processes taking place in the urban canopy layer influence the upper layers by slowing down the flow and increasing turbulence production. On the other hand, local flow conditions, notably wind direction and speed, are directly conditioned by regional-scale meteorology. This makes air pollutant dispersion an inherently multiscale phenomenon and represents another grand challenge for modeling. For instance, microscale studies are used to parameterize the effect of the urban environments in regional-scale models (Nazarian et al. 2020; Nagel et al. 2023), while mesoscale data are required to define boundary conditions of microscale dispersion models (García-Sanchez and Gorlé 2018). In Sect. I.2.2, we present in greater detail the uncertainties associated with this coupling of scales.

I.1.2 Atmospheric flow and dispersion at microscale

I.1.2.1 Governing equations

The dispersion of pollutants in the atmosphere is intrinsically driven by atmospheric flow, which is governed by the equations of fluid mechanics. These are the famous Navier-Stokes equations that account for the fundamental principle of mass, energy, and momentum balance. In the following, we present how these equations can be simplified at the microscale to describe the local atmospheric flow and dispersion of passive species. This concise theoretical presentation is based on Stull's book (1988) and Seinfeld and Pandis' one (1998). When writing the equations, we adopt the index notation, including Einstein's rule of summation.

Conservation of mass Adopting an Eulerian description, the law of conservation of mass applied to an elementary volume of air leads to the continuity equation:

$$\frac{D\rho}{Dt} + \rho \frac{\partial u_i}{\partial x_i} = 0, \tag{I.1}$$

where ρ and $\mathbf{u} = (u_i, u_j, u_k)$ are the fluid density and velocity vector fields, and where $\frac{D\rho}{Dt} = \frac{\partial \rho}{\partial t} + u_i \frac{\partial \rho}{\partial x_i}$ denotes the material derivative of density. In the surface layer, the first term in Eq. I.1 is negligible compared to the second term (Stull 1988), which yields the incompressible formulation of the mass conservation equation

$$\frac{\partial u_i}{\partial x_i} = 0, \tag{I.2}$$

Conservation of momentum In the surface boundary layer, the flow is seen in a plane tangent to the Earth, and the Coriolis force can be neglected. Assuming that air viscosity μ is constant in space and using the incompressibility assumption already made ($\frac{D\rho}{Dt} = 0$), the conservation of momentum equation can be written as

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = \underbrace{-\frac{1}{\rho}\frac{\partial p}{\partial x_i}}_{\text{(i)}} + \underbrace{\nu \frac{\partial^2 u_i}{\partial {x_j}^2}}_{\text{(ii)}} \underbrace{- \delta_{i3} g}_{\text{(iii)}}, \tag{I.3}$$

This equation means that atmospheric flow motions are governed by three actions: the pressure gradient (i); the effects of air viscosity (ii), with $\nu = \mu/\rho$ ($\mathrm{m^2\,s^{-1}}$) the kinematic viscosity of air; and gravity (iii), with $g = 9.81\,\mathrm{m\,s^{-2}}$ the gravitational acceleration on Earth and δ_{ij} the Kronecker function defined as $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise.

Conservation of energy The first law of thermodynamics, which describes the conservation of enthalpy, is usually written for the potential temperature θ when describing atmospheric flows. The potential temperature describes the temperature that a parcel of air initially at temperature T and pressure p would have if adiabatically compressed or

expanded to a reference pressure level p_0, usually defined as 10^5 Pa. Using the ideal gas state equation:

$$p = \rho R^* T \tag{I.4}$$

with $R^* = R/M_a$ the specific gas constant of air where R is the universal gas constant and M_a is the molar mass of air, the potential temperature of air is defined as:

$$\theta = T \left(\frac{p_0}{p} \right)^{\frac{R^*}{c_p}}. \tag{I.5}$$

By making it possible to compare parcels of air at different heights, the potential temperature is convenient for expressing the conservation of energy, which is then expressed as:

$$\frac{\partial \theta}{\partial t} + u_j \frac{\partial \theta}{\partial x_j} = \frac{1}{\rho c_p} \frac{\partial}{\partial x_j} \left(\lambda \frac{\partial \theta}{\partial x_j} \right) + S_\theta, \tag{I.6}$$

with c_p ($\mathrm{J\,kg^{-1}\,K^{-1}}$) the specific heat at constant pressure of the fluid, λ ($\mathrm{W\,m^{-1}\,K^{-1}}$) its molecular thermal conductivity, and S_θ a source term associated with latent heat released during phase change and radiation effects.

Boussinesq approximation In the surface layer of the ABL, we can apply the Boussinesq approximation, which consists of neglecting density variations in momentum and energy conservation equations except in the gravity term in Eq. I.3 (Stull 1988). This leads to the simplified Navier-Stokes equations for the atmospheric surface layer:

$$\begin{cases} \dfrac{\partial u_i}{\partial x_i} = 0, & \text{(I.7a)} \\ \dfrac{\partial u_i}{\partial t} + u_j \dfrac{\partial u_i}{\partial x_j} = -\dfrac{1}{\rho_0} \dfrac{\partial p'}{\partial x_i} + \dfrac{1}{\rho_0} \dfrac{\partial \tau_{ij}}{\partial x_j} + \delta_{i3} \left(\dfrac{\theta'}{\theta_0} \right) g, & \text{(I.7b)} \\ \dfrac{\partial \theta}{\partial t} + u_j \dfrac{\partial \theta}{\partial x_j} = \dfrac{1}{\rho_0 c_p} \dfrac{\partial}{\partial x_j} \left(\lambda \dfrac{\partial \theta}{\partial x_j} \right) + S_\theta, & \text{(I.7c)} \end{cases}$$

where hydrostatic equilibrium is assumed, i.e $\dfrac{\partial p_0}{\partial z} = -\rho_0 g$ with ρ_0 the reference density, p_0 and θ_0 the associated pressure and potential temperature using the equation of state for ideal gases (Eq. I.4), and where p' and θ' are perturbation around the reference state.

Transport of pollutant Under the hypothesis described above, the mass conservation principle applied to a pollutant species of volume concentration c leads to the transport equation:

$$\frac{\partial c}{\partial t} + \underbrace{u_j \frac{\partial c}{\partial x_j}}_{\text{(I)}} = \underbrace{D \frac{\partial^2 c}{\partial x_j^2}}_{\text{(II)}} + R + S, \tag{I.8}$$

where D ($\mathrm{m^2\,s^{-1}}$) is the mass diffusivity of the species in the atmosphere, R is a chemical reaction term, and S is the rate of emission (resp. removal) of the pollutant sources

(resp. sinks). In this thesis, we restrict ourselves to the study of passive species both in the chemical sense, i.e. that do not react with other chemical compounds in the atmosphere ($R = 0$) on microscale time lengths, and in the mechanical sense, i.e. that do not influence the flow (otherwise, an additional buoyancy contribution would appear in Eq. I.7b). Neither source momentum effects nor deposition are studied since we are focusing exclusively on the dispersion of gaseous species released passively. As a gateway to these complex phenomena, we refer to Seinfeld and Pandis' book (1998).

The advection term (I) in Eq. I.8 accounts for the fact that the species is carried by the atmospheric flow, while the diffusion term (II) corresponds to the natural movement of the species from areas of higher concentration to areas of lower concentration. The latter results from random molecular motion, and therefore occurs even in the absence of bulk fluid movement. These are the two fundamental transport processes, which is why Eq. I.8 is also known as the advection-diffusion equation. In the ABL, molecular diffusion is several orders of magnitude smaller than the advection term and can be neglected (Stull 1988).

I.1.2.2 Atmospheric turbulence and impact on dispersion

Surface atmospheric flows are highly turbulent (Stull 1988) and are therefore characterized by the development of three-dimensional vortices, called eddies, whose size, location and orientation are constantly changing. In the ABL, the length scale of these eddies varies from a few kilometers down to the Kolmogorov scale on the order of millimeters, where kinetic energy is dissipated as heat. These vortices cause fluid particles to move in chaotic patterns, as opposed to laminar flows, making the prediction of atmospheric flows particularly challenging because of the resulting scale disparity.

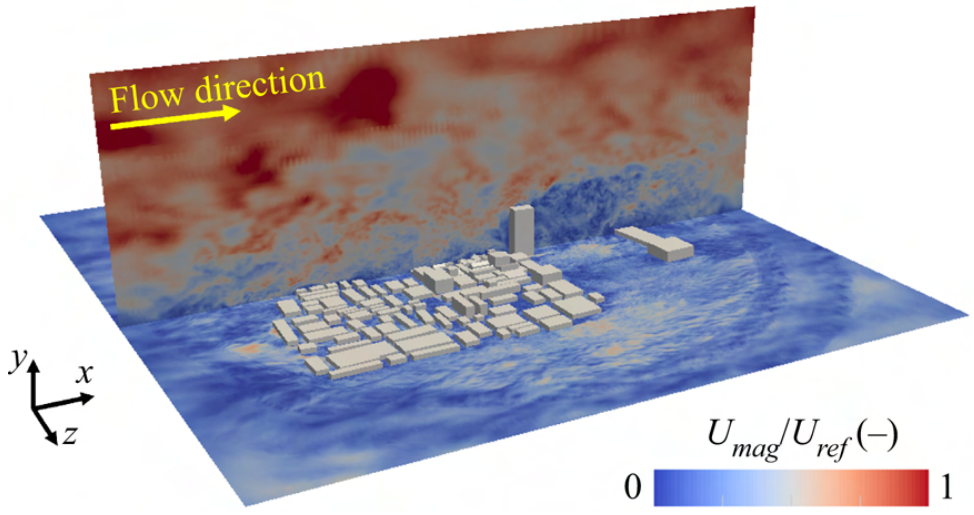


Figure I.4: *Large-eddy simulation from Hwang and Gorlé (2023) of the instantaneous wind normalized velocity U_{mag}/U_{ref}, with U_{ref} a reference wind speed, in a vertical plane and a horizontal plane of the urban boundary layer. The area studied is a neighborhood slum in Dhaka (Bangladesh).*

As a result, the instantaneous wind field is highly heterogeneous, featuring considerable variability in both space and time. Figure I.4 shows the disordered aspect of velocity in the urban boundary layer and the wide range of structure scales. In addition, the urban canopy increases flow turbulence, generating eddies up to several times the size of the buildings, as evidenced in the wake of the tallest building in Fig. I.4. Finally, we note that turbulence contributes to the coupling of the scales of motion mentioned in Sect. I.1.1 by transferring energy back and forth between eddies of different sizes.

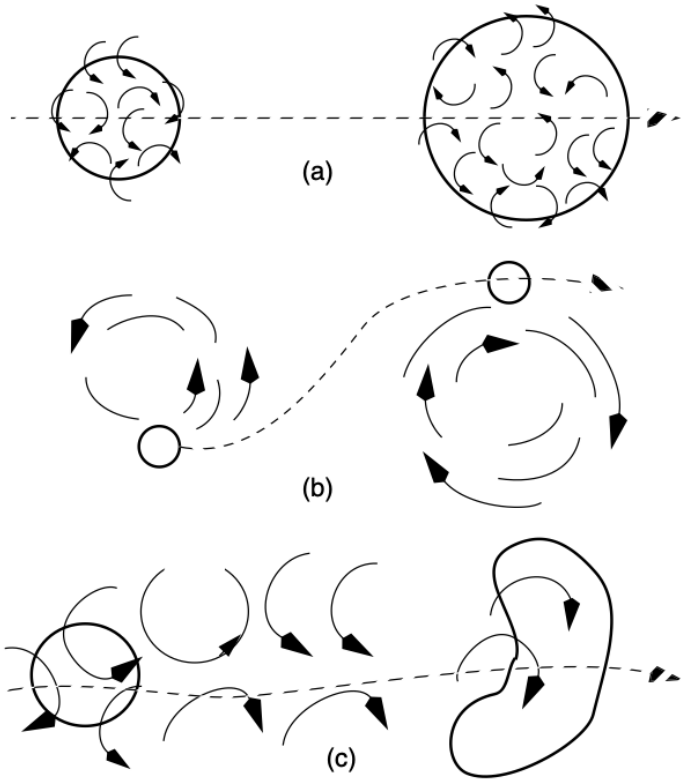


Figure I.5: *Interaction regime between pollutant dispersion and turbulent eddies of size smaller (a), larger (b), and comparable (c) to the characteristic plume scale. Extracted from Seinfeld and Pandis (1998).*

Turbulence also has a significant and complex impact on dispersion. First, it should be stated that species are dispersed much more efficiently in a turbulent flow than in a comparable laminar flow (Pope 2000). How turbulent eddies influence dispersion depends on their size relative to that of the plume, as shown in Fig. I.5. Eddies significantly smaller than the plume essentially spread the plume (Fig. I.5a). This is often called turbulent diffusion as it acts similarly to molecular diffusion. Conversely, eddies much larger than the plume advect the plume, without modifying the internal configuration of the plume (Fig. I.5b). Finally, turbulent structures of a size comparable to the plume reshape it and expand its edge, as illustrated in Fig. I.5c.

Because of the importance of these phenomena, accurately representing urban boundary layer turbulence in all its complexity is essential for dispersion modeling.

I.1.2.3 Atmospheric stability and impact on dispersion

Atmospheric stability corresponds to the equilibrium state of air parcels in the atmosphere from a thermal point of view. The potential temperature θ (Eq. I.5) provides a simple way of characterizing it:

- if $\frac{\partial \theta}{\partial z} > 0$, the atmosphere is stable. The air near the surface has a lower potential temperature than the air aloft, and air parcels that are displaced vertically tend to return to their original position, resisting vertical motion. This is often the case during clear nights when the ground loses heat, cooling the air near the surface.
- if $\frac{\partial \theta}{\partial z} < 0$, the atmosphere is unstable. The air near the surface is much warmer than the air aloft, in terms of potential temperature, and air parcels moving vertically tend to keep rising.
- if $\frac{\partial \theta}{\partial z} = 0$, the atmosphere is neutral. Potential temperature is constant over altitude, which means that the decrease in absolute temperature with altitude is in equilibrium with the dry adiabatic lapse rate (the rate at which unsaturated air cools as it rises). Air parcels displaced vertically neither rise nor sink because of buoyancy; their behavior depends on other forces or processes.

By modifying the mean wind velocity profile and the turbulence mixing, the atmospheric stability strongly influences pollutant dispersion, as illustrated in the case of continuous release in Fig. I.6. In a stable atmosphere, vertical motions of air parcels are attenuated and dispersion is reduced (Fig. I.6c). Conversely, under unstable atmospheric conditions, large convective cells favor dispersion and are responsible for the plume meandering (Fig. I.6a). In a neutral atmosphere, there is no buoyancy effect and the turbulence generated by the surface is not attenuated, leading to a steady increase in plume cross-section (Fig. I.6b). Note that the potential temperature profile is not necessarily monotonic, which can lead to other dispersion regimes such as lofting (Fig. I.6d) and fumigation (Fig. I.6e).

In this thesis, we focus on neutral atmospheric conditions, for which the governing equations can be simplified as shown in the next section.

I.1.2.4 Similarity theory under neutral stratification conditions

The dynamic similarity between two systems refers to the fact that the ratio of forces involved is the same in both systems, ensuring similar contributions to flow dynamics. Similarity for the pollutant transport is defined in the same way. These similarity conditions can be obtained by writing the governing equations in terms of dimensionless quantities. It informs on how the physical quantities of interest scale which is essential in wind-tunnel studies (Bezpalcová 2007), but also for the modeler as it provides a theoretical basis for model validation and simplification. It can also be used to reduce the number of predictions required to describe the behavior of the system (Sousa et al. 2018). In this section, we limit the discussion to neutral stratification conditions.

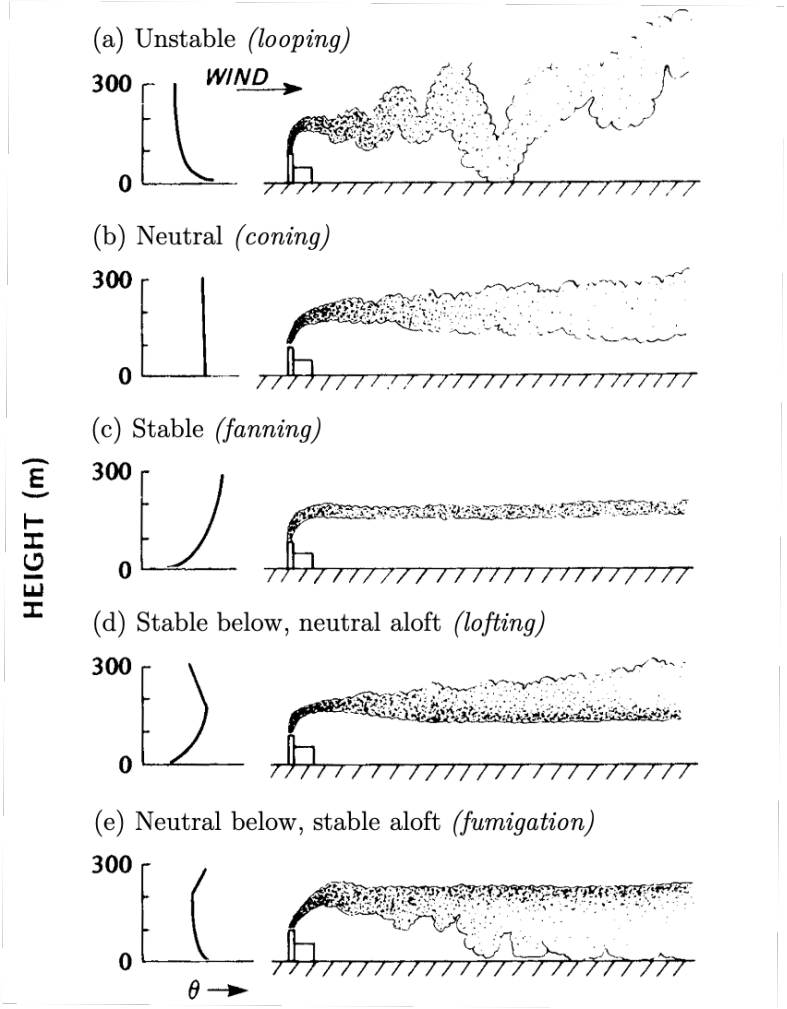


Figure I.6: *Effect of the atmospheric stability on the plume shape and trajectory, adapted from Oke (1987). The profiles on the left correspond to the vertical profiles of potential temperature θ.*

We first introduce the dimensionless flow quantities

$$\mathbf{u}^* = \frac{\mathbf{u}}{U_0}, \quad \mathbf{x}^* = \frac{\mathbf{x}}{L_0}, \quad t^* = \frac{tU_0}{L_0}, \quad p'^* = \frac{p'}{\rho_0 U_0^2}, \tag{I.9}$$

where the subscript 0 denotes the characteristics scales. These scales are typical measures of the problem such as the obstacle length for L_0 and reference speed at a given height for U_0. Their definition is arbitrary but must be consistent when comparing different flows. We also introduce the dimensionless pollutant concentration

$$c^* = \frac{cU_0L_0^2}{Q_{source}}, \tag{I.10}$$

with Q_{source} the volumetric release rate. The dimensionless source term S^* is defined similarly.

Under neutral stratification conditions, the potential temperature is spatially uniform implying that mass and momentum conservation equations (Eqs. I.7a, b) become independent from the energy conservation equation which is trivially satisfied (Eq. I.7c). The governing equations in terms of dimensionless quantities are then written as follows

$$\begin{cases} \dfrac{\partial u_i^*}{\partial x_i^*} = 0, & \text{(I.11a)} \\ \dfrac{\partial u_i^*}{\partial t^*} + u_j^* \dfrac{\partial u_i^*}{\partial x_j^*} = \dfrac{\partial p'^*}{\partial x_i^*} + \left[\dfrac{\nu}{U_0 L_0}\right] \dfrac{\partial^2 u_i^*}{\partial x_j^{*2}} & \text{(I.11b)} \\ \dfrac{\partial c^*}{\partial t^*} + u_j^* \dfrac{\partial c^*}{\partial x_j^*} = \left[\dfrac{\nu}{L_0 U_0}\right] \left[\dfrac{D}{\nu}\right] \dfrac{\partial^2 c^*}{\partial x_j^{*2}} + S^*, & \text{(I.11c)} \end{cases}$$

which reveals the Reynolds number

$$Re = \frac{U_0 L_0}{\nu}, \tag{I.12}$$

and the species Schmidt number

$$Sc = \frac{\nu}{D}. \tag{I.13}$$

In the atmospheric surface layer, the Schmidt number depends on the transported species but is typically of the order of 0.2 to 2 depending on the size of the molecule, while the Reynolds number is usually greater than 10^6 (Stull 1988). This implies that i) surface-layer atmospheric flows are inherently turbulent which induces the effects presented in Sect. I.1.2.2, ii) the viscous effects in Eq. I.11b become negligible, iii) the transport of species is dominated by advection, as the molecular diffusion term in Eq. I.11c becomes negligible. This weak dependence of flow and dispersion on Reynolds number effects implies that flow velocity and species concentration scale linearly with the reference velocity U_0.

It is also possible to establish the similarity theory using the Buckingham π-theorem which leads to the well-known Monin-Obukhov similarity (Monin and Obukhov 1954), see for example Stull's book (1988). In this formulation, all the mean and turbulent wind and temperature of the atmospheric surface layer depend on four scaling scales: the surface roughness length z_0, the displacement length d, the Monin-Obukhov length L_{MO}, and the friction velocity u_*. z_0 and d are related to surface roughness elements, including the buildings that compose the urban canopy layer, while L_{MO} describes thermal stratification effects ($L_{MO} \to \infty$ under neutral conditions). Finally, the friction velocity is defined as:

$$u_* = \frac{1}{\rho_0}\sqrt{\tau_0} = \sqrt{\overline{u'w'}}, \tag{I.14}$$

with τ_0 the surface Reynolds stress, and $\overline{u'w'}$ the correlation between horizontal and vertical wind velocity fluctuation which represents the vertical transport of momentum by turbulence. The friction velocity can be assumed to be constant in the roughness sublayer. Moreover, mean and turbulent velocities are proportional to u_* in this layer. We therefore use it as reference scaling throughout this thesis.

I.1.3 A brief overview of dispersion modeling approaches

Modeling microscale dispersion of pollutants makes it possible to predict the effects of an accidental release of pollutants in an urban environment, and also to carry out upstream design studies, without the complexity of the real world or wind-tunnel experiments (Blocken 2015). However, it is a challenging task because of the complex interactions between urban boundary layer flow and dispersion (see Sect. I.1.2). In this section, we briefly introduce the standard dispersion modeling approaches. For a detailed overview, the reader is referred to Holmes and Morawska (2006) and Leelőssy et al. (2014).

Gaussian and operational models Assuming steady and spatially homogeneous wind flow, the solution of the transport equation (Eq. I.8) for a point source S which releases a passive species at a constant-in-time flow rate Q is a plume where the mean concentration $\overline{c}$ follows a Gaussian distribution in both vertical and cross-flow directions (Seinfeld and Pandis 1998):

$$\overline{c}(x,y,z) = \frac{Q}{2\pi U \sigma_y \sigma_z} \exp\left(-\frac{(y-y_s)^2}{2\sigma_y^2} - \frac{(z-z_s)^2}{2\sigma_z^2}\right), \tag{I.15}$$

where U is the uniform velocity in the streamwise direction (x-direction), $\mathbf{x}_s = (x_s, y_s, z_s)$ is the source location, and where σ_y and σ_z are spread parameters that depend on the distance from the source $(x - x_s)$ and on the atmospheric stability according to the so-called Pasquill-Grifford curves (Turner 1969; Hanna et al. 1982). Other parameterizations can be used to account for additional complex physical processes such as buoyancy, chemistry or deposition (Seinfeld and Pandis 1998). The Gaussian model is a standard approach that is used in a large number of operational models (Carruthers et al. 1994; Cimorelli et al. 2005; Perry et al. 2005).

To take into account the effect of the urban canopy, Gaussian models describe the urban canopy using roughness parameters that affect the spread terms (Hanna et al. 2003; Philips et al. 2013). However, this mesoscopic approach does not allow explicit modeling of microscale dispersion within the urban canopy (Carruthers et al. 1994; Scaperdas 2000), making these models unsuitable for the microscale application context of this thesis. Still, it should be noted that more complex models are developed to allow the description of pollutants in the urban canopy, hybridizing Gaussian and simple box models with local parameterizations of flow and turbulence (Soulhac et al. 2011, 2012).

Computational fluid dynamics (CFD) models solve the pollutant concentration transport equation (Eq. I.8) based on the velocity field obtained from the Navier-Stokes equation (Eqs. I.2–I.6). In contrast to operational models, they explicitly take into account the effect of the urban canopy on flow, therefore they are less restricted in terms of assumptions and less sensitive to parameterization. Consequently, there is a growing consensus for using CFD models in atmospheric dispersion modeling at microscale (Holmes and Morawska 2006; Blocken et al. 2013). However, CFD models require very fine spatial discretization of the equations and hence considerably higher computational

cost compared to established operational models. This problem is further detailed in Sect. I.2.1.

The biggest challenge that comes with solving the Navier-Stokes equations is representing turbulence which, as explained in Sect. I.1.2.2, is highly nonlinear, chaotic and occurs on a wide range of scales. The three main CFD approaches correspond to three different paradigms for taking turbulence into account:

- the **Direct Numerial Simulation (DNS)** approach solves the Navier-Stokes equations directly for all spatial and temporal scales making it ideal for capturing fine-scale turbulence and detailed flow structure. This is the most accurate approach, but also the most expensive, since it requires spatial resolution down to the Kolmogorov scale. In an urban canopy context, some studies have used DNS to simulate the flow around one or a few cubical obstacles in reduced-scale configuration (Coceal et al. 2006; Rossi et al. 2010). But to date, this approach is considered unrealistic for simulating dispersion at the real scale.

- the **Reynolds-Averaged Navier-Stokes (RANS)** approach solves the Navier-Stokes equations in a statistically-averaged sense. This implies that the effect of turbulence on the flow has to be fully modeled, using for example the standard k-ϵ model (Jones and B.E 1972). Because of its reasonable computational cost (see Sect. I.2.1), RANS is historically the most widely used CFD approach in microscale applications (Meroney et al. 1999; Calhoun et al. 2004; Hanna et al. 2004; Milliez and Carissimo 2007; Koutsourakis et al. 2012; Tominaga and Stathopoulos 2013; Toparlar et al. 2017).

- **Large-Eddy Simulation (LES)** provides a balance between DNS and RANS approaches, simulating large turbulent structures explicitly while modeling the smaller scales. This compromise limits the computational cost compared to DNS and the need for parameterization compared to RANS. The unsteady nature of LES also makes it possible to present the temporal variability of the ABL. For these reasons, LES is also becoming popular in microscale studies (Gousseau et al. 2011; Moonen et al. 2012; Tominaga and Stathopoulos 2013; Blocken 2014; García-Sanchez et al. 2018).

Most of the CFD solvers adopt an Eulerian approach to solve the Navier-Stokes equations. This means that a fixed reference frame is used to describe the flow, and the equations are solved in terms of fields, discretized on a fixed grid. But it is also possible to adopt a Lagrangian approach, which consists of tracking fluid parcels along their trajectories. This approach is particularly natural to describe pollutant dispersion and has interesting properties such as being non-dispersive, which improves the representation of plume fronts, and being intensively parallelizable. Note that Eulerian and Lagrangian approaches can be combined, for example, a Lagrangian approach can be used to model the transport of a species in the flow field obtained by an Eulerian CFD model (Jicha et al. 2000; Bahlali et al. 2019). In this thesis, we adopt an Eulerian viewpoint.

It is important to note that due to their computational cost, CFD approaches limit the resolution of the governing equations to the microscale area of interest. However,

as discussed in Sect. I.1.1, although mainly influenced by interaction with the built environment, microscale flow and dispersion are also directly by larger-scale atmospheric motions. This is why the accuracy of CFD dispersion models is strongly linked to the representativeness of boundary conditions, both for large-scale atmospheric conditions and for the geometry of the urban canopy. Uncertainties related to the representativeness of the boundary conditions of microscale CFD models are discussed in detail in Sect. I.2.2.

I.1.4 Experiments of dispersion in urban environment

Validating models by confronting them with the real world has always been at the core of the scientific method. This is how the overall quality of a model can be assessed and how limitations to its use can be defined. Although CFD models are theoretically very accurate, they are no exception to the fundamental principle of validation (Meyers et al. 2008). In the past decades, particular efforts have been conducted in the field of microscale atmospheric flow and dispersion to harmonize validation procedures and to provide best practice guidelines for CFD modeling (Franke et al. 2007; Schatzmann et al. 2010; Tominaga and Stathopoulos 2013; Blocken 2015). It should be noted that the quality of a model is multifactorial and its appreciation depends on the purpose of the model (Blocken and Gualtieri 2012). In the following, we give an overview of some dispersion experiments in urban environments commonly used for validation.

Reduced-scale laboratory experiments can be used to validate dispersion models. Their advantage over full-scale campaigns is twofold: i) through greater instrumentation, typically using laser Doppler anemometry and flame ionization detectors (Bezpalcová 2007; Garbero 2008), they can provide measurements of velocity and concentration at high spatial and temporal resolution for validation, which provides an accurate description of the operating boundary conditions; ii) they enable repeatability of experiments and the acquisition of very long time series to obtain converged flow statistics, which is not feasible with field campaigns (Schatzmann and Leitl 2011). These experiments are carried out in wind tunnels (Meroney et al. 1999; Kastner-Klein et al. 2004; Garbero et al. 2010; Hajra and Stathopoulos 2012) or water channels (Macdonald and Ejim 2002; Hilderman et al. 2008), which both rely on a thorough dimensional analysis to ensure that the reduced-scale flow has the same Reynolds number (Eq. I.12) as its full-scale equivalent. However, it is important to note that reduced-scale experiments cannot represent the complete range of interactions that occur in real urban boundary layers. Besides, most of the wind tunnels cannot reproduce stable and unstable atmospheric stratification conditions (Blocken 2014).

Full-scale dispersion experiments in real urban environments are essential to represent the full complexity of dispersion processes in the urban boundary layer but are time- and resource-intensive.

At the street scale, numerous full-scale experiments have been carried out to characterize air pollution induced by the traffic in street canyon (Pfeffer et al. 1995; Namdeo et al. 1999; Väkevä et al. 1999; Chan and Kwok 2000; Vachon 2001). At the neighborhood

scale, experiments of various passive tracer released have been notably carried out in the cities of Copenhagen (Denmark) (Gryning and Lyck 1984), Salt Lake City (U.S.) and Oklahoma City (U.S.) as part of the URBAN 2000 and JOINT URBAN 2003 campaigns (Allwine et al. 2002, 2004), Birmingham (U.K.) (Britter et al. 2002), Basel (Switzerland) for the BUBBBLE campaign (Rotach et al. 2004), and London (U.K.) in the framework of the DAPPLE project (Wood et al. 2009).

These experiments are fully representative of the urban canopy complexity but cannot be transposed to the study of other cities because of the differences in city architectures. In order to distinguish the general effects of the urban canopy from those specific to a given site, other dispersion experiments were conducted with simplified representations of the urban canopy: by an array of cubes in the field experiments of Davidson et al. (1995), and by an array of shipping containers during the MUST (Biltoft 2001) and Jack Rabbit II (Fox et al. 2022) field campaigns. In the Kit Fox (Hanna and Chang 2001) and ATREUS-PICADA (Idczak et al. 2007) experiments, the urban canopy (respectively a simplified industrial plant and an arrangement of canyon streets) was represented at a reduced scale while maintaining real meteorological conditions.

I.2 The limits of computational fluid dynamics dispersion models

Despite their substantial computational cost and great promises, the accuracy of microscale CFD models can still be improved (Blocken 2014; Dauxois et al. 2021). Some studies even find limited gain in overall accuracy when using CFD models compared to operational/empirical models (Neophytou et al. 2011; Antonioni et al. 2012). This is because the higher the complexity of the model, the more it gets sensitive to input data uncertainty (Hanna 1989a). In particular, the internal variability of the ABL implies that field measurements are the result of stochastic processes and thus inherently uncertain. This variability affects validation data and propagates to model predictions as observation data are used to define the model boundary conditions. This is why the internal variability of the ABL is often referred to as one of the main reasons for the discrepancies between CFD models and field experiments (Neophytou et al. 2011; Antonioni et al. 2012; García-Sanchez et al. 2018; Dauxois et al. 2021).

In this section, we take a closer look at the problem of the computational cost of CFD models (Sect. I.2.1), before proposing an insightful classification of the uncertainties that limit the accuracy of CFD models (Sect. I.2.2). In our view, computational cost and uncertainty are the two main limitations of CFD microscale dispersion models, and overcoming them is the main issue addressed in this thesis.

I.2.1 The computational burden of CFD models

As previously mentioned, the computational cost of CFD models stems from the fact that they require a very fine level of discretization. It is important to have a mesh fine enough to accurately account for the effect of small-scale turbulence on the flow and dispersion. The computational cost of CFD models therefore depends heavily on how they treat turbulence.

In DNS, all the eddie scales down to the Kolmogorov scales must be solved. For wall-bounded flows, such as in the ABL, this leads to a number of grid points proportional to $Re^{9/4}$ with Re the Reynolds number based on the integral scale of the flow (Pope 2000). In terms of computational cost, assuming that the number of operations scales with the number of grid points, and that the time step is determined by a standard CFL stability condition, the total cost rises to $\mathcal{O}(Re^3)$ (Reynolds 1990). Considering the very high Reynolds of the ABL ($Re \sim 10^6$–10^8), this challenges the computing power of the world's most powerful supercomputers to date, which have just broken the 10^{18} flop barrier (Trystram 2022).

For LES, grid requirements are less important since the effect of smaller scales is modeled. For instance, Choi and Moin (2012) estimate the computational cost of an LES of a turbulent boundary layer to $\mathcal{O}(Re^{2.5})$. Using wall-laws to model the effect of the inner part of the boundary layer on the flow can cut this cost down to $\mathcal{O}(Re^{4/3})$ (Piomelli 2020), making it suitable for atmospheric flow modeling.

Finally, the computational cost of RANS simulations is way smaller since the necessary number of grid points does not depend on the Reynolds number in the streamwise and

spanwise directions, and scales as $\ln(Re)$ along the wall-normal direction (Pope 2000).

In concrete terms, to simulate flow and dispersion within the urban canopy for a few hundred seconds, a RANS model typically takes between 100 and 1 000 CPU core hours[1], depending on the size of the simulated domain and the grid resolution (Hanna et al. 2004). LES is much more demanding, with computational costs approximately one to two orders of magnitude higher than with RANS (Cheng et al. 2003; Xie and Castro 2006; Salim et al. 2011).

The cost of CFD microscale dispersion models has a direct effect on their use. Indeed, it is currently not possible to use a CFD model for real-time prediction of pollutant dispersion in an urban area, since the computational time is far greater than the time scales involved at the microscale. As a result, CFD models are used more as research tools to improve scientific understanding of the processes involved, but also to formulate parametrization for operational air quality models (Philips et al. 2013; Hertwig et al. 2018; Grylls et al. 2019).

This could change as computing power continues to increase, driven by the race between the world's largest supercomputers (Trystram 2022), but also thanks to advances in solver parallelization (Afzal et al. 2017) and the capabilities offered by new technologies such as GPUs (Graphics Processing Units) (Muñoz-Esparza et al. 2021). But this comes at a cost as, to perform these simulations, high-performance computing centers require a considerable amount of energy in electricity and cooling, as well as substantial investment in infrastructure, hardware and maintenance. It also means that simulations have a high carbon footprint (Berthoud et al. 2020) and therefore contribute to global warming. Note that this is not mitigated by recent gains in energy efficiency, as they are more than offset by the increase in computing volume, which is a typical example of the rebound effect (Trystram 2022). To raise awareness of this issue, an estimate of the emissions associated with the calculations carried out in this thesis is provided in Appendix C.

I.2.2 Uncertainties in microscale atmospheric CFD models

Model uncertainties can be divided into two categories: aleatory uncertainties, and epistemic uncertainties (Kiureghian and Ditlevsen 2009; Beven et al. 2018). Aleatory uncertainties are inherent to the stochastic nature of the physical system under consideration and are therefore irreducible. For example, the statistical behavior of a die is perfectly known, but the outcome of a roll will always remain uncertain. Epistemic uncertainties relate to a lack of knowledge, either about the physics of the phenomenon or about the model inputs. This implies that they can be reduced, for example by using a higher model resolution, a richer representation of the phenomenon, or more numerous and precise data to define the model inputs. Tackling these uncertainties is one of the modeler's main challenges, as this can improve the model in terms of accuracy, preci-

[1]Corresponds to the total wall-clock time of the simulation multiplied by the number of processor cores used. It corresponds to the actual cost of the simulation in terms of computational resources/energy consumption and allows for a fair comparison between simulations performed using different supercomputers.

sion and robustness. In the following, we present a brief overview of uncertainties in the specific context of microscale atmospheric CFD models.

I.2.2.1 Epistemic uncertainties

The epistemic uncertainties of a microscale atmospheric CFD model are reduced compared to an analytical model that heavily depends on parameterization, but are still significant. In this thesis, we propose to divide these uncertainties into two subcategories:

a) Boundary condition uncertainties. They are due to the fact that microscale CFD models only solve the governing equations over a limited computational domain. Interactions with the exterior of the domain are represented using boundary conditions. Representativeness errors can thus arise from the boundary modeling assumptions as well as from uncertainty in the data used to calibrate the boundary conditions. For CFD dispersion models, these uncertainties relate in particular to:

- The meteorological conditions which correspond to the wind velocity and direction forcing imposed at the boundaries to account for the operating conditions but also for the larger scales of motions of the ABL (see Sect. I.1.1). These uncertainties have been extensively studied (García-Sánchez et al. 2014; Lucas et al. 2016; Wise et al. 2018) and they are known to be a major source of uncertainty in atmospheric flow, and therefore in pollutant dispersion.

  In the specific urban environment context, these uncertainties can be exacerbated by strong non-linear response between local flow and boundary conditions. In particular, small deviations in wind direction can lead to large changes in flow patterns in the urban canopy. In contrast, in some areas, the flow is mostly determined by the local buildings, and hence insensitive to larger-scale conditions (Fig. I.7).

- The urban geometry uncertainties arise from errors in the position and shape of buildings. They may come from a lack of measurements or errors in the map, but also from the fact that geometrical details are often omitted or simplified to reduce the computational cost of models (Franke et al. 2007).

  In the urban canopy layer, these uncertainties have significant impacts. For instance, Montazeri and Blocken (2013) shows that building facade details such as balconies drastically influence the flow pattern and the overall pressure distribution on the building. Gromke et al. (2016) demonstrate that taking into account hedgerows has a significant impact on the mean concentration predicted within a street canyon. Note that even in cases featuring simplified geometries, such as the array of containers used in the MUST field experiment, small geometry irregularities can have a significant impact on the microscale features of the flow as demonstrated by Santiago et al. (2010).

- The pollutant source can also be uncertain both in terms of location and intensity. This makes the task of predicting pollutant dispersion significantly more complex, and this requires observations to locate the source (Winiarek

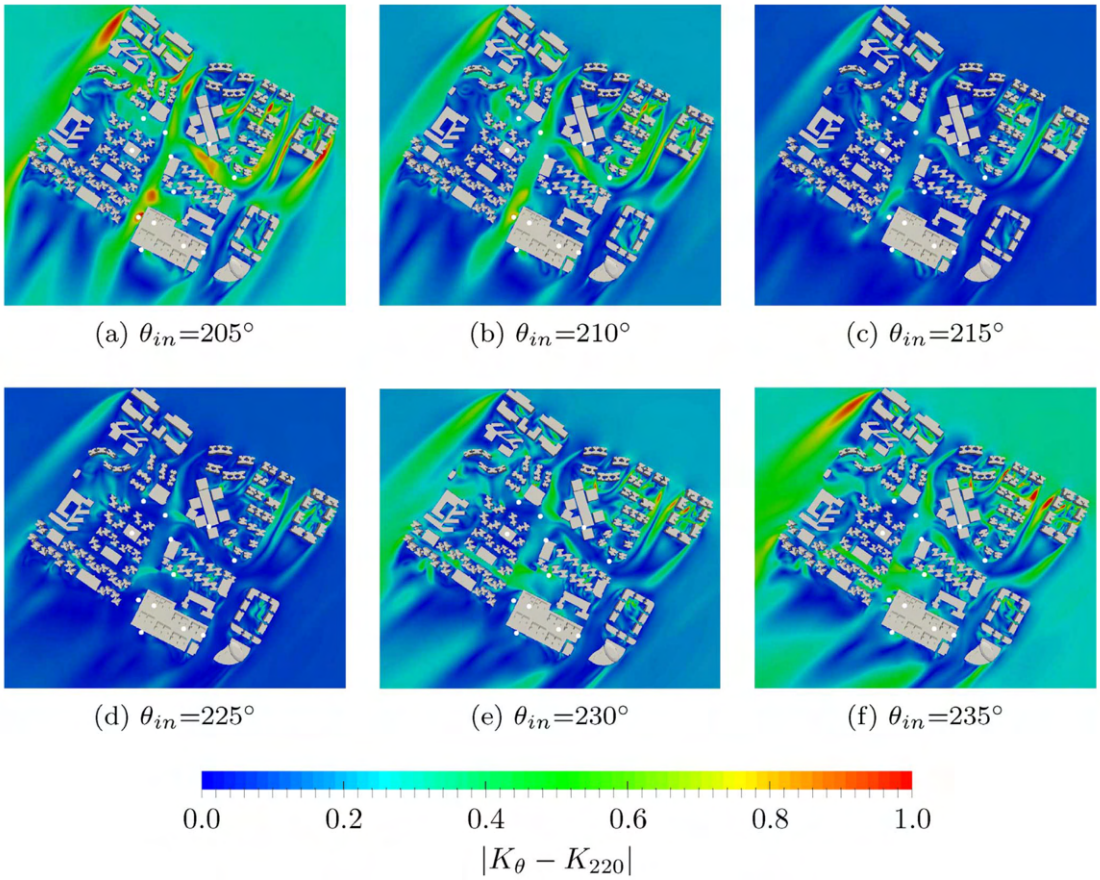


Figure I.7: *Effect of minor changes in the wind direction on the local flow in a neighborhood of Singapore, from Wise et al. (2018). Contours represent the ratio K_θ between the local wind speed at 20 meters above the ground and a reference speed, predicted by RANS simulations for varying incoming wind direction θ.*

et al. 2012; Lucas et al. 2016). In addition, the pollutant release rate may change over time, and there may be uncertainties related to source modeling (Spicer and Tickle 2021).

b) Structural uncertainties correspond in this thesis to all the uncertainties that are inherent to the model solver and its underlying modeling assumptions, outside of the boundary conditions. In CFD models, this mainly concerns turbulence modeling and wall modeling errors. It is a standard research area that has also been investigated in the specific context of microscale atmospheric flows in urban environments (Blocken et al. 2008; Yue Yang and Wang 2008; Tominaga and Stathopoulos 2009; Gorlé and Iaccarino 2013; Gorlé et al. 2015; Xiao et al. 2016). Note that LES is expected to have reduced uncertainties related to turbulence modeling compared to RANS (Gousseau et al. 2011; García-Sanchez et al. 2018). In the classification we propose, structural uncertainty also includes numerical errors, which depend on the solver and numerical scheme used, as well as the model sensitivity to the computational grid used.

I.2.2.2 Aleatory uncertainty arising from the internal variability of the ABL

Definition and characterization As shown in Fig. I.8, the atmospheric flow is unsteady with strong variations occurring over a wide range of frequencies corresponding to the time-scales of the atmospheric eddies. When considering averages over finite periods, this internal variability yields sampling errors and is therefore a source of aleatory uncertainty. This uncertainty is assumed to explain part of the discrepancies between field-scale experiments on the one hand, and wind-tunnel experiments and CFD model predictions on the other hand (Schatzmann and Leitl 2011; Neophytou et al. 2011; Antonioni et al. 2012; García-Sanchez et al. 2018). We emphasize that this uncertainty does not just concern CFD model predictions since it is inherent to the phenomenon under study.

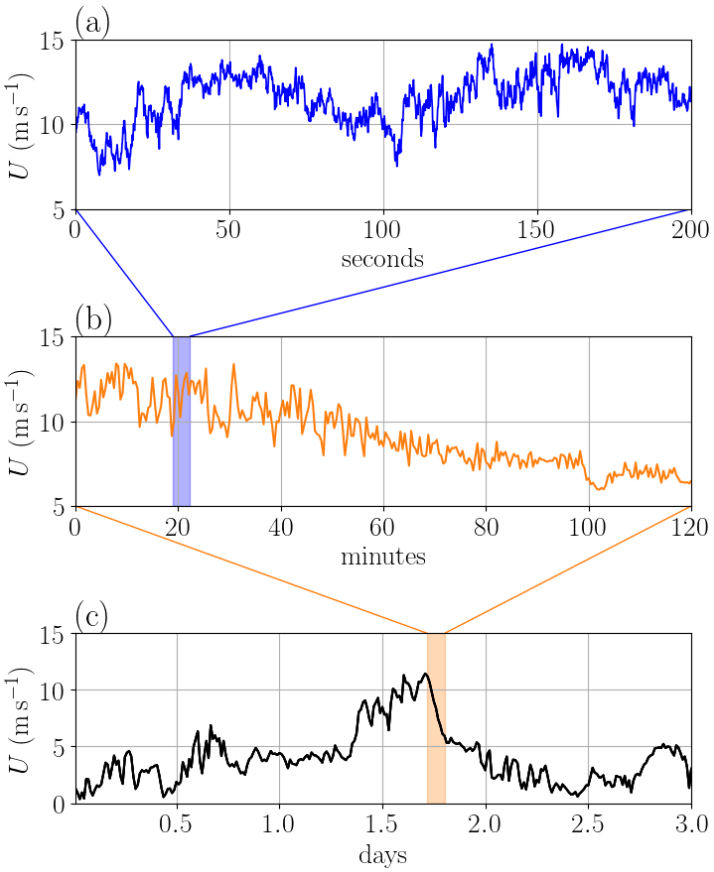


Figure I.8: *Measurements of the horizontal wind velocity at 10 meters above the ground during the MUST field experiment (Biltoft 2001) over 200 seconds (a), 120 minutes (b), and 3 days (c).*

In this thesis, we distinguish the microscale internal variability, which is related to microscale turbulence fluctuations in the surface layer induced by buoyancy, wind shear and interaction with the obstacles (Fig. I.8a), from the mesoscale variability induced by large-scale fluctuations occurring in the ABL (Figs. I.8b, c).

To limit the effect of the microscale internal variability of the ABL, it is necessary to achieve statistical convergence of the flow and transport phenomena. This requires the

acquisition and simulation of periods significantly longer than the time scale characteristic of internal variability. It is possible in wind-tunnel experiments where stationary wind conditions can be imposed, but not in field-scale experimental campaigns (Schatzmann and Leitl 2011). Longer acquisitions are indeed affected by transient phenomena such as large-scale fluctuations of the ABL or day-night cycle (Figs. I.8b, c). In microscale studies, it is therefore common to select short analysis periods to minimize the influence of the mesoscale variability. For instance, 200-second quasi-stationary periods have been extracted from each trial of the MUST experiment (Yee and Biltoft 2004). When studying unsteady phenomena like puff pollutant emissions, useful acquisition times are even shorter (Biltoft 2001; Allwine et al. 2004). In all these cases, statistics are computed over relatively short periods and are thus affected by microscale internal variability.

Internal variability in CFD models This source of uncertainty in physical system observations can also affect CFD model predictions, depending on the approach adopted. In the RANS framework, all the scales of turbulence are modeled to predict ensemble-averaged flow and dispersion fields, which implies that microscale internal variability is not represented in RANS predictions, which are deterministic. Meanwhile, in DNS and LES, most or all turbulent scales are explicitly resolved. Their predictions therefore reproduce the microscale fluctuations observed in the ABL (Fig. I.8) and are thus affected by internal variability (Sood et al. 2022). As in wind-tunnel experiments, it is possible to limit the resulting uncertainty by running long simulations (Piomelli 1999) at the expense of increasing the already substantial computational cost (see Sect. I.2.1).

At this point, it is important to note that the distinction between aleatory uncertainty and epistemic uncertainty is sometimes blurred. For instance, observational data used to define the wind velocity boundary conditions are also subject to aleatory uncertainty related to the internal variability of the ABL. This is why the effect of mesoscale internal variability on CFD predictions is often studied as a form of a boundary condition uncertainty using an ensemble of simulations (García-Sánchez et al. 2014; García-Sanchez and Gorlé 2018) or simulations with time-varying inflow conditions (Tominaga and Stathopoulos 2017). In this thesis, we also adopt this point of view that considers mesoscale internal variability as a boundary condition uncertainty.

In contrast to mesoscale variability, the effect of microscale internal variability in CFD predictions has not been explored in detail. Schatzmann et al. (2010) show that microscale variability significantly affects observations from MUST field experiment, while Sood et al. (2022) give error bars associated with the internal variability of LES predictions in a wind resource estimation context. In both cases, this uncertainty is significant but not central to the study. In our view, assessing microscale internal variability would therefore be a significant methodological advance for robust atmospheric CFD model validation when data are acquired over limited periods. It can address the need to go beyond deterministic point-wise model/observations comparison (Schatzmann and Leitl 2011; Harms et al. 2011; Dauxois et al. 2021). It also finds application in model sensitivity analysis and multi-model comparisons as it provides a reference to assess whether changes in CFD model predictions are significant or not.

I.3 Thesis objectives and approach

In this section, we introduce two methods of applied mathematics tailored to address the limitations of microscale CFD models: model reduction in Sect. I.3.1, and data assimilation in Sect. I.3.2. Model reduction allows to greatly speed up CFD models, while data assimilation provides a way to control and reduce part of their uncertainty using observations. In Sect. I.3.3, we present the flowchart of the modeling system that has been designed in this thesis, combining model reduction and data assimilation to enhance microscale pollutant dispersion prediction. Implementing and evaluating this system is the main objective of this thesis.

I.3.1 Accelerating predictions using reduced-order model

Model reduction refers to all the techniques employed to build reduced-order models that emulate the response surface of a costly numerical model. A reduced-order model should be as computationally efficient as possible, while also maximizing its accuracy in reproducing the reference model predictions and avoiding the introduction of non-physical artifacts in the predictions of the quantities of interest. This is very useful in contexts where real-time prediction is required and is an active research topic in fluid mechanics to cope with the heavy computational cost of CFD models (Lassila et al. 2014; Vinuesa and Brunton 2022). It is also used for uncertainty quantification and sensitivity analysis to allow model multi-queries (Cheng et al. 2020). For these reasons, model reduction is particularly in line with the issues discussed in Sect. I.2.

The historical approach to reducing the cost of a numerical model is to substitute it with a simplified model, which adopts a less complex description of the system based on assumptions and simplifications. For pollutant dispersion, this can be done using a Gaussian plume model instead of a CFD model. Information from high-fidelity simulations can then be used to parameterize the simplified model (Philips et al. 2013). However, the accuracy of such an approach remains limited by the assumptions used by the simplified model. For instance, representing the complex local interactions of the flow with the built environment remains out of reach for a Gaussian model.

In this thesis, we focus on model reduction approaches that emulate the full complexity of the CFD model response surface by learning from a database of predictions corresponding to different input parameters. This problem can be divided into two steps (Chinesta et al. 2011):

1. a **reduction step**, which consists of determining a low-dimensional latent space onto which model predictions can be projected without losing their key characteristics (Fig. I.9a),

2. and a **regression step** to characterize the evolution of the reduced-basis projected components with respect to model input parameters.

These two steps pertain to machine learning since they aim at establishing links between the inputs and outputs of a system based on data. Once they are trained, the reduction

and regression models are used successively to predict the emulated field for any new parameter (Fig. I.9b). Note that it is also possible to build a reduced-order model by Galerkin projection of the equations that govern the system physics onto the latent space (Vervecken et al. 2015; Qian et al. 2020; Deng et al. 2021a), but this approach was not considered in this study.

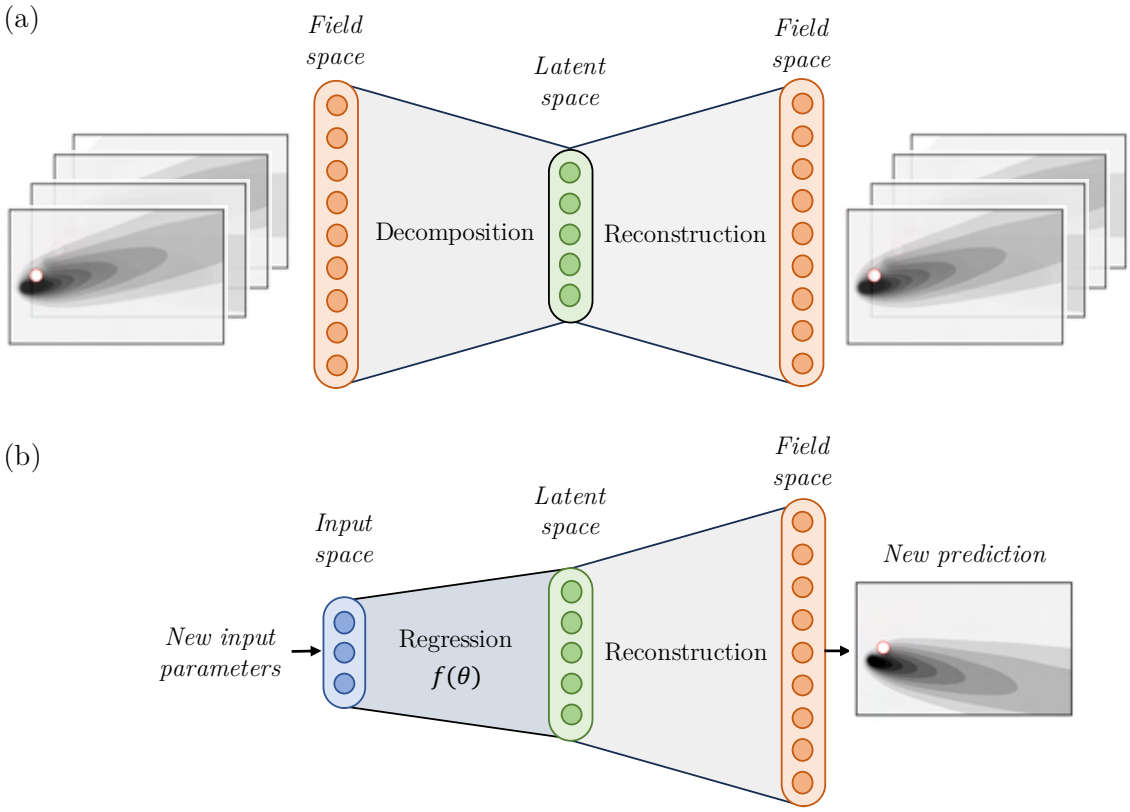


Figure I.9: *Machine-learning-based model reduction principle, adapted from Vinuesa and Brunton (2022). The reduction step (a) consists of learning a latent space to reduce the model output dimension using a database of precomputed predictions. The aim is to preserve as much as possible the key features of the predictions when reconstructing them. The regression step (b) aims at predicting the field of interest for new input parameters. This is done using a pre-trained regressor that directly estimates the decomposition in the latent space associated with the new input parameters.*

The data-driven model reduction approach we adopt has the advantage of being non-intrusive, implying that: i) it does not require modification of the complex CFD code; and ii) it can be adapted to different physical systems, which widens the field of applications of the global framework presented in this thesis (see Sect. I.3.3). The counterpart is that i) it requires the construction of a costly training base, and therefore substantial preparatory work before it can be applied to operational contexts; and ii) it has a limited generalization capability as it cannot predict flow configurations that are significantly different from the ones of the training base. An additional concern is that predictions may not satisfy physical constraints, since they are not directly governed by equations.

Incorporating these constraints into a reduced-order model is an active field of research (Swischuk et al. 2019; Wang et al. 2020a).

In the following, we give a short overview of the different dimension reduction and regression methods used in fluid mechanics, before discussing applications to microscale atmospheric modeling.

Dimension reduction techniques Proper orthogonal decomposition (POD) (Sirovich 1987; Berkooz et al. 1993) is one of the most commonly used reduction algorithms in fluid mechanics (Taira et al. 2017; Vinuesa and Brunton 2022). Also known as principal component analysis or singular value decomposition, POD has become a basic tool for data analysis (Jolliffe and Cadima 2016). It provides a linear subspace of orthogonal modes of decreasing importance to approximate field data. It can be extended to handle time evolution with dynamic mode decomposition (Schmid 2022). More advanced nonlinear techniques such as neural-network autoencoder (Bourlard and Kamp 1988; Hinton and Zemel 1993) have also shown promising results in fluid mechanics applications as they require less number of modes to represent strongly nonlinear dynamics (Milano and Koumoutsakos 2002; Murata et al. 2020; Nony et al. 2023b).

Regression models Basic interpolation may fail to predict latent space components (Brunton and Kutz 2019), but machine learning offers a wide variety of algorithms to solve this problem such as polynomial chaos expansion (PCE) (Wiener 1938), Gaussian process regression (GPR) (Rasmussen et al. 2006), decision trees (Hastie et al. 2009), and neural networks (Specht 1991). The choice of best-suited method depends on the context, on the available data and also on the explicability of the predictions. For instance, GPR models provide probabilistic predictions, while PCE coefficients can be directly used for sensitivity analysis (Sudret 2008), which makes them natural candidates for risk assessment applications as well as for dealing with the uncertainty problem introduced in Sect I.2.2.

Applications to microscale atmospheric flows Such reduced-order model has been used to learn the dependence of CFD wind and dispersion predictions in the urban canopy layer over a wide range of parameters, including meteorological boundary conditions parameters and/or pollutant source parameters (García-Sánchez et al. 2014, 2017; Margheri and Sagaut 2016; Xiao et al. 2019; Lamberti and Gorlé 2021). In particular, Nony (2023) compares numerous types of models to emulate an LES response surface in a canonical dispersion case (i.e. boundary-layer flow interacting with a surface-mounted obstacle) including POD and autoencoders for the reduction step, and GPR, PCE, and decision trees for the regression step, thus providing guidelines for this specific context. One of the main takeaways from Nony's work (2023) is that taking into account the pollutant source position is a difficult task for model reduction that requires a large learning base and a high latent space dimension, especially for POD-based strategies.

These models can then be integrated into a more complex framework for uncertainty quantification (García-Sánchez et al. 2014, 2017), and uncertainty reduction through data assimilation (Mons et al. 2017; Sousa et al. 2018; Sousa and Gorlé 2019). For instance,

García-Sánchez et al. (2014) and García-Sanchez and Gorlé (2018) use PCE to emulate RANS predictions and study how uncertainties on inlet wind velocity and direction as well as ground surface roughness length propagate on flow and pollutant dispersion quantities in downtown Oklahoma City. The same model reduction approach is then used by Sousa et al. (2018) and Sousa and Gorlé (2019) to estimate wind direction and velocity using an ensemble-based data assimilation method (see Sect. I.3.2). Note that the reduced-order model they use does not include the dimension reduction step (i.e. predictions are made at each point of the domain independently), and thereby does not benefit from correlation structures in the fields and could be improved in terms of efficiency. Mons et al. (2017) go one step further, using an advanced POD–GPRs reduced-order model in a global sensitivity analysis to define optimal sensor placement strategies for data assimilation.

These studies fully motivate the use of reduced-order models when the objective is to take into account the uncertainties of microscale CFD dispersion models, in addition to reducing their computational cost.

I.3.2 Data assimilation: harnessing observations to improve predictions

Data assimilation (DA) refers to a series of methods for combining observational data with numerical models of complex systems along their uncertainties in order to improve their predictions, as schematically outlined in Fig. I.10. DA originated in weather forecasting as a response to the challenge posed by the inherently chaotic nature of weather, which inevitably leads to divergent model predictions. Within this context, observational data represent a means of rectifying the model trajectory, but they are often noisy and sparse in time and space. The main challenge of DA therefore lies in the optimal and efficient combination of numerical models and observations, two complementary yet inherently uncertain sources of information. Over time, DA has become a research field in its own right, making it possible to reduce uncertainty on numerical model predictions in a wide range of application fields. It is essential in weather and oceanography forecasting, where it is employed in operational forecast services (Rabier 2005; Martin et al. 2015). It also finds application in land surface sciences for improving environmental monitoring such as the estimation of snow cover (Largeron et al. 2020). Furthermore, DA is used in climate science for producing reanalyses of past climates based on historical observations enhanced by contemporary meteorological models (Hersbach et al. 2020).

As mentioned by Asch et al. (2016), DA is a specific case of an inverse problem, which is defined as the general problem of determining parameters, inputs or structures of a mathematical model that would best explain the observed data (Tarantola 2005). Inverse problems are encountered in a wider range of scientific and engineering applications, including for example control theory, medical imaging, and seismology. To bring together DA and inverse problem techniques under the same formalism, it is convenient to introduce the control vector, which is defined as the vector of the estimation targets, i.e. the variables that are inferred using observations.

In historical DA applications, the control vector corresponds to the state of the system under study, which is used as the initial condition for the numerical model (Fig. I.10),

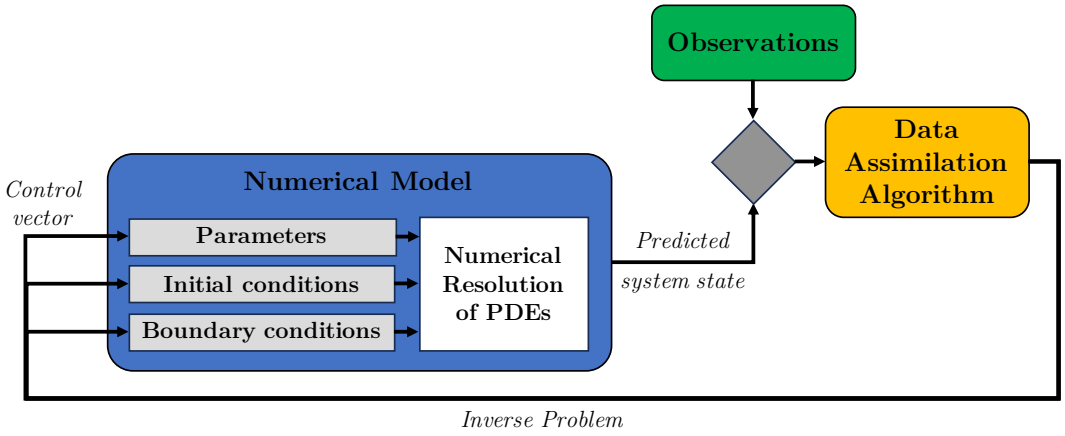


Figure I.10: *General DA framework, adapted from Rochoux et al. (2014b).*

for example, global temperature and wind speed fields in numerical weather prediction. This is known as DA for state estimation. In contrast, we speak of parameter estimation, when the control vector corresponds to model input parameters such as pollutant source position or subgrid process parameters. Recently, DA methods have been developed to jointly estimate state and parameters (Evensen 2009; Smith et al. 2013).

Two types of DA methods are classically employed:

- **Variational methods** formulates DA as an optimization problem, where the goal is to minimize the difference between model predictions and observations by adjusting the control vector. 3D–Var and 4D–Var are the most common variational methods (Asch et al. 2016).

- **Statistical methods** seek the most likely probability distribution of the control vector given the observations and a priori knowledge of the system (background), both of which are also viewed probabilistically. Among statistical methods, ensemble-based methods, such as the particle filter (Gordon et al. 1993), and the ensemble Kalman filter (EnKF) (Evensen 1994, 2003; Houtekamer and Mitchell 1998), are very popular for their ability to either deal with model non-linearity and very large control vector dimension. These methods represent probability distributions through random sampling and run multiple model simulations to represent model uncertainty. These ensemble members are then weighted based on their agreement with observations to generate the updated control vector estimate.

Variational methods offer optimal estimates but may be computationally expensive and sensitive to model errors. In contrast, statistical methods, and especially ensemble-based approaches, are easy to implement, highly parallelizable, and more robust to model errors, but can suffer from sampling errors and require careful consideration of ensemble size and quality. As they provide probabilistic estimates, statistical methods are well-suited for uncertainty quantification and ensemble forecasting. Quite recently, hybrid ensemble-variational methods have been developed to combine the advantages of both approaches (Hamill and Snyder 2000; Liu et al. 2008; Bocquet and Sakov 2014). These methods are

now substituting standard DA methods in operational weather forecast systems (Raynaud et al. 2011; Bonavita et al. 2011; Buehner et al. 2015). For a more detailed overview of the existing DA methods, we refer to Asch et al. (2016) and Carrassi et al. (2018).

To sum up, using observational data, DA enables reducing the uncertainty on the control vector, possibly correcting prior errors and improving the likelihood of the model predictions (Fig. I.10). From this perspective, DA methods seem tailored to address the problem of uncertainty in microscale atmospheric CFD models mentioned in Sect. I.2.2. In the following, we present an overview of DA studies in the specific context of microscale atmospheric dispersion and CFD modeling.

Data assimilation to improve atmospheric CFD models In contrast to fields like meteorology and oceanography, the use of DA techniques to enhance microscale or local-scale atmospheric dispersion predictions is still in its early stages. In our opinion, this is explained by the fact that the community's efforts were initially focused on building and improving CFD models, and the idea of adopting a probabilistic representation due to the involved uncertainties is relatively recent (García-Sanchez and Gorlé 2018; Dauxois et al. 2021). In addition, the cost of CFD models and the scarcity of reliable observational data at the microscale represent barriers that need to be overcome.

Nevertheless, some recent studies have demonstrated the full potential of DA to improve microscale atmospheric CFD models (Mons et al. 2017; Sousa et al. 2018; Sousa and Gorlé 2019; Aristodemou et al. 2019; Defforge et al. 2019, 2021; Bauweraerts and Meyers 2021). A cross-analysis of the main studies we have identified in the literature is presented in Chapter V. This analysis highlights the importance of the control vector definition. In particular, Defforge (2019) shows that, given the typical temporal and length scales involved at the microscale, it is more relevant to correct wind boundary conditions than initial wind conditions. This is verified by the study of Aristodemou et al. (2019), which shows very limited persistence of state corrections in an LES dispersion model.

These studies also show that it can be relevant to use reduction techniques to address the issues of model cost and/or large field size. For instance, Sousa et al. (2018) use a reduced-order model to substitute the CFD model, while Aristodemou et al. (2019) and Bauweraerts and Meyers (2021) use techniques similar to POD to reduce the state dimension. Another point that emerges from these studies is that sensor location has a critical effect on DA accuracy (Mons et al. 2017; Sousa et al. 2018). In particular, positioning sensors in areas of the urban canopy layer where the flow is mostly determined by surrounding buildings can lead to ill-posed problems, as it does not contain much information on the large-scale atmospheric condition. Finally, we have identified that these works could be improved when it comes to the modeling of errors. Observation and prior errors are often arbitrarily chosen and may underestimate the impact of internal variability, while model error is most of the time simply ignored due to a lack of knowledge.

Data assimilation for source estimation The problem of determining the location and characteristics of a pollutant release source based on observed data of pollutant concentrations is another typical inverse problem. This is of particular importance in emergency response to accidental release. It is for example used to estimate radionuclide

release (Winiarek et al. 2012; Saunier et al. 2019; Dumont Le Brazidec et al. 2023) or to retrieve volcanic ashes properties (Francis et al. 2012). At the microscale, Keats et al. (2007) retrieve source location and release rate using measurements from the MUST and JOINT URBAN 2003 field campaigns.

The source estimation problem is inherently ill-posed, as multiple sources or combinations of parameters can yield similar concentration patterns. Markov chain Monte Carlo is one of the most popular approaches for solving this complex problem (Gilks et al. 1995). For a detailed overview of source estimation methods, we refer the reader to Hutchinson et al. (2017). We also note that Allen et al. (2007) and Lucas et al. (2016) demonstrate the benefit of jointly estimating source parameters and wind conditions to reduce the uncertainty on the source location.

Other data assimilation studies involving CFD models It is important to bear in mind that microscale atmospheric flows represent only a small fraction of the flows for which CFD models are used, implying that DA methods can be applied in other application contexts. For instance, DA systems have been developed to improve CFD model predictions in aeronautics (Misaka et al. 2008; Kato et al. 2015), combustion (Jahn et al. 2012; Gao et al. 2017; Labahn et al. 2019; Wang et al. 2020b), or safety engineering for the prediction of gas leakage in terminal and utility tunnel (Wu et al. 2021; Cai et al. 2022). An interesting application for microscale simulations is the reduction of uncertainties in turbulence modeling, for example, Xiao et al. (2016) use an EnKF for the assessment and mitigation of turbulence model uncertainties in RANS simulations of flow over periodic hills. It is also possible to go further by using DA to jointly correct the state of the flow alongside turbulence and/or boundary condition parameters (Gronskis et al. 2013; Kato et al. 2015; Kumar et al. 2019).

I.3.3 Design of an efficient data assimilation system

The general objective of this thesis is to design and validate a reduced-cost DA system that: i) uses in-situ observational data to assess and reduce uncertainty in microscale CFD dispersion predictions; and ii) relies on model reduction techniques to provide probabilistic predictions at almost no computational cost.

We focus the DA estimation problem on the correction of the boundary condition parameters that account for large-scale meteorological conditions. Meteorological boundary conditions are indeed often uncertain because of the scarcity of measurements available to define them, and because of the use of measurements from remote weather stations that can introduce representativeness errors (Pimont et al. 2017). Even in field experiments, where considerable instrumentation efforts can be made, meteorological conditions remain uncertain due to the internal variability of the ABL (Schatzmann and Leitl 2011). This uncertainty is all the more problematic as microscale CFD models are highly sensitive to changes in boundary conditions (García-Sánchez et al. 2014; Lucas et al. 2016; Wise et al. 2018). For this reason, using DA to reduce uncertainty on the large-scale meteorological boundary conditions has been shown to significantly improve the accuracy of microscale CFD models (Mons et al. 2017; Sousa and Gorlé 2019; Defforge et al. 2021).

Furthermore, we employ a reduced-order model as a surrogate for the microscale CFD model in the DA system. This approach is a way to drastically reduce the computational cost of the DA process (Rochoux et al. 2014a; Sousa et al. 2018; Sousa and Gorlé 2019). The reduced-order model is trained in an offline step to emulate the fields of interest for varying meteorological boundary conditions from a database of precomputed CFD simulations and then used online to accelerate DA. This surrogate modeling approach enables the DA system to be used in real time and combines particularly well with ensemble-based DA algorithms to provide probabilistic predictions that represent the uncertainties involved. The counterpart is that it introduces a new source of error in the DA scheme. Rochoux et al. (2014a) show that, given limited model reduction error, using a reduced-order model does not impair DA performance, and may even improve DA accuracy by allowing more model predictions. This is also valuable to carry out a large number of trials to adjust and optimize the DA system, which is particularly relevant in this thesis, as our DA system is implemented from A to Z.

The architecture of the reduced-cost DA system designed for this thesis is presented in Fig. I.10. Thanks to in-situ measurements, an ensemble-based DA algorithm is used to correct and reduce the uncertainty on prior large-scale meteorological boundary condition parameters, such as wind direction and velocity. The inferred parameters can then be used as inputs for the reduced-order model in order to provide a corrected prediction of the quantity of interest, such as the pollutant concentration levels in an urban environment, along with the associated uncertainty. A more exhaustive specification of the DA system is given in Chapter V.

We highlight that the DA system presented in Fig. I.11 can reduce errors on the boundary conditions induced by the mesoscale variability of the ABL. However, in contexts where data are acquired over limited periods, this system cannot address the aleatory uncertainty due to microscale internal variability associated with the smaller eddies of

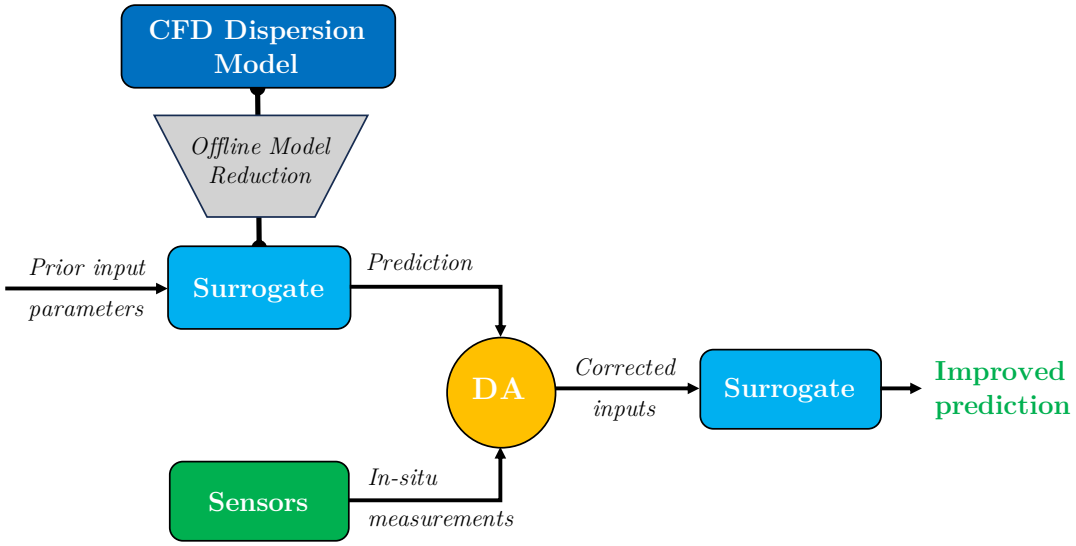


Figure I.11: *The data-driven modeling system designed to address the problem of this thesis. It is a specification of the basic system shown in Fig. I.10. To reduce computational cost, a reduced-order model is used as a surrogate of the CFD dispersion model. Model predictions associated with uncertain prior of the model input parameters are compared to in-situ measurements of the system state using a DA algorithm. This algorithm corrects input parameters, from which the reduced-order model can predict an improved estimation of the system state. Uncertainty on prior knowledge, measurements and model predictions are explicitly modeled and taken into account to ensure the robustness of DA estimates.*

the ABL, as it is intrinsically irreducible. To our knowledge, this source of error has not yet been explored by this kind of DA system, although it may significantly impact both model and measurement accuracy. In our opinion, incorporating this uncertainty in the DA system would considerably improve its robustness by preventing it from giving too much confidence in observations and model predictions, and the representativeness of its uncertainty estimation. Hence, this thesis is a first attempt to realistically account for this aleatory uncertainty in a DA system for microscale dispersion. This also motivates the choice of an LES approach for the CFD model because of its ability to represent the microscale internal variability of the ABL, whereas this important information is not accessible with the averaged formalism of the RANS approach.

To sum up, the main objectives of the thesis are to:

1. build the proposed reduced-cost DA system and verify its ability to make fast estimations of microscale pollutant dispersion based on LES while reducing meteorological boundary condition uncertainty,

2. take into account the observational and prediction uncertainty induced by the microscale internal variability of the ABL in this DA system, to ensure its robustness,

3. realistically represent the effect of the main uncertainties involved in the final pollutant dispersion estimations, in order to go beyond standard point-wise predictions which are intrinsically unsuited to the microscale atmospheric context.

I.3.4 Implementation strategy

To build and validate the reduced-cost DA system presented in Fig. I.11, which is the backbone of this thesis, we follow a four-stage action plan including all necessary new developments.

i) Set up an LES model of a real field dispersion experiment

Among the different dispersion field experiments presented in Sect. I.1.4, we choose to reproduce one trial of the MUST field experiment (Biltoft 2001) corresponding to neutral stratification conditions, using an LES modeling approach. This choice is also motivated by the internal variability problem, as it is shown to be particularly significant in this experiment (Schatzmann et al., 2011, p. 88–101). In addition, a large number of studies comparing CFD model predictions with experimental measurements are already reported in the literature and can be used as a basis for comparison.

ii) Estimate the LES model main uncertainties

We explore the main sources of uncertainty in microscale LES dispersion prediction, in particular parameter perturbation tests are carried out to investigate the model sensitivity to boundary condition parameters, but also part of the model structural uncertainties. A significant effort is made to estimate the effect of microscale internal variability not only on LES predictions but also on observations. To do so, we design and propose a bootstrap approach. The LES model is then validated in a robust way, by taking these uncertainties into account when comparing its predictions to experimental data. Source and geometry uncertainties are not studied in this thesis, considering that they are well-controlled in the MUST field campaign.

iii) Build a reduced-order model of the LES model

First, we generate a large ensemble of 200 LES predictions to represent different possible scenarios of meteorological boundary conditions, sampling the parameters to which the model is most sensitive. The LES mean concentration response surface is then emulated using a reduced-order modeling approach called POD–GPRs, which relies on proper orthogonal decomposition (POD) for the reduction step, and Gaussian process regressors (GPRs) for the regression step. This approach has been successfully applied to a simplified 2–D dispersion case by Nony (2023). We extend it to the MUST real case, which requires adaptations to deal with the very large field dimension, and with the wide range of pollutant concentration levels involved. In addition, we use our estimation of the aleatory uncertainty related to microscale internal variability to make an informed choice on the number of reduced-basis modes used by the POD–GPRs model, in order to avoid overfitting noisy structures and thus minimize model reduction errors.

iv) Assemble and validate the reduced-cost DA system

Finally, we assemble the reduced-cost DA system presented in Fig. I.11. To answer the need for probabilistic prediction, the standard EnKF is chosen as the DA

algorithm. We assess its ability to correct meteorological boundary condition parameters using local mean concentration measurements. The main innovation of our approach is that we propose advanced and realistic error models for i) the prior boundary condition parameters, thanks to a microclimatology study involving a large amount of observational data; ii) the observation and model predictions, by accounting for the aleatory uncertainty due to the microscale internal variability of the ABL thanks to our bootstrap approach; and iii) the reduced-order model predictions compared to the ones of the LES model. In our opinion, this is significant progress towards more realistic and robust DA systems for microscale atmospheric applications. Finally, the DA system thus obtained is first set up by assimilating synthetic observations (OSSE experiments), then validated using the real tracer concentration measurements from the MUST field campaign.

I.3.5 Structure of the manuscript

A thesis chapter is devoted to each of the four main stages previously introduced. Figure I.12 gives a schematic view of the general outline of the manuscript. This figure also illustrates the dependency structure between chapters.

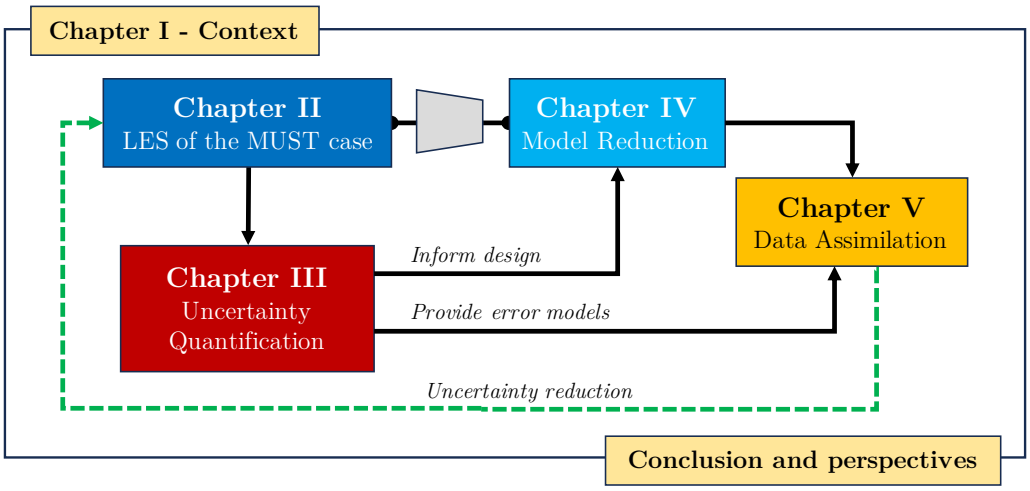


Figure I.12: *Overview of the manuscript structure. Each block corresponds to a thesis chapter. Black arrows represent dependencies. The green dashed line highlights the general problematic of the thesis.*

The LES microscale atmospheric dispersion model, which is the starting point for all other work, is presented in Chapter II. Robust validation of this model is then carried out in Chapter III by taking into account its uncertainties and confronting its flow and dispersion predictions with experimental measurements. This also helps determine which model inputs have the most effect on the predictions and hence should be prioritized for uncertainty reduction. Based on this knowledge, the POD–GPRs reduced-order model is built and validated as a substitute for the LES model in Chapter IV. Once all these prerequisites are in place, we demonstrate in Chapter V the ability of the data-driven modeling system designed in Sect. I.3 to answer the general problematic of the thesis. Finally, the last part of the manuscript summarizes our contributions and opens up interesting perspectives.

Thesis-related publications

- Part of the content of Chapters II and III is the subject of an article published in the journal Boundary-Layer Meteorology (Lumet et al. (2024). *Assessing the internal variability of large-eddy simulations for microscale pollutant dispersion prediction in an idealized urban environment*).
- Preparatory works for Chapters IV and V were presented during the 21st International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (HARMO 21) in Aveiro (Portugal) in September 2022.

Chapter II

Microscale atmospheric dispersion large-eddy simulation

This chapter presents the large-eddy simulation (LES) model used in this thesis to predict microscale pollutant dispersion, and how it is implemented to reproduce one trial of the Mock Urban Setting Test (MUST) field campaign experiment (Biltoft 2001).

We first introduce the LES theoretical framework, with a particular focus on microscale atmospheric flow and dispersion. We also explain how the chosen solver, originally developed for reactive and compressible flows, is adapted to efficiently simulate atmospheric flows.

Next, we present the MUST pollutant dispersion field campaign experiment chosen to validate the LES model. We briefly describe the array of shipping containers used to mimic a simplified urban canopy, before providing an overview of the main sensors installed during the experiment. A specific pollutant release trial, the #2681829 which corresponds to neutral atmospheric conditions, is chosen as the case study for this thesis.

The construction and adaptation of the LES model for this specific trial of the MUST campaign is then presented. Proper definitions of the computational domain and boundary and initial conditions are given. Particular emphasis is placed on the implementation of turbulence injection to realistically reproduce the fluctuations of the atmospheric boundary layer (ABL), as this is one of the new contributions of this thesis to the model.

In the last part of the chapter, a preliminary study in a free field configuration, i.e. without obstacles, is carried out to investigate the consistency between the inflow conditions and the wall law imposed as the ground boundary condition, as well as the development of the turbulent fluctuations injected at the inlet.

Chapter outline

II.1 Introduction 47
II.2 Large-eddy simulation dispersion modeling 50
II.2.1 Large-eddy simulation principle 50
II.2.2 LES-filtered governing equations 51
II.2.3 Subgrid-scale modeling 52
II.2.4 Solver features and specifications 54
II.2.4.1 Numerical scheme 55
II.2.4.2 Pressure gradient scaling to improve computational efficiency 55
II.2.4.3 Conversion of species mass fraction into concentration 56
II.3 The Mock Urban Setting Test field campaign 57
II.3.1 Experimental site description 57
II.3.2 Available experimental data 57
II.3.3 Selected case 60
II.4 Large-eddy simulation model of the MUST trial 2681829 . 61
II.4.1 Computational domain and spatial discretization 61
II.4.2 Boundary conditions 62
II.4.2.1 Inflow boundary conditions 62
II.4.2.2 Wall boundary conditions 63
II.4.2.3 Other boundary conditions 64
II.4.3 Turbulence injection method in the context of ABL flow 64
II.4.4 Tracer source modeling 67
II.4.5 Initial condition and spin-up time definition 67
II.5 Preliminary verification: simulation of a free-field case . . . 70
II.6 Summary 74

II.1 Introduction

The aim of this section is to introduce the numerical model used for the prediction of microscale pollutant dispersion throughout the thesis. In particular, we discuss in detail the choices made in terms of CFD approach, solver, and validation case study.

Motivation of the LES modeling approach In this thesis, we focus on microscale CFD dispersion models because of their ability to explicitly account for the complex interactions between the built environment and atmospheric flow and dispersion. From the main CFD approaches introduced in Sect. I.1.3, only the RANS and LES approaches are affordable at the neighborhood scale, which takes into account ensembles of buildings in order to be representative of the urban canopy layer in a real city. In this thesis, we adopt the LES approach because it provides instantaneous realizations of the most energetic atmospheric eddies, and can thereby reproduce the effect of the microscale internal variability of the ABL on the model predictions. This important knowledge, which is central to the objectives of this thesis, is not accessible with the average formalism of the RANS approach. LES instantaneous realizations can also be directly used to estimate the concentration statistics such as peak concentrations and fluctuations, which are of importance for pollutant dispersion studies (Tominaga and Stathopoulos 2013), whereas RANS requires additional modeling assumptions to estimate these statistics (Milliez 2006). In addition, LES is expected to reduce the uncertainties associated with turbulence modeling compared with RANS (Gousseau et al. 2011; García-Sanchez et al. 2018), allowing us to focus on other forms of uncertainty, and in particular those associated with large-scale meteorological conditions. The downside is that LES has a significantly higher computational cost than RANS as discussed in Sect. I.2.1. However, this drawback will be mitigated by the reduced-order model developed in Chapter IV.

The AVBP solver choice In this thesis, we use the AVBP[1] solver developed by CERFACS to perform LES of compressible, reactive, turbulent flows on massively parallel architectures. It is a reference code in the High-Performance Computing (HPC) community and it is widely used in the field of combustion, both for research purposes and as a design tool for industry (Vermorel et al. 2017; Pérez Arroyo et al. 2020). It is also used, to a lesser extent, to resolve non-reacting flows to study turbomachinery (Wang et al. 2014), wind turbine (Dabas et al. 2022), and pollutant formation and dispersion (Poubeau et al. 2016; Paoli et al. 2020).

This choice of solver is motivated by its high scalability and unstructured capability, which simplifies the generation of meshes representing the urban canopy. In the longer term, AVBP offers interesting prospects, such as the modeling of reactive phenomena and realistic sources. This choice therefore opens the door to a broader safety environment perspective, with applications in modeling industrial fires, wildland fires, and hydrogen or liquified gas leaks.

As AVBP was not initially developed to solve environmental fluid flows and plume

[1]See documentation at http://www.cerfacs.fr/avbp7x/.

dispersion in open areas, one of the contributions of this thesis is to validate AVBP in this kind of context. To do so, we draw on model development work initiated by Rochoux et al. (2021). The adaptation of the compressible formulation of the Navier-Stokes equations used in AVBP to atmospheric flow is detailed in Sect. II.2.4. A particular development carried out in this thesis is the addition of a turbulence injection method to improve the representativeness of the inlet wind profile (Sect. II.4.3).

Choice of validation case study Among the different dispersion validation experiments presented in Sect. I.1.4, we choose the MUST field campaign (Biltoft 2001) as the case study for this thesis. MUST is an attractive test case to assess the accuracy and reliability of microscale CFD dispersion models because i) it features an idealized urban canopy made up of a regular array of shipping containers, simplifying the computational grid construction; ii) the experimental test site is isolated, allowing the atmospheric boundary layer to develop properly, which simplifies the definition of meteorological boundary conditions for the model; iii) the pollutant species used in the experiment, propylene, can be considered as passive chemical species and is released passively (Biltoft 1995, 2001), which simplifies its modeling; and iv) a large number of observations of wind, turbulence and tracer concentration are available at different locations throughout the field. This allows good control of the meteorological boundary conditions of the model and offers the possibility of in-depth validation. More details on the experimental field campaign are given in Sect.II.3.

MUST was selected as one of the reference case studies for the European COST[2] Action 732 on the assurance of the quality of microscale dispersion models (Franke et al. 2007). In this exercise, including a dozen CFD models, measurements from the wind tunnel reproduction of the MUST experiment from Bezpalcová (2007) are used as validation data instead of the field campaign data. This choice is motivated by the fact that wind tunnels enable validation data to be acquired over very long acquisition windows, thus avoiding flaws in the model validation and comparison because of the internal variability of the ABL (Schatzmann et al. 2010). In this thesis, we propose not to bypass the problem of internal variability and to actually take it into account in the validation. The approach developed to quantify the effect of internal variability will be presented in Chapter III.

A large number of studies comparing the predictions of RANS and LES models with experimental measurements from the MUST field campaign are reported in the literature (see Table II.1). These studies can be used as a basis for building and evaluating our LES model. Overall, these studies show acceptable agreement between CFD models and experimental observation, but local differences between models and observations remain and are not fully explained (Milliez and Carissimo 2007; König 2014; Nagel et al. 2022). The choice of this case study is therefore relevant to the overall objective of this thesis, which is to quantify and reduce uncertainties in microscale CFD models.

[2]COST (European Cooperation in Science and Technology) is an intergovernmental European organization that promotes research and innovation networks. See `https://www.cost.eu/`.

Table II.1: *Overview of the CFD modeling studies of the MUST field campaign. CFD studies that use data from the wind tunnel experiment from Bezpalcová (2007) for validation are not reported in this table. For an overview of the models used in COST Action 732, please refer to Olesen et al. (2008) and Schatzmann et al. (2010). The blue color indicates the studies that reproduce the trial # 2681829.*

Study	CFD approach	Solver	Trial(s)
Hanna et al. (2004)	RANS	FLACS	37 trials
Camelli et al. (2005)	VLES[1]	FEFLO–URBAN	# 2682353
Hsieh et al. (2007)[2]	RANS	STREAM	/
Milliez and Carissimo (2007)	RANS	Code–Saturne	20/21 trials[3]
Donnelly et al. (2009)	RANS	WinMISKAM	19/21 trials[3]
Efthimiou et al. (2011)	RANS	ADREA	# 2682256
Antonioni et al. (2012)	RANS/LES	FLUENT	# 2682353
König (2014)	LES	ASAM	# 268182 and # 2682353
Kumar et al. (2015)	RANS	Fluidyn-PANACHE	20/21 trials[3]
Bahlali et al. (2019)	RANS	Code–Saturne	# 2681829 and # 2692157
Nagel et al. (2022)	LES	Meso-NH	# 2681829

[1] Very Large Eddy Simulation.
[2] Validation with data from the water channel modeling of MUST by Yee et al. (2006).
[3] Among the 21 trials that were selected by Yee and Biltoft (2004) for their high quality.

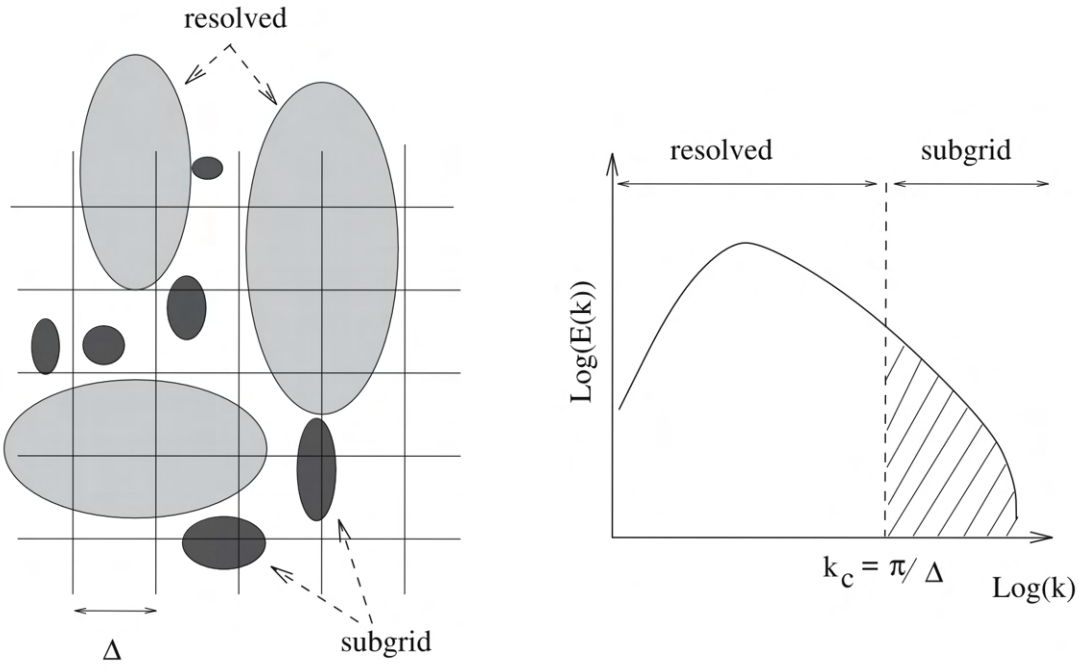
This chapter is limited to the presentation of the case study and of the implementation of the LES model used to reproduce it. The LES model will be validated in Chapter III, in light of the various uncertainties involved.

II.2 Large-eddy simulation dispersion modeling

In this section, we present the LES theoretical basis, and how it can be used to solve the equations governing atmospheric flows and pollutant transport at microscale. We also describe in more detail how the AVBP solver is adapted to solve these flows efficiently. This section has been synthesized based on the lecture notes from Piomelli (2020), and on Sagaut's book (2005).

II.2.1 Large-eddy simulation principle

LES is a numerical approach for simulating flows governed by the Navier-Stokes equations. It consists of explicitly solving the larger-scale turbulence and modeling the effect of smaller scale. This makes it possible to use coarser discretization than with direct numerical simulations (DNS), which resolve all turbulence scales, and thereby reduce simulation computational cost.



(a) *Physical space* **(b)** *Fourier space*

Figure II.1: *Large-Eddy Simulation implicit filtering principle illustrated. The filter is defined by the mesh resolution* Δ *(a): the eddies larger than* Δ *are resolved while the effect of the smaller eddies is taken into account by the subgrid-scale model. In the Fourier space (b), the separation of the resolved scales from the subgrid scales, at the cutoff wave number* k_c*, is illustrated for the turbulent energy spectrum associated with a homogeneous isotropic turbulent field. Figure excerpt from Sagaut (2005).*

In contrast to the RANS (Reynolds-Averaged Navier-Stokes) approach, LES provides

a representation of the flow field with spatio-temporal resolution of the largest scales of turbulent fluctuation. This limits turbulence modeling errors as they only concern the small-scale subgrid motions provided that a sufficient fraction of the turbulent kinetic energy is explicitly resolved, usually at least 80% (Pope 2000). Most of the time this separation between the resolved and modeled scales is implicitly achieved thanks to the grid used to discretize the Navier-Stokes equations as illustrated in Fig. II.1. Note that the LES tends towards explicit DNS resolution when the filter size tends towards the smallest scale of turbulence in the flow: the Kolmogorov scale.

II.2.2 LES-filtered governing equations

In this section, we explain how the governing equations for atmospheric flow and pollutant dispersion at the microscale introduced in Sect.I.1.2 are filtered with the LES paradigm. Since our focus in this thesis is on reproducing a specific MUST test carried out under neutral atmospheric stratification conditions, we present the simplified equations for neutral conditions.

The separation between the larger scales and the smaller scales involved in LES can be conceptualized as a filter of a given characteristic length Δ. In this thesis, we only use implicit filtering which means that the cut-off scale of the filter is of the order of the spatial resolution of the grid as shown in Fig. II.1. Applying this filtering operation to the incompressible Navier-Stokes equations using Boussinesq approximation (Eqs. I.7) and under neutral stratification conditions, as well as to the passive species transport equation (Eq. I.8) gives:

$$\begin{cases} \dfrac{\partial \tilde{u}_i}{\partial x_i} = 0, & \text{(II.1a)} \\ \dfrac{\partial \tilde{u}_i}{\partial t} + \tilde{u}_j \dfrac{\partial \tilde{u}_i}{\partial \tilde{x}_j} = -\dfrac{1}{\rho_0}\dfrac{\partial \widetilde{p'}}{\partial x_i} + \dfrac{1}{\rho_0}\dfrac{\partial}{\partial x_j}\left(\widetilde{\tau}_{ij} + \widetilde{\tau}^t_{ij}\right) & \text{(II.1b)} \\ \dfrac{\partial \tilde{c}}{\partial t} + \tilde{u}_j \dfrac{\partial \tilde{c}}{\partial x_j} = D\dfrac{\partial^2 \tilde{c}}{\partial {x_j}^2} - \dfrac{\partial \widetilde{J}^t_i}{\partial x_j} + S, & \text{(II.1c)} \end{cases}$$

where $\tilde{\mathbf{u}}$, $\widetilde{p'}$ and $\tilde{c}$ are the filtered velocity, pressure variation and species concentration fields respectively, and where $\widetilde{\tau}_{ij}$ is the filtered viscous stress tensor, defined as:

$$\widetilde{\tau}_{ij} = 2\rho_0 \nu \widetilde{S}_{ij}, \tag{II.2}$$

with $\widetilde{S}_{ij} = \dfrac{1}{2}\left(\dfrac{\partial \widetilde{u}_i}{\partial x_j} + \dfrac{\partial \widetilde{u}_j}{\partial x_i}\right)$ the resolved large-scales strain-rate tensor. In addition, the filtering brings out two new terms compared with the original equations: the subgrid-scale (SGS) turbulent stress tensor $\widetilde{\tau}^t_{ij}$ and the SGS diffusive flux vector of the transported species $\widetilde{J}^t_i$. These terms account for the effect of unresolved scales on flow and species transport and are defined as follows

$$\widetilde{\tau}^t_{ij} = -\rho_0 \left(\widetilde{u_i u_j} - \widetilde{u}_i \widetilde{u}_j\right), \tag{II.3a}$$

$$\widetilde{J}^t_i = \widetilde{u_i c} - \widetilde{u}_i \widetilde{c}, \tag{II.3b}$$

Because of the filtered product terms in Eq. II.3, such as $\widetilde{u_i}\widetilde{u_j}$, that cannot be obtained from the resolved filtered quantities $\widetilde{u_i}$, $\widetilde{p}$, and $\widetilde{c}$, additional closure equations are required to solve the system (Eq. II.1). These closure equations are of primary importance for LES because they are the ones that model the effects of the subgrid turbulent scales (see Fig II.1). The principle of subgrid-scale modeling and some fundamental models are presented in the next section.

Note that some authors consider that it is more rigorous to call $\widetilde{\tau}_{ij}^t$, and $\widetilde{J}_i^{\,t}$ the *subfilter-scale* turbulent terms (Pope 2000; Piomelli 2020), but we decided to stick with the *subgrid-scale* terminology because it is more common, and because we only use the filter implicitly defined by the discretization grid.

II.2.3 Subgrid-scale modeling

A wide range of SGS models for the terms (Eq. II.3) have been developed to adequately reproduce the effect of filtered-scale turbulence in LES. They can have various levels of complexity from a scalar equation with one constant to a dynamic procedure that adapts the subgrid-scale contribution in space and in time. In this section, we present the historic Smagorinsky-Lilly model (Smagorinsky 1963; Lilly 1967) as well as the dynamic Smagorinsky model (Germano et al. 1991) and wall-adaptative local eddy-viscosity model (Nicoud and Ducros 1999) which are the two SGS models mainly used in this thesis for the SGS turbulent stress tensor $\widetilde{\tau}_{ij}^t$ (Eq. II.3a). We also present the modeling strategy adopted for $\widetilde{J}_i^{\,t}$ (Eq. II.3b).

The Smagorinsky-Lilly model from Smagorinsky (1963) and Lilly (1967) is one of the most common subgrid-scale models because of its seniority, simplicity and ability to model correctly the action of the unresolved scales. It is an eddy-viscosity model whose principle is to model the subgrid-scale turbulent stress tensor analogously to the viscous stress tensor (Eq. II.2) by introducing a subgrid-scale turbulent viscosity ν_t:

$$\widetilde{\tau}_{ij}^t - \frac{\delta_{ij}}{3}\widetilde{\tau}_{kk}^t = 2\rho_0\nu_t\widetilde{S}_{ij}. \tag{II.4}$$

To ensure closure, the Smagorinsky model uses one algebraic equation for the subgrid-scale turbulent viscosity:

$$\nu_t = (C_s\Delta)^2|\widetilde{S}|, \tag{II.5}$$

with Δ the filter cutoff length, C_s the Smagorinsky constant and $|\widetilde{S}| = \sqrt{2\widetilde{S}_{ij}\widetilde{S}_{ij}}$ the magnitude of the resolved strain-rate tensor. A relationship between this constant and the Kolmogorov constant C_κ has been derived by Lilly (1967), for $C_\kappa = 1.41$ it gives $C_s = 0.18$. In the solver we use, the filter scale Δ is taken equal to the cubic root of nodal volume[3].

[3]The nodal volume corresponds to the volume associated with the node in the dual graph of the original mesh.

This model accounts for the dissipation of energy from the subgrid scales as it always predicts $\nu_t \geq 0$. Nevertheless, it is well suited for isotropic turbulent flows in which the turbulent cascade energy is well verified. Deardorff et al. (1970) showed that in the presence of shear and thus especially around obstacles the C_s constant must be reduced. In addition, this model should be adjusted when the grid size is not homogeneous, for instance when refining towards solid boundaries (Piomelli 2020). This is usually done by applying the van Driest damping function (van Driest 1956). Finally, it is known that it can be too dissipative (Sagaut 2005). For these reasons, we avoid using the Smagorinsky-Lilly model for ABL flows.

The dynamic Smagorinsky model from Germano et al. (1991) is a more accurate version of the standard Smagorinsky model which requires an additional calculation step to estimate and adapt the constant C_s during the simulation. To do so, a test filter is introduced with a cutoff length scale $\hat{\Delta}$ larger than the one from the primary filter scale Δ, and C_s is expressed using the Germano inequality following Lilly's procedure (1992) as:

$$C_s^2 = \frac{1}{2}\frac{\langle L_{ij} M_{ij} \rangle}{\langle M_{ij} M_{ij} \rangle}, \tag{II.6}$$

where $L_{ij} = \widehat{\widetilde{u}_i \widetilde{u}_j} - \widehat{\widetilde{u}_i}\widehat{\widetilde{u}_j}$ and $M_{ij} = \hat{\Delta}^2 |\widehat{\widetilde{S}}| \widehat{\widetilde{S}}_{ij} - \Delta^2 \widehat{|\widetilde{S}| \widetilde{S}_{ij}}$, and where the hat designates the variables filtered with the test filter. An additional averaging denoted $\langle \cdot \rangle$, often spacewise, may be used to smooth local variations of the coefficient C_s (Piomelli 2020).

The wall-adaptive local eddy-viscosity (WALE) model introduced by Nicoud and Ducros (1999) for wall-bounded flows in an attempt to recover the scaling laws of turbulent viscosity in the near-wall region. As the Smagorinsky model, it is an eddy-viscosity model so the subgrid-tensor turbulent stress tensor is expressed as in Eq. II.4. But the subgrid turbulent viscosity ν_t is estimated by the following parametric closure equation:

$$\nu_t = (C_w \Delta)^2 \frac{\left(s_{ij}^d s_{ij}^d\right)^{3/2}}{\left(\tilde{S}_{ij}\tilde{S}_{ij}\right)^{5/2} + \left(s_{ij}^d s_{ij}^d\right)^{5/4}}, \tag{II.7}$$

where C_w is a model constant set to 0.5 (Nicoud and Ducros 1999), and s_{ij}^d is the traceless symmetric part of the square of the velocity gradient tensor:

$$s_{ij}^d = \frac{1}{2}\left(\widetilde{g}_{ij}^2 + \widetilde{g}_{ji}^2\right) - \frac{1}{3}\widetilde{g}_{kk}^2 \delta_{ij}, \text{ with } \widetilde{g}_{ij} = \frac{\partial \widetilde{u}_i}{\partial x_j}. \tag{II.8}$$

In this thesis, we chose to mainly use the WALE model because it is particularly suited for our objective of performing wall-modeled LES of ABL flows. In addition, even with a simple algebraic model for ν_t, the SGS modeling errors are expected to have a limited effect on the overall accuracy of the model as i) the subgrid-scale stress only accounts for a limited fraction of the total stress (Fig II.1), ii) the small scales tend to be more homogeneous and isotropic than the large ones, so they are more likely to be easily

represented by simple models (Piomelli 2020). This was verified in our case by comparing the WALE model with the dynamic Smagorinsky model. Results of the comparison are presented in Sect. III.5.3, page 112.

The SGS species diffusive flux model. The effect of the subgrid-scales of turbulence on the tracer transport is simply modeled using a gradient-diffusion hypothesis:

$$\widetilde{J}_i^t = -\frac{\nu_t}{\mathrm{S}_c^t}\frac{\partial \widetilde{c}}{\partial x_j}, \tag{II.9}$$

with ν_t the subgrid-scale turbulent viscosity in Eq. II.4 calculated by the SGS model used for the subgrid-scale turbulent stress tensor, and S_c^t the turbulent Schmidt number defined by analogy with the standard Schmidt number (Eq. I.13). Modeled in this way, unresolved small-scale eddies affect species transport in a way analogous to molecular diffusion (Fig. I.5a). In our simulations, we use a value of $\mathrm{S}_c^t = 0.6$ in line with solver best practice. Sensitivity tests to the value of S_c^t are carried out in Sect. III.5.3, page 112 and show very limited impact on model predictions.

II.2.4 Solver features and specifications

The solver used in this thesis, AVBP[4], solves on unstructured grids the filtered compressible Navier-Stokes equations for flow dynamics and the filtered passive transport equation (Schönfeld and Rudgyard 1999; Gicquel et al. 2011):

$$\begin{cases} \dfrac{\partial \overline{\rho}}{\partial t} + \dfrac{\partial}{\partial x_j}(\overline{\rho}\widetilde{u}_j) = 0, & \text{(II.10a)} \\ \dfrac{\partial}{\partial t}(\overline{\rho}\widetilde{u}_j) + \dfrac{\partial}{\partial x_j}(\overline{\rho}\widetilde{u}_i\widetilde{u}_j) = -\dfrac{\partial}{\partial x_j}\left[\overline{P}\delta_{ij} - \overline{\tau}_{ij} - \overline{\tau}_{ij}^t\right] - \delta_{i3}\overline{\rho}g, & \text{(II.10b)} \\ \dfrac{\partial \overline{\rho}\widetilde{E}}{\partial t} + \dfrac{\partial \overline{\rho}\widetilde{E}\widetilde{u}_j}{\partial x_j} = -\dfrac{\partial}{\partial x_j}\left[\overline{u_i(P\delta_{ij} - \tau_{ij})} + \overline{q}_j + \overline{q}_j^t\right], & \text{(II.10c)} \\ \dfrac{\partial \overline{\rho}\widetilde{w}}{\partial t} + \dfrac{\partial \overline{\rho}\widetilde{w}\widetilde{u}_i}{\partial x_i} = D\dfrac{\partial^2 \overline{\rho}\widetilde{w}}{\partial {x_j}^2} - \dfrac{\partial \overline{\rho}\overline{J}_i^t}{\partial x_i} + S, & \text{(II.10d)} \end{cases}$$

where $P = \rho RT/W$ is the pressure given by the equation of state for an ideal gas with $R = 8.314\,\mathrm{J\,mol^{-1}\,K^{-1}}$ and W the air molecular weight. E denotes the total energy per unit mass ($E = e + \frac{1}{2}u_iu_i$, with e the internal energy), q_j the heat flux vector, and w the mass fraction of the transported species. Note that in AVBP, $\overline{J}_i^t$ and S are defined in terms of mass fraction, not volume concentration. The standard filtration operation is denoted by a bar, while the tilde indicates mass-weighted filtering from (Favre 1965) ($\widetilde{f} = \overline{\rho f}/\overline{\rho}$). Compared to the filtered incompressible Navier-Stokes equations under neutral conditions (Eqs. II.1a and b), AVBP additionally solves the energy conservation equation (Eq. II.10c), which requires a new SGS model for the SGS heat flux vector $\overline{q}_j^t$.

[4]See footnote 1, page 47.

II.2.4.1 Numerical scheme

Different numerical schemes are implemented in AVBP to discretize and solve the filtered LES equations (Eq. II.10). In most of this thesis, we use the second-order Lax-Wendroff (LW) finite-volume scheme (Schönfeld and Rudgyard 1999). This explicit scheme in time, and centered scheme in space, was chosen for its reduced computational cost when compared to higher-order schemes. A comparison between LW and the two-step Taylor-Galerkin (TTGC) scheme (Colin and Rudgyard 2000) is presented in Sect. III.5.3, page 112. As a third-order in space and time, explicit scheme, TTGC reduces the numerical errors but increases the simulation computational cost by a factor of approximately 1.8 in return.

II.2.4.2 Pressure gradient scaling to improve computational efficiency

The LES code AVBP is based on a fully compressible explicit formulation of the Navier-Stokes equation. Therefore, the time step used for the time advancement of the simulation is constrained by the Courant-Friedrichs-Lewy (CFL) number condition, which must be lower than 0.9 for the stability of the explicit time integration. For the compressible formulation used in the numerical solver, this constraint reads

$$\mathrm{CFL} = \frac{\Delta t(u+c)}{\Delta x} < 0.9\,, \tag{II.11}$$

where u is the velocity norm, c is the speed of sound and Δx is the characteristic grid size. The key idea of the pressure gradient scaling (PGS) (Ramshaw et al. 1986) is to artificially reduce the speed of sound c, which reduces the constraint on the time step. Compressibility effects and acoustic propagation are modified by this transformation, but they are not relevant for ABL flows, which are in the limit of very low Mach number. In this limit, it is demonstrated that the solution of the Navier-Stokes is preserved (Ramshaw et al. 1986). The transformation is done in practice by scaling the specific gas constant r by a factor $\alpha^2 < 1$ in the equation of state

$$r^* = \alpha^2 r\,, p^* = \rho r^* T = \rho \alpha^2 r T\,, \tag{II.12}$$

where the superscript * denotes artificially scaled quantities and T is the temperature. The internal energy e and enthalpy h are linked by the relation

$$h = e + \frac{p}{\rho} = e + rT\,, \tag{II.13}$$

the internal energy is also rescaled by the same factor α^2 to maintain consistency between internal energy variation and mechanical work

$$e^* = \alpha^2 e, \tag{II.14}$$

so that

$$h^* = e^* + \frac{p^*}{\rho} = \alpha^2 \left(e + \frac{p}{\rho} \right) = \alpha^2 h. \tag{II.15}$$

The resulting speed of sound c^* is therefore artificially reduced by a factor α

$$c^* = \sqrt{\gamma\, r^*\, T} = \sqrt{\gamma\, \alpha^2 r\, T} = \alpha c\,. \tag{II.16}$$

This approach significantly alleviates the constraint on the time step due to the CFL condition. For the hydrodynamic problem to remain unchanged, the artificial Mach number $\mathrm{Ma}^* = \mathrm{Ma}/\alpha$ must remain small (typically $< 0.2 - 0.4$) (Ramshaw et al. 1986). The value retained for this study is $\alpha = 0.22$ leading to $\mathrm{Ma}^* < 0.15$. This results in a four-time increase of the unsteady time-step Δt and hence of the total computational cost, which makes the solver competitive with incompressible or anelastic approaches more conventionally used for ABL flows.

II.2.4.3 Conversion of species mass fraction into concentration

In practice, AVBP solves the transport equation (Eq. II.10d) for the species mass fraction w, while pollutant concentrations are often measured in terms of volume concentration c. Both quantities can be related using the ratio ρ_s/ρ between the density of the species and that of the air:

$$c = \frac{w}{w + (1-w)\left(\frac{\rho_s}{\rho}\right)}.$$

For most atmospheric dispersion applications, the considered species concentrations are very low, implying that $w << 1$ which yields

$$c \approx \frac{\rho}{\rho_s} \times w.$$

In the remaining of the thesis, concentration refers to volume concentration and is expressed in parts per million (ppm) by volume as in the literature (Yee and Biltoft 2004).

II.3 The Mock Urban Setting Test field campaign

II.3.1 Experimental site description

MUST is a field-scale experiment performed in September 2001 at the US Army Dugway Proving Ground test site in the Utah desert (USA) (Fig. II.2). Its objective was to provide extensive measurements in the short-to-medium range of a plume within an urban-like canopy in support of the development and validation of urban dispersion models (Biltoft 2001; Yee and Biltoft 2004).

The idealized urban canopy is mimicked by an array of 10×12 regularly-spaced shipping containers covering an area of about $200\times200\,\mathrm{m}^2$ (Fig. II.3). The array (in its x-y coordinate system) makes an angle of 30° to the north. The containers are 12.2-m long, 2.42-m wide, and 2.54-m high. The average distance between two containers is 12.9 m along the x-axis and 7.9 m along the y-axis. The terrain is flat and homogeneous with a mix of sparse greasewood and sagebrush ranging from 0.4 to 0.75 m high. It is worth mentioning that the geometry of the idealized canopy was slightly irregular, as the containers were not all perfectly aligned, and one container was replaced by a van (Biltoft 2001). Their impact on the flow field was studied in Santiago et al. (2010), but we consider this study a regular case as in Milliez and Carissimo (2007) and Nagel et al. (2022).

During the experiments of the MUST field campaign, a non-reactive gas (propylene) was released, passively, at different horizontal and vertical locations and for different atmospheric conditions (in terms of wind direction, wind speed, and atmospheric stability condition). The propylene density is $\rho_s = 1.81\,\mathrm{kg\,m^{-3}}$ but the negative buoyancy effects of the gas are insignificant (Yee and Biltoft 2004), and it can be considered a passive tracer.

II.3.2 Available experimental data

The MUST experimental dataset includes wind velocity and tracer concentration measurements within and outside the container array. This section summarizes the data that are used in this thesis, for the design and evaluation of the LES model, as well as validation of the data assimilation approach (Chapter V). The reader should refer to Biltoft (2001) for full details of the instruments used during the MUST field campaign.

For wind velocity measurements, two-dimensional and three-dimensional sonic anemometers were provided by the Dugway Proving Ground's West Desert Test Center (WDTC) and deployed vertically on different masts (triangles in Fig. II.3): four anemometers were mounted on the central tower T at $z = 4$, 8, 16 and 32 m; and three anemometers were mounted at $z = 4$, 8 and 16 m on each of the towers S and N (located 30 m upstream and downstream from the canopy, respectively). WDTC three-dimensional sonic anemometers were also positioned within the canopy at $z = 1.15$ m on four tripods V (yellow plus-symbols in Fig. II.3). An additional three-dimensional sonic anemometer, provided by the Arizona State University (ASU), measured wind velocity upstream of the containers, near tower S, at $z = 1.6$ m (purple plus-symbol in Fig. II.3). In addition, a network of meteorological data collection stations called Surface Atmospheric Measurement System

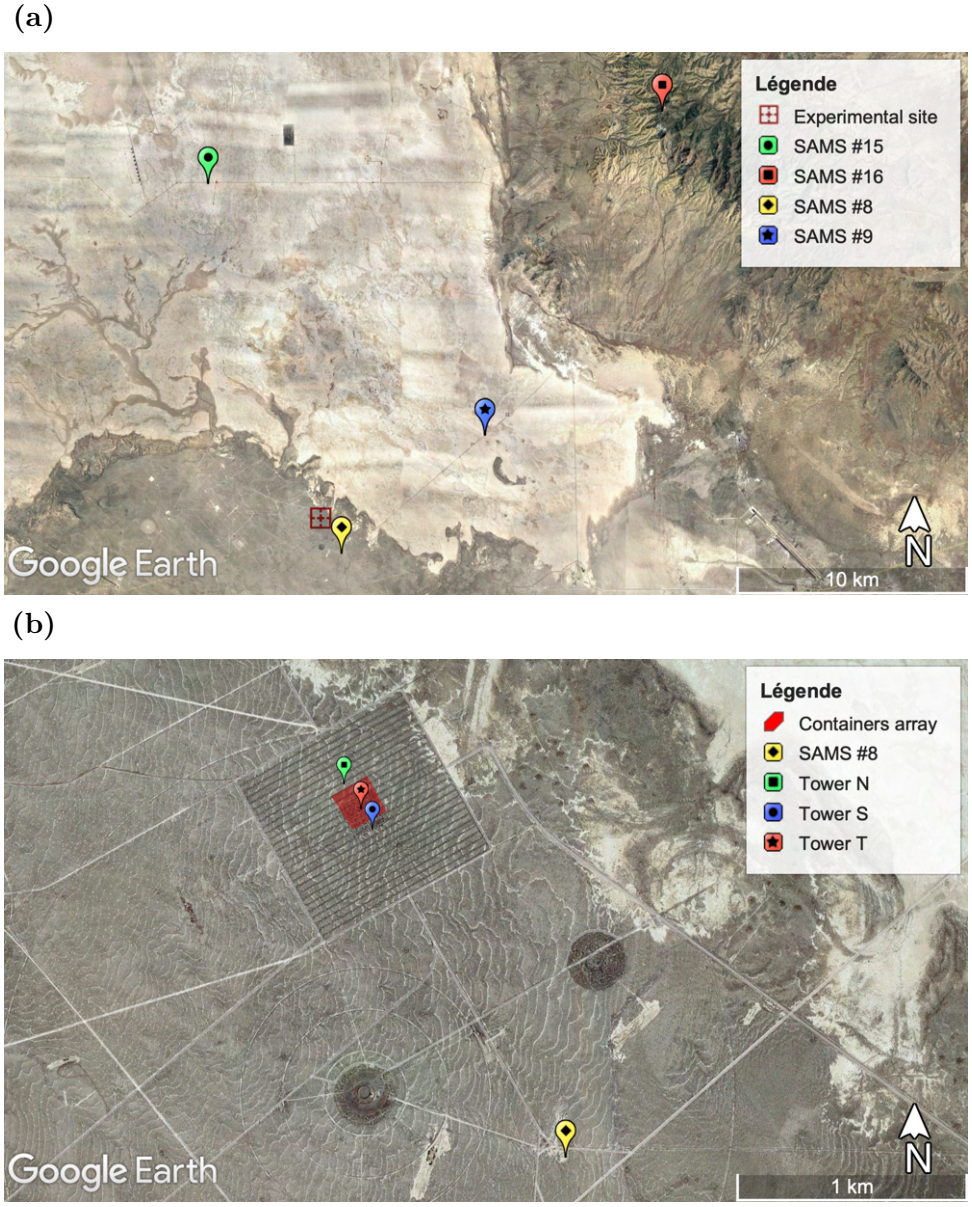


Figure II.2: *Satellite images of Dugway Proving Ground test site which hosted the MUST field campaign. (a) Wide view of the surrounding area, the experimental site is represented by the red crosshair, while the location of the SAMS meteorological stations #8, #9, #15, and #16 are indicated by the yellow, blue, green and red markers respectively. (b) Close-up view showing in red the area occupied by the containers. The towers S, T, and N equipped with anemometers are represented with the blue, red and green markers respectively.*

(SAMS) is installed across the Dugway Proving Ground. Each station measures wind velocity at $z = 10\,\mathrm{m}$, as well as temperature, humidity and pressure at $z = 2\,\mathrm{m}$ with acquisition times ranging from 5 to 15 minutes. The location of the SAMS stations for which we have access to measurements (#8, #9, #15, #16) is indicated in Fig. II.2a. The SAMS station #8 is the closest to the container array, about 1 600 meters southeast.

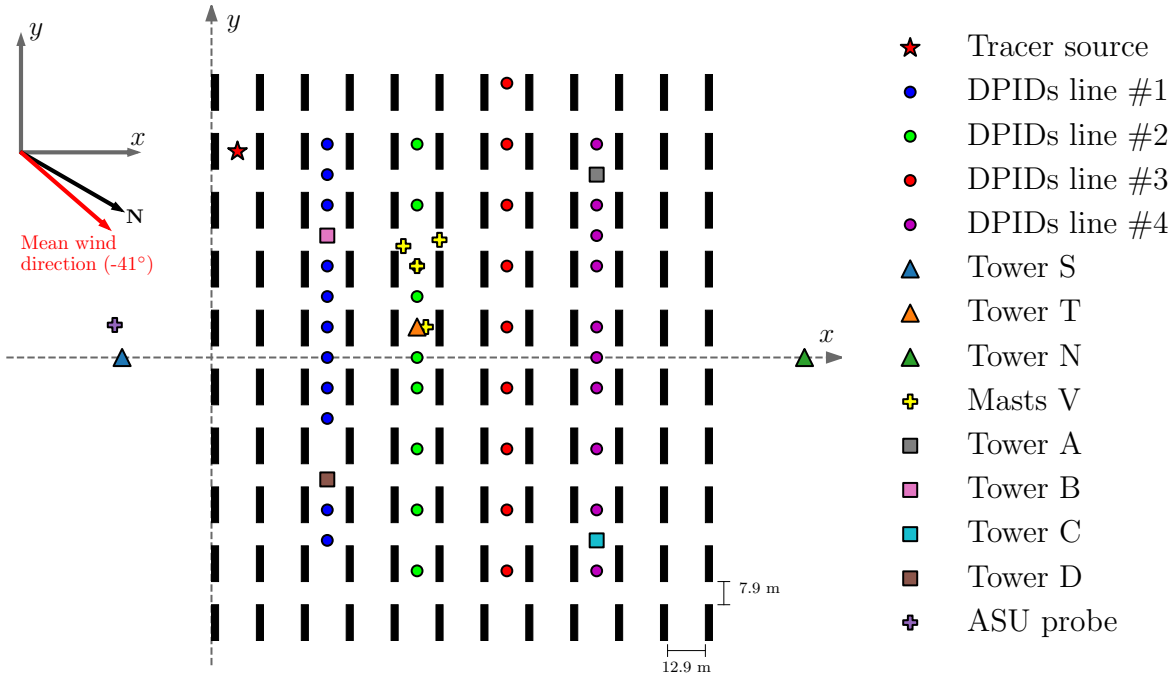


Figure II.3: *Schematic view of the MUST array configuration adapted from Kumar et al. (2015), the coordinate system used is the same as in Yee and Biltoft (2004) such that north corresponds to the angle of -30° in the x-y coordinate system. Black rectangles represent the shipping containers used to mimic the urban canopy. Triangles correspond to the anemometers mounted on towers S, T, and N. Yellow plus symbols correspond to the masts V equipped with WDTC anemometers, and the purple plus symbol to the ASU anemometer. Colored circles correspond to the DPID concentration samplers (one color for each line of sensors), and colored squares correspond to the UVIC concentration samplers mounted on towers A, B, C, and D (note that there is a DPID sampler at the same location as tower D). The upstream mean wind direction (red arrow) and the propylene-source location (red star) of trial 2681829 retained for the present study are also indicated.*

For tracer concentration measurements, 48 digi-photoionization detectors (DPIDs, colored circles in Fig. II.3) were used as well as 24 ultraviolet ion collectors (UVICs, colored squares in Fig. II.3). The DPIDs have a detection threshold of 0.04 ppm against 0.01 ppm for the UVICs. 40 DPIDs were placed within the canopy at $z = 1.6$ m, forming four sensor lines aligned with the y-axis (referred to as the DPIDs lines in Fig. II.3); eight were placed on the central tower T at $z = 1$, 2, 4, 6, 8, 10, 12 and 16 m. Also, six UVICs were mounted within the canopy on each of the four 6-m towers A, B, C, and D at $z = 1$, 2, 3, 4, 5, 5.9 m to obtain vertical concentration profiles.

We acknowledge the Defense Threat Reduction Agency (DTRA) for providing access to the MUST dataset.

II.3.3 Selected case

From all the available observations, 21 trials were chosen by Yee and Biltoft (2004) for their high quality (i.e. tracer detection on the tower T and for three of the four DPID lines). In addition, Yee and Biltoft (2004) extracted a 200-s quasi-stationary (in the statistical sense) period in each 15-minute experiment that minimizes the effect of mesoscale meteorological fluctuations on the tracer concentration time series. This time window (referred to as the analysis period in the following) was chosen as the sequence with the smallest variation in mean wind speed and direction at the upstream tower S for each trial.

In this work, we simulate one of the 21 trials referred to as 2681829, which has also been studied in other studies using both LES and RANS approach as shown in Table II.1. The trial main characteristics extracted from the data of Yee and Biltoft (2004) are summarized in Table II.2. This case is a configuration with neutral atmospheric conditions (i.e. afternoon transition from unstable to stable conditions), characterized by a high value of the surface Obukhov length L_o ($L_\mathrm{o} \gg 2\,500\,\mathrm{m}$), no latent and sensible heat fluxes, and a weak influence of buoyancy. The time-averaged wind speed u_4 and direction α_4 at $z = 4\,\mathrm{m}$ at the upstream tower S are respectively $7.93\,\mathrm{m\,s^{-1}}$ and -41° (this angle is defined with respect to the x-axis of the container array indicated in Fig. II.3, the north direction corresponding to an angle of -30°). The gas was released at $z_s = 1.8\,\mathrm{m}$ near the inlet of the canopy (red star symbol in Fig. II.3) with a constant flow rate $Q = 225\,\mathrm{L\,min^{-1}}$. For this selected trial, the analysis period is between 300 and 500 s after the acquisition start.

Table II.2: *Main characteristics of the MUST trial 2681829 (Yee and Biltoft 2004): α_4 and u_4 are respectively the time-averaged wind direction and wind speed at $z = 4$ m of the upstream tower S, L_o is the surface Obukhov length estimated by Yee and Biltoft (2004) using the vertical fluxes of temperature and velocity measured at $z = 4$ m of the central tower T, Q is the constant tracer release rate at the source, and (x_s, y_z, z_s) are the source coordinates. The flow statistics are computed on the [300; 500 s] analysis period*

Trial	Local start time (UTC - 6 h)	α_4 (°)	u_4 (m s^{-1})	L_o (m)	Q (L min^{-1})	x_s (m)	y_s (m)	z_s (m)
2681829	2001/09/25 1830	-41	7.93	28 000	225	8.87	70.35	1.8

II.4 Large-eddy simulation model of the MUST trial 2681829

II.4.1 Computational domain and spatial discretization

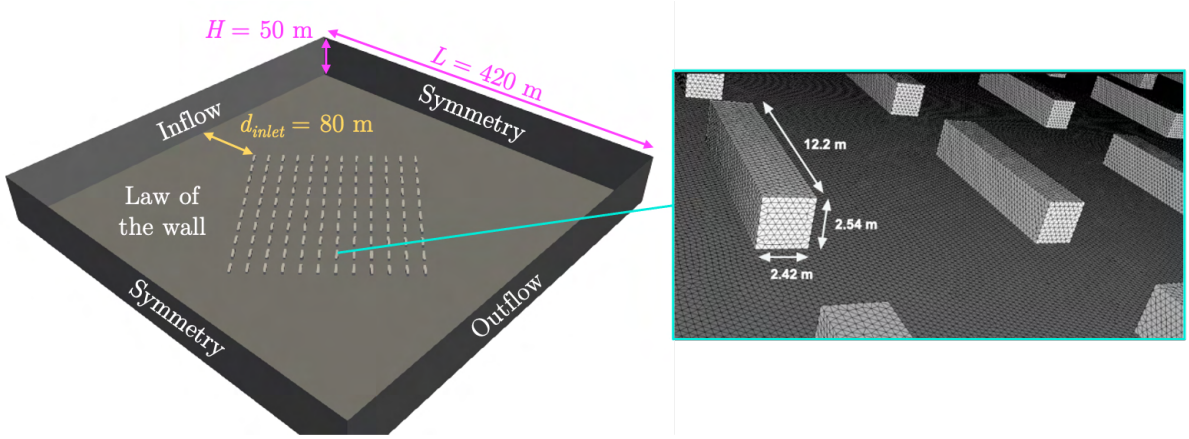


Figure II.4: *Computational domain of the LES model of the MUST experiment and view of the reference mesh used in the obstacle region.*

The computational domain in which the Navier-Stokes and the tracer transport equations are solved is a rectangular cuboid oriented so that the inlet boundary is normal to the mean upstream wind direction (Fig. II.4). In the x-y plane, the domain is a $420 \times 420\,\mathrm{m}^2$ square centred on the container array. Along the z-axis, the height of the domain is 50 m. To avoid lateral or vertical confinement effects, the distance between the lateral boundaries and the container array is at least 80 m (corresponding to $30\,H$, with $H = 2.54\,\mathrm{m}$ the container height), and the distance between the top boundary and the top of the containers is 18 H. This geometry ensures compliance with the guidelines for CFD simulation of urban atmospheric flows (Tominaga et al. 2008; Franke et al. 2007), with a blockage ratio of 3.1% and sufficient distances between the obstacles and the boundary conditions. The results of a sensitivity test to the height of the domain are given in Sect. A.2, page 233.

An unstructured and boundary-fitted mesh of 91 million tetrahedra is used to discretize the computational domain. In the region of interest (in a box of $246 \times 266 \times 3.6\,\mathrm{m}^3$ that contains the full container array), the mesh is uniform with a resolution equal to $\Delta x = \Delta y = \Delta z = 0.3\,\mathrm{m}$. In the rest of the domain, the mesh has a resolution of 0.3 m at the ground level except near the outlet and the lateral boundaries, where the resolution was coarsened to 2 m to reduce the number of cells. On the vertical, the mesh is gradually stretched to reach a 5-m resolution at the top boundary. The mesh is generated using Centaur[5] software with a maximum stretching ratio of 1.7. The main characteristics of the resulting mesh are given in Table A.1, page 232.

[5]https://www.centaursoft.com/

The mesh resolution used in this study is in line with the resolutions used in LES modeling of the MUST experiment in the literature, which typically ranges from 50 cm in König (2014) to 30 cm in Nagel et al. (2022). We also try to comply as much as possible with the mesh guidelines required by the use of wall laws presented in Sect II.4.2.2:

- for the rough wall law used to model the effect of the ground (Eq. II.19), the first node height z_1 should $z_1 > 50z_0$ with z_0 the aerodynamic roughness length (Basu and Lacser 2017).
- for the smooth wall law used for the obstacles (Eq. II.20), the first cell should be located within the log-layer of the boundary layer velocity profile. For atmospheric flows in urban environments, Franke et al. (2007) recommend that $30 < z^+ < 500$, with $z^+ = \rho u_\tau z_1/\mu$ the height of the first node in wall units based on the local friction velocity u_τ.

The first node height of the mesh used in this study is $z_1 = 0.12\,\text{m}$ which is equivalent to approximately $5z_0$ and 12×10^3 wall units. This means that the mesh is too coarse for the smooth wall laws near the obstacles and too refined for the rough wall law used for the ground. It is a problem that cannot be solved as the two guidelines cannot be matched simultaneously in this particular case. A compromise is made in the zone of interest by using approximately 8 cells over the height of the obstacle (Fig II.4) to explicitly account for the effect of the obstacles on the flow while keeping a reasonable model computational cost. We demonstrate in Appendix A.1, page 231 that this spatial resolution is enough to reach convergence of the LES predictions.

With the CFL condition on the LES time-step equal to 0.9 with the LW numerical scheme (Eq. II.11), the time step associated with this mesh is equal to $7.9 \times 10^{-4}\,\text{s}$ when applying the PGS technique presented in Sect. II.2.4.2, compared to $1.9 \times 10^{-4}\,\text{s}$ without. We therefore estimate that PGS reduces the computational cost of the LES model by a factor of 4.

II.4.2 Boundary conditions

II.4.2.1 Inflow boundary conditions

One challenge in LES of near-field pollutant dispersion relates to the modeling of inflow boundary conditions (Muñoz-Esparza et al. 2014; Dauxois et al. 2021). In field-scale applications, there is usually a limited amount of information available to represent the complexity of actual microscale inflow conditions that are influenced by the ABL variability. One way to represent the mesoscale/microscale interactions is to perform a dynamical downscaling of the atmospheric flow using a multi-scale meteorological model based on grid nesting (Wiersema et al. 2020; Nagel et al. 2022). This multi-scale approach resulted in a significant improvement of the microscale flow velocity and tracer concentration predictions for the Oklahoma City Joint Urban 2003 experiment (Wiersema et al. 2020). However, this finding did not hold for the MUST idealized urban environment, where a standalone microscale LES configuration based on idealized inflow boundary conditions achieved the same level of accuracy as a multi-scale approach (Nagel et al. 2022). Therefore, we represent the turbulent inflow boundary condition using an idealized approach in this work.

The logarithmic wind profile from Richards and Hoxey (1993), representing a fully developed neutral atmospheric surface layer, is imposed at the inlet. This description is sufficient as i) the selected trial corresponds to a neutral stratification condition; and ii) we focus on the near-surface flow inside and just above the canopy. The mean horizontal wind velocity $\overline{u_{inlet}}$ at height z reads

$$\overline{u_{inlet}}(z) = \frac{u_*}{\kappa} \ln\left(\frac{z+z_0}{z_0}\right), \tag{II.17}$$

where z_0 (m) is the aerodynamic roughness length equal to 0.045 ± 0.005 m according to observations (Yee and Biltoft 2004), κ is the von Kármán constant equal to 0.4, and u_* ($\mathrm{m\,s^{-1}}$) is the friction velocity. The parameter u_* is calibrated here by fitting the profile (Eq. II.17) through a least-square regression on wind speed measurements available at the upstream tower S and for the ASU anemometer (these data are described in Sect. II.3.2), which leads to a value of $u_* = 0.73\ \mathrm{m\,s^{-1}}$. The corresponding vertical profile for the inlet mean wind is shown in Fig. II.5a along with the measurements used for regression.

A constant wind direction α_{inlet} is imposed on the vertical at the inlet so that the inlet wind vector reads

$$\overline{\mathbf{u}} = \left(\overline{u_{inlet}}\,\cos(\alpha_{inlet}),\ \overline{u_{inlet}}\,\sin(\alpha_{inlet}),\ 0\right)^{\mathrm{T}}, \tag{II.18}$$

in the MUST frame of reference (see Fig. II.3). The constant wind direction is obtained by spatially averaging the four wind direction measurements available at tower S and for the ASU anemometer. This leads to $\alpha_{inlet} = -41°$, the same value as measured at $z = 4\,\mathrm{m}$ (see Table II.2).

II.4.2.2 Wall boundary conditions

The ground boundary is modeled as a rough surface with imposed shear stress τ according to the Monin Obukhov similarity theory:

$$\tau/\rho = \left(\frac{\kappa u_h}{\ln\left(\frac{z_1+z_0}{z_0}\right)}\right)^2, \tag{II.19}$$

where u_h is the horizontal velocity measured at height z_1 of the first vertex above the ground, ρ is the air density, $\kappa = 0.4$ is the von Kármán constant. Concerning the aerodynamic roughness length we use the same value as for the inlet wind profile (Eq. II.17). By using this law of the wall we assume that the ABL is well-established and in equilibrium. We show in Sect. II.5 that the LES model preserves this profile rather well in free-field conditions.

For the surfaces of the obstacles, the shear stress is imposed to match the law of the wall for a smooth surface based on a viscous length (Larsson et al. 2016):

$$u^+ = \frac{1}{\kappa} \ln\left(Ez^+\right), \tag{II.20}$$

where $u^+ = u_h/u_\tau$ is the normalized horizontal velocity measured at the first vertex above the obstacle surface in wall units, $z_1^+ = \rho z_1 u_\tau/\mu$ is the normalized height of this vertex, and E a constant set equal to 9.2.

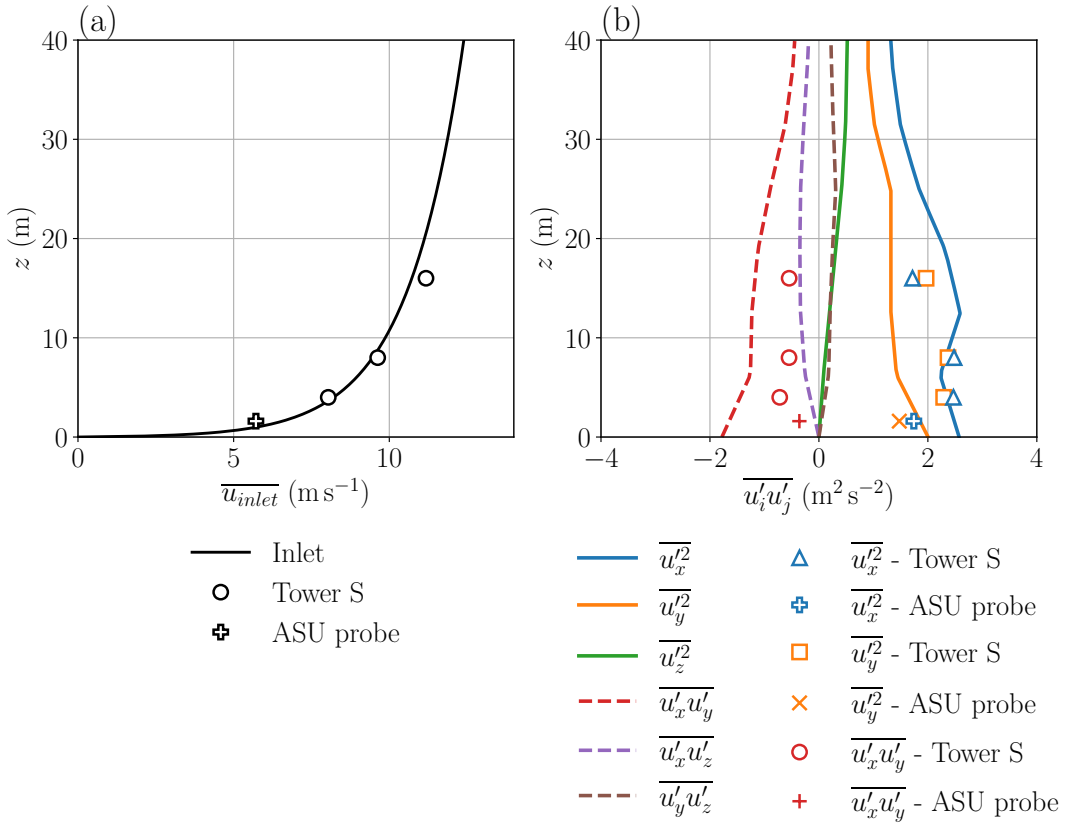


Figure II.5: *Vertical profiles (solid lines) of the (a) inlet mean wind speed $\overline{u_{inlet}}(z)$, and the (b) inlet wind speed fluctuations $\overline{u_i'u_j'}(z)$ predicted by a precursor simulation and used to defined the LES inflow boundary condition. Symbols correspond to experimental data.*

II.4.2.3 Other boundary conditions

At the outlet and top boundaries, the static pressure is imposed in a soft manner (with a relaxation towards the target) to evacuate acoustic waves (Poinsot and Lele 1992). Symmetry boundary conditions are used for the lateral boundaries as shown in Fig. II.4. In addition, a 5-meter sponger layer with increased artificial viscosity is defined along the outlet surface to prevent numerical instabilities.

II.4.3 Turbulence injection method in the context of ABL flow

The AVBP LES model of the MUST trial 2681829 studied in this thesis was originally developed by Rochoux et al. (2021) as i) a validation test case for environmental fluid flows in open areas as AVBP is not usually used in this context, ii) part of a multi-model comparison to assess LES dispersion models structural uncertainties. However, for the sake of comparison between the different codes, turbulence injection was not used since the codes rely on different turbulence injection methods, which would add discrepancies in the comparison. This implies that the prescribed inlet flow was uniform and steady and thus an oversimplification of the real turbulent ABL.

One of the main contributions of this thesis to the AVBP model of the MUST case is to rely on a turbulence injection method to have a more realistic inlet boundary condition, i.e. representative of the turbulent nature of the ABL behavior. Various methods are available in the literature for this purpose. They can be crudely classified into three categories: periodic precursor methods, recycling methods, and synthetic turbulence injection methods (Dhamankar et al. 2018). We opt for the approach proposed by Vasaturo et al. (2018) which consists of using synthetic turbulence injection with prescribed turbulence statistics obtained from a periodic precursor. The approach has two main advantages: firstly, it is relatively inexpensive, as the precursor simulation is only performed once at a preliminary stage; secondly, as an objective of the study is to perform parametric variations on the inlet quantities, this is done easily in this framework by rescaling properly the obtained turbulence statistics, without additional computations.

Concerning the synthetic turbulence injection, we use the Kraichnan-Celik (Kraichnan 1970; Smirnov et al. 2001) method already implemented in the AVBP solver. In this method, temporal flow fluctuations $\mathbf{u}'$ are added to the mean inlet wind profile (Eq. II.18), according to Reynolds' decomposition: $\mathbf{u}(\mathbf{x},t) = \overline{\mathbf{u}(\mathbf{x})}+\mathbf{u}'(\mathbf{x},t)$, with x the position vector and t the time. The fluctuations are first expressed using a three-dimensional Fourier transform in space, and the Taylor assumption to link temporal and spatial fluctuations. The solution is then approximated by a truncated Fourier series. The coefficients of the series are adequately randomly drawn to satisfy the incompressibility assumption, and to match the turbulence spectrum (Eq. II.21) from Passot and Pouquet (1987):

$$E(k) = 16\frac{u_{RMS}^2}{k_e}\sqrt{\frac{2}{\pi}}\left(\frac{k}{k_e}\right)^4 \exp\left(-2\left(\frac{k}{k_e}\right)^2\right), \text{ with } k_e = 2\pi/\lambda_e. \tag{II.21}$$

The Celik extension (Smirnov et al. 2001) of the Kraichnan method allows prescribing anisotropic and heterogeneous profiles of the velocity fluctuations at the inlet. This is done by an orthogonal transformation of the fluctuation field generated using the Kraichnan method based on the target Reynolds stress tensor over the inlet surface:

$$\mathbf{R}_{ij}(\mathbf{x}) = \rho\overline{u_i'u_j'}(\mathbf{x}). \tag{II.22}$$

The components of the Reynolds stress tensor (Eq. II.22), used to prescribe the target level of fluctuations in the Kraichnan-Celik method, are estimated using a preliminary simulation with the same surface roughness but without obstacles, and with periodic boundary conditions at the inlet and outlet (Keating et al. 2004; Munters et al. 2016; Vasaturo et al. 2018). In this periodic setup, an additional pressure gradient term is added in Eq. II.1b in order to drive the flow. This periodic simulation is run at a 6.25-m resolution over a $400 \times 400 \times 250\,\mathrm{m}^3$ computational domain and a 2-hour period to obtain converged velocity fluctuation statistics. The resulting mean velocity fluctuations are shown in Fig. II.5b alongside fluctuation measurements. Even though experimental measurements of fluctuations were not used to calibrate the precursor simulation, it is overall consistent with the level of fluctuations measured at tower S upstream of the containers. The $\overline{u_x'^2}$ fluctuation profile is accurately predicted, however, the precursor tends

to underestimate $\overline{u_y'^2}$ and overestimate $\left|\overline{u_x' u_y'}\right|$, especially below 5 m. These fluctuations statistics form the Reynolds stress tensor (Eq. II.22) imposed at the inlet of the microscale domain.

Table II.3: *Parameters values used in this study to calibrate the turbulence injection method. N_{modes} is the finite number of Fourier modes used to represent the fluctuations, λ_e is the most energetic length of the turbulent spectrum, L_c is the cutoff length, U_{bulk} is the averaged bulk velocity in the inlet section, and $\mathbf{R}_{ij}$ is the Reynolds stress tensor.*

N_{modes}	λ_e	L_c	U_{bulk}	$\mathbf{R}_{ij}$
500	25 m	6 m	11.0 m s^{-1}	*computed using precursor*

Table II.3 summarizes the turbulence injection parameter values used in this study. The bulk velocity is calculated analytically by integrating the velocity mean profile (Eq. II.17) over the cross-section of the inlet. The most energetic length of the turbulent spectrum λ_e which parameterizes the turbulence spectrum injected (Eq. II.21) is taken to its maximum limit which is half the height of the inlet surface. We refer the reader to Boudin (2021) for additional details on the choice of parameters. The limitation on λ_e leads to an undestimation of the integral scale of turbulence $\Lambda = \lambda_e/\sqrt{2\pi}$ when compared to the real ABL (Fig. II.6). The domain height could be increased to be able to inject larger structures, but the method would remain imperfect because it assumes a uniform integral scale. This is typically not the case within the ABL (Fig. II.6), for which the integral scale increases linearly with height. Still, the retained value for λ_e allows it to be consistent with the experimental data in the region of interest close to the ground.

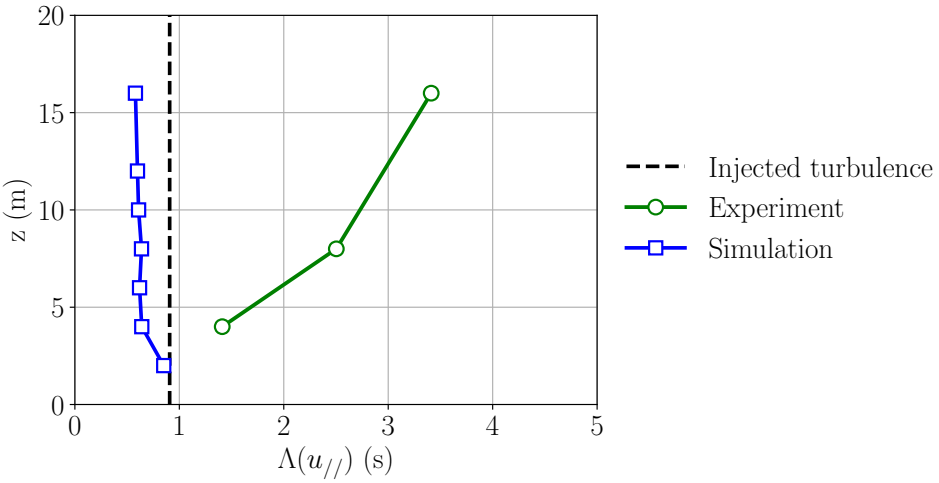


Figure II.6: *Turbulence integral time scale based on the velocity alongside the streamwise direction $u_{//}$. The simulation and experimental data are represented as blue and red lines respectively. The integral time scale prescribed at the inlet is depicted as a dashed black line.*

A detailed evaluation of the behavior of the injection method is investigated in Sect. II.5, assessing how the injected mean and fluctuation statistics of velocity evolve

through an obstacle-free computational domain. In addition, we assess how adding turbulence injection impacts the LES model dispersion predictions in Appendix A.3, and the model sensitivity to the level of turbulence prescribed at the inlet is investigated Sect. III.5.2, page 109.

II.4.4 Tracer source modeling

The propylene source is modeled as a volumetric source term S in the transport equation (Eq. II.1c). The momentum of the released gas is not taken into account as it exits the outlet pipe at a velocity below $1\,\mathrm{m\,s^{-1}}$ and Biltoft (1995) showed that it has a negligible effect on the plume elevation, especially for the strong wind conditions considered in this trial. To avoid strong concentration discontinuities, a Gaussian shape of radius r_s in space is used, so for each vertex with coordinate $\mathbf{x}$ of the computational domain, the source term reads

$$\phi_r(\mathbf{x}) = \frac{Q}{V_s}\omega(\mathbf{x})e^{\frac{-\|\mathbf{x}-\mathbf{x}_s\|^2}{2r_s^2}},$$

with Q the experimental tracer release rate ($\mathrm{m^3\,s^{-1}}$), and $\mathbf{x}_s = (x_s,\ y_s,\ z_s)$ the source location. Values of Q, x_s, y_s, and z_s corresponding to the selected trial are given in Table II.2. The dual volume of the vertex at location $\mathbf{x}$ is denoted $\omega(\mathbf{x})$, and V_s is the total volume of the source model:

$$V_s = \sum_{\mathbf{x}\in\Omega} \omega(\mathbf{x})e^{\frac{-\|\mathbf{x}-\mathbf{x}_s\|^2}{2r_s^2}}\mathrm{d}\mathbf{x},$$

so that the total volumetric flow rate of the injected pollutant matches the experimental one:

$$\sum_{\mathbf{x}\in\Omega} \phi_r(\mathbf{x})\mathrm{d}\mathbf{x} = Q.$$

A source half-width r_s of 0.45 m was chosen so the Gaussian-shaped source is covered by approximately 6 cells in each direction.

II.4.5 Initial condition and spin-up time definition

The LES simulation is initialized using a homogeneous flow field in the horizontal directions equal to the prescribed inlet mean field (Sect. II.4.2). A spin-up time of 60 s, which corresponds to approximately 17 times the LES turnover time H/u_*, is used so that first- and second-order statistics of the flow and the tracer reach a stationary state. A 200-s time window corresponding to the [300; 500 s] analysis period (Sect. II.3.3) is then simulated, from which statistics of the flow and tracer concentration variables can be collected. At probe location, outputs are saved with a resolution of 0.05 s.

Figure II.7 shows different examples of time series of horizontal wind speed magnitude and propylene concentration predicted by the LES model over the complete simulation duration. It visually demonstrates that the 60-s spin-up time is sufficient to reach a stationary state. In addition, we verify that the spin-up is sufficiently greater than the

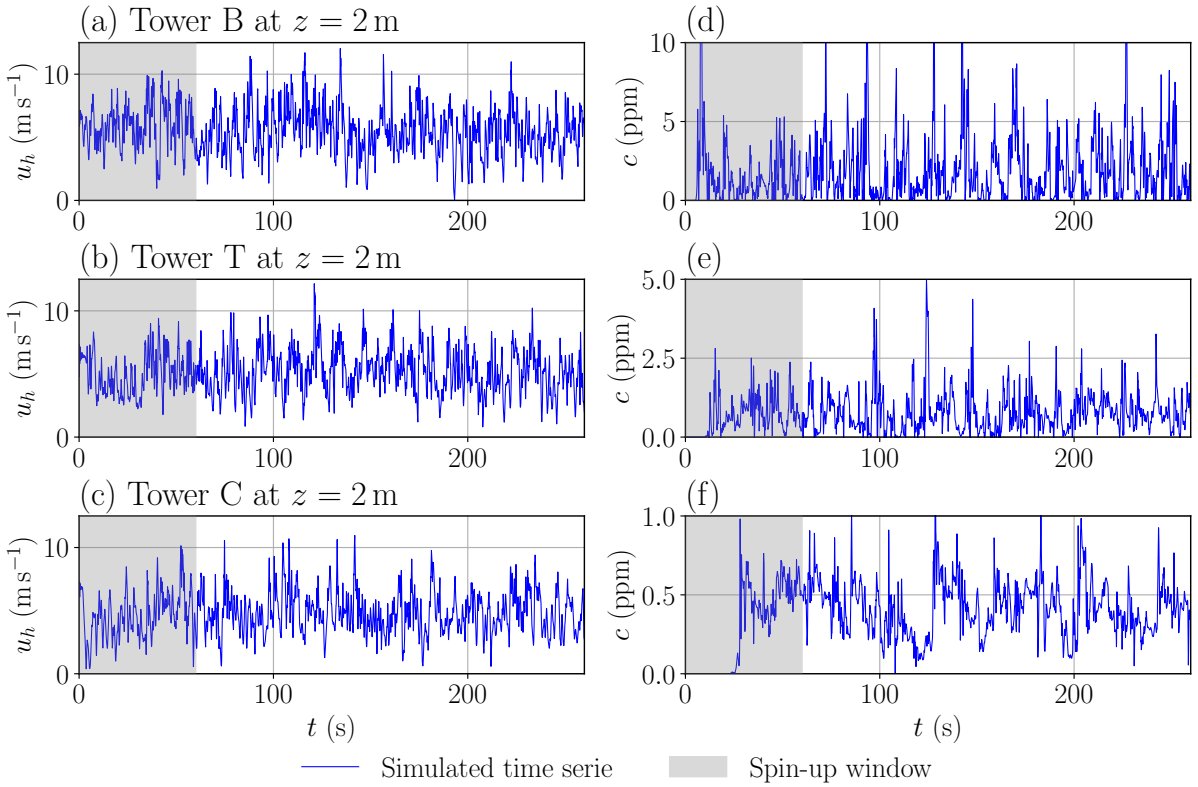


Figure II.7: *LES predictions of the horizontal wind speed magnitude and propylene concentration times series over at tower B (a, d), tower T (b, e), and tower C (c, f) at $z = 2\,\mathrm{m}$. The location of these towers is depicted in Fig. II.3. Shaded grey areas correspond to the 60-s spin-up.*

convective time scale, since the tracer reaches tower C, located at the edge of the zone of interest, after about 30 s, i.e. half the spin-up time (Fig. II.7e). It also shows the highly volatile nature of the quantities of interest, which motivates the study not only of their mean values but also of the statistics of their temporal distributions.

In terms of computational cost, simulating a 260-s physical time period (including the spin-up and the 200-s analysis period) for this MUST configuration costs approximately 20 000 core hours, using 1 344 CPU cores on the TGCC Irene SKL supercomputing facility (Intel Skylake architecture). More details on the computing resources used during this thesis are given in Table IV.1, page 142.

Illustrative examples of instantaneous flow and tracer concentration fields obtained by the LES model after the spin-up period are given in Fig. II.8a, b. The transition from large-scale turbulence provided by the turbulent inlet forcing to small-scale turbulent structures induced by the containers is visible in Fig. II.8a. The resulting instantaneous tracer concentration field is shown in Fig. II.8b, highlighting the scale disparity of local tracer concentration values that can reach 10 ppm near the emission source. Figure II.8c shows the time-averaged propylene concentration over the 200-s analysis period within the canopy. It highlights the deviation of the mean plume centerline from the incident

mean wind direction because of the wind channeling effect induced by the obstacles.

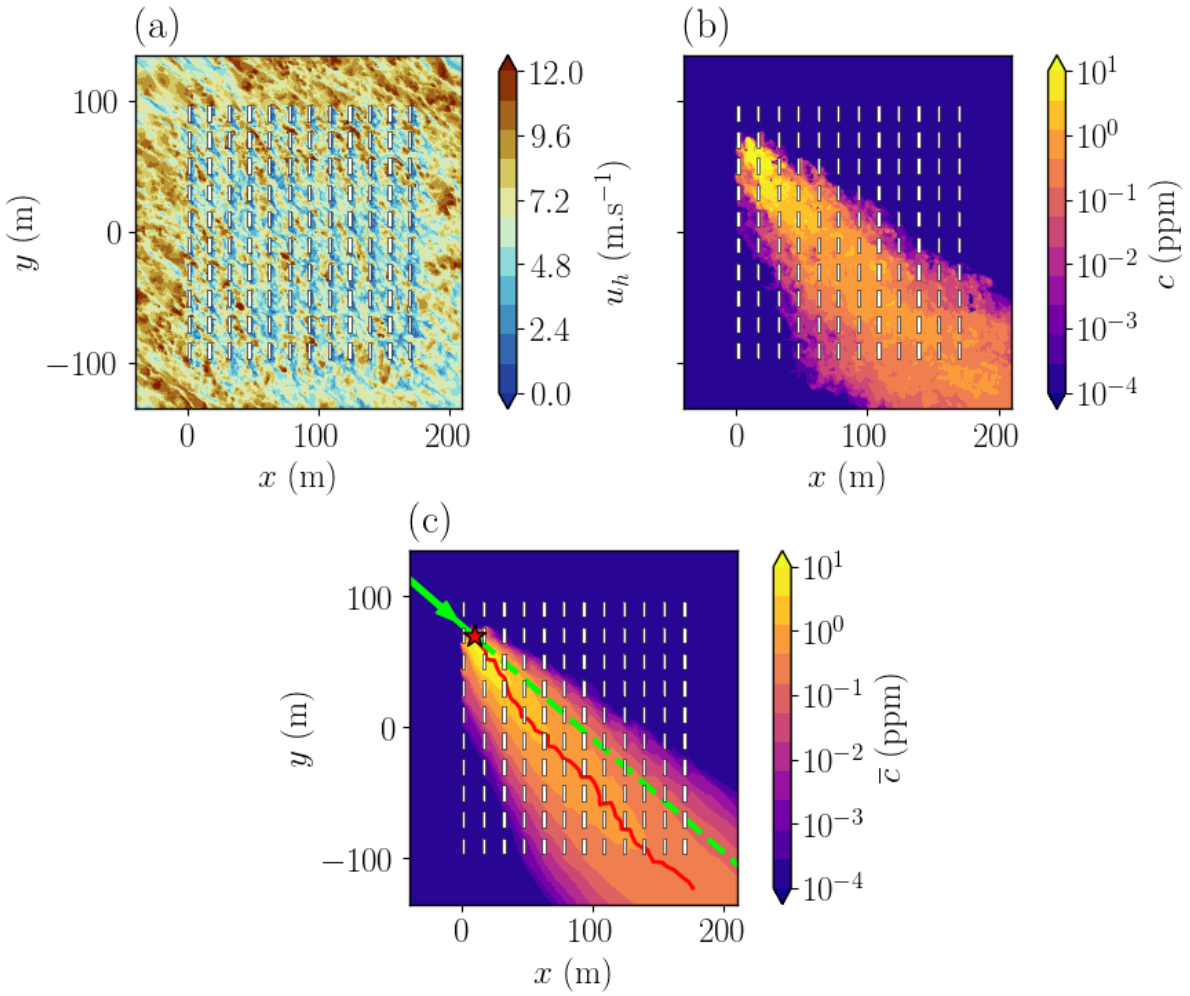


Figure II.8: *Horizontal cuts at $z = 1.6$ m of instantaneous (a) horizontal wind speed magnitude u_h ($m\,s^{-1}$) and (b) propylene concentration c (ppm) at $t = 60$ s. (c) Horizontal cut of the time-averaged concentration over the 200-s analysis period. White rectangles represent containers. The red star represents the tracer source, and the green line represents the mean wind direction imposed at the inlet. The plume centerline, identified by the positions of the mean concentration maximum on lines orthogonal to the incident wind angle, is represented as a red line (c).*

II.5 Preliminary verification: simulation of a free-field case

The goal of this preliminary study is to investigate how the profiles imposed at the inlet boundary condition of the LES model evolve through the domain. In particular, we aim to verify the consistency between inflow conditions and the ground wall boundary condition (Eq. II.19), in an empty computational domain (without the obstacle array). Inconsistencies between inlet and surface boundary conditions can indeed lead to unintended streamwise gradients in the vertical profiles of mean wind speed and turbulence quantities that might deform the incident profile on the zone of interest compared to that imposed in the inlet (Vasaturo et al. 2018). Such preliminary free-field simulation is considered best practice in CFD modeling of atmospheric surface flows both for RANS approach (Franke et al. 2007; Blocken 2015) and LES approach (Vasaturo et al. 2018).

For the verification in the free-field configuration to be representative of the target MUST case, the LES setup relies on the same computational domain and boundary conditions (Fig. II.4), but without obstacles. Dimensions of the computational domain are the same as for the model with obstacles i.e. $420 \times 420 \times 50\,\mathrm{m}^3$ (Fig. II.4). Compared to the mesh used in the MUST case, the mesh is still progressively stretched from a 0.3-m resolution at ground level to a 5-m resolution at the top boundary, but the additional refinement found in the obstacle array is omitted for this case, enabling the number of grid cells to be divided by a factor of two compared to the MUST case. The 0.3-m resolution corresponds to a ratio between the first off-wall grid point and the roughness length $z_1/z_0 = 6.7$, which violates the guideline $z_1/z_0 > 50$ from Basu and Lacser (2017) followed for the precursor simulation. This overly fine grid resolution may lead to a misprediction of fluctuations which can degrade statistics predicted by LES (Basu and Lacser 2017). However, this grey zone where the logarithmic law is not strictly valid is required to ensure a smooth grid resolution transition to the fine grid resolutions required in the obstacle-resolved region. Improving the representation of the rough surface in this grey zone is still an active topic (Arthur et al. 2019; Maronga et al. 2020).

Aware of this modeling limitation, in the following, we examine the evolution of the vertical profiles of the wind velocity statistics. For this purpose, probes have been placed every 25 m after the inlet in the streamwise direction. We are particularly interested in the profile obtained at a distance of around 80 m from the inlet since it corresponds to where the first obstacle is located in the reference model (Fig. II.4), in order to verify the incident mean and fluctuation velocity statistics on the zone of interest.

Figure. II.10 shows the evolution of the mean wind velocity vertical profile with increasing distance from the inlet. The profiles appear to remain nearly constant in the first 100 meters. The profile then deteriorates with acceleration close to the ground. It was verified that this is due to the over-refinement of the mesh given the rough wall law used ($z_1 < 50z_0$, as mentioned in Sect. II.4.1). Nevertheless, the profile is overall well-preserved until the area of interest of the complete LES model of MUST (which starts at $d_{inlet} \approx 80$ m).

Regarding higher order statistics, we also investigate how the velocity fluctuations imposed at the inlet, with the turbulence injection approach presented in Sect. II.4.3,

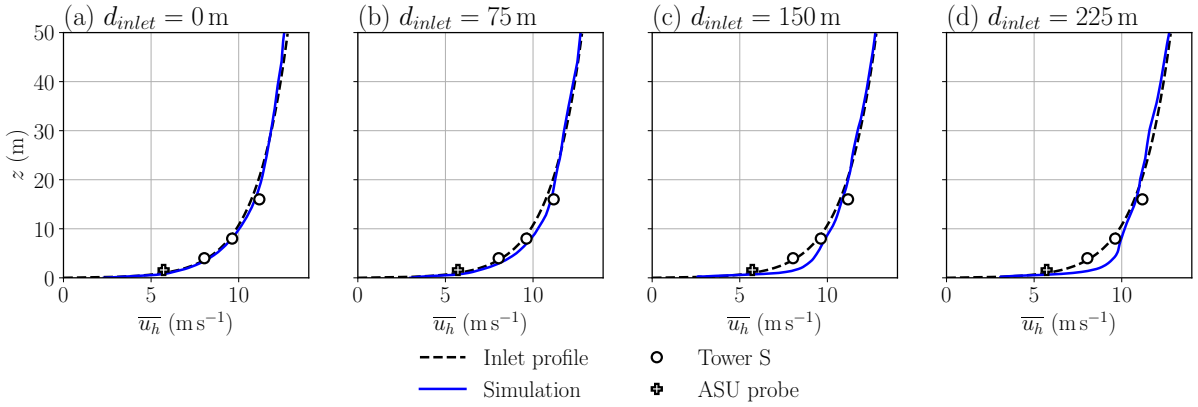


Figure II.9: *Spatial evolution of the mean horizontal velocity $\overline{u_h}$ vertical profile, at a distance from the inlet ranging from 0 to 225 m. The blue lines correspond to the predicted profile through the domain, while the profile imposed at the inlet is shown as black dashed lines. Symbols correspond to tower S and ASU anemometer measurements.*

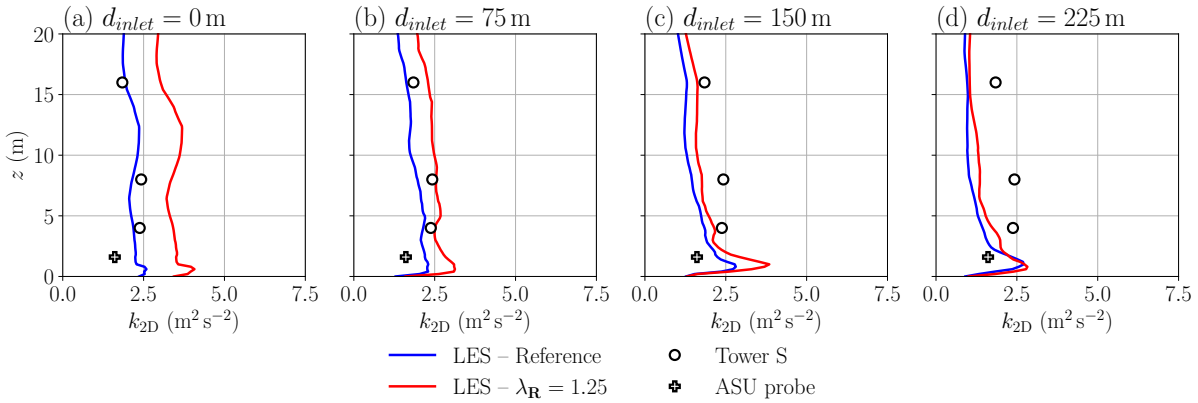


Figure II.10: *Spatial evolution of the horizontal turbulent kinetic energy k_{2D} vertical profile, at a distance from the inlet varying from 0 to 225 m. The blue and red lines correspond respectively to the reference LES and an additional LES with rescaled Reynolds tensor (Eq. II.23) to increase the level of intensity of fluctuations injected. Symbols correspond to tower S and ASU anemometer measurements.*

evolve through the empty domain. The evolution of the horizontal turbulent kinetic energy $k_{2D} = \frac{1}{2}(\overline{u_x'^2}+\overline{u_y'^2})$ with increasing distance from the inlet is shown in Fig. II.10. We do not compare the 3–D turbulent kinetic energy as tower S anemometers did not measure the vertical velocity during the MUST field campaign. The horizontal turbulent kinetic energy above the ground is fairly consistent at the first two locations with experimental level and is coherent with ABL theory (which provides a rough estimate for the 3D turbulent kinetic energy as $k \simeq u^{*2}/0.09^{0.5} = 1.8$ m^2/s^2 (Richards and Hoxey 1993)). Concerning the conservation of the injected fluctuation statistics, we find that turbulent kinetic energy increases significantly near the ground for $d_{inlet} > 150$ m, which might be an effect of the overly fine resolution at the ground. It also tends to dissipate in the upper

part of the domain as turbulence is damped by the outflow boundary condition imposed at the top of the domain. The effect of the top boundary condition is further investigated in Sect. A.2, page 233.

To investigate the capability of the turbulence injection method to represent perturbations of turbulent kinetic energy, we carried out an additional simulation by rescaling the Reynolds tensor prescripted at the inlet (Eq. II.22) as follows:

$$\mathbf{R}_{ij} \leftarrow {\lambda_{\mathbf{R}}}^2 \times \mathbf{R}_{ij}. \tag{II.23}$$

The evolution of the profile resulting from a 25% increase in injected fluctuations is shown in Fig. II.10. We find that the increased turbulent kinetic energy profile, which would correspond to a non-equilibrium situation, tends to return over a rather short distance to the equilibrium levels.

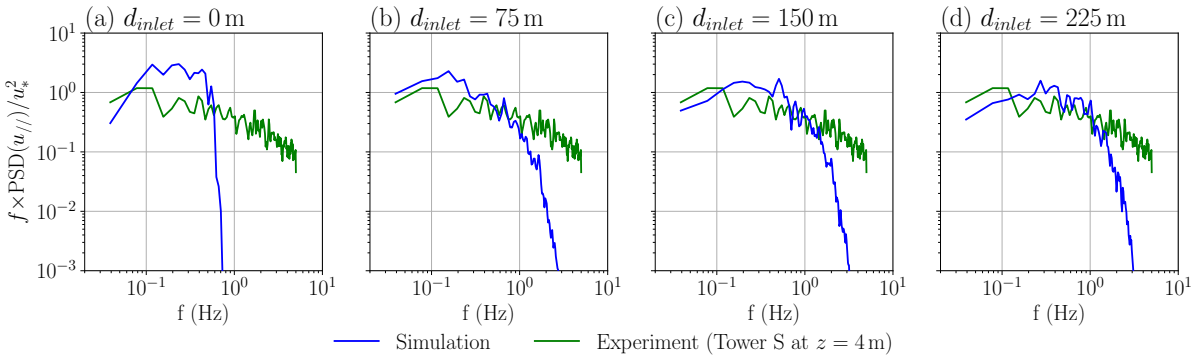


Figure II.11: *Spatial evolution of the turbulent energy spectrum evaluated from the streamwise velocity at z=4 m, at a distance from the inlet varying from 0 to 225 m. The predicted spectrum is shown in blue, while the green spectrum is computed from the tower S anemometer measurement.*

Looking at the turbulent scales, it is shown in Fig. II.11 that the injected synthetic spectrum (Eq. II.21) lacks high-frequency content. Still, it develops well within the empty computational domain until a fully-developed turbulence energy cascade is obtained at around 100 m from the inlet. Moreover, the developed turbulent energy spectrum reproduces rather well the measured turbulence spectrum near the ground. Note that low-frequency fluctuation scales are filtered because of the limit on the integral scale λ_e of the turbulence injection method used (Fig. II.6).

The results obtained above in terms of mean velocity, turbulent kinetic energy and frequency content show that the constructed LES setup is well suited to reproduce the idealized boundary layer in an equilibrium state, and is therefore weakly sensitive to the parameters of the turbulent injection method, provided that a sufficient distance is imposed between the inlet and the zone of interest, to allow for the cascade of the initial synthetic spectra to a more physical state. In particular, this explains why LES predictions in the area of interest are barely affected by turbulence injection parameters (Sect. III.5.2, page 109). Further methodical work would be needed to define a proper way to represent non-ideal effects, such as non-equilibrium between the flow and the ground surface.

Still, a slight deviation of the vertical profiles of mean and fluctuation velocity statistics is observed, which can be attributed to the over-refinement of the mesh near the ground required to resolve the container region. The top boundary condition imposed at the top also has a detrimental impact on the turbulence statistics, as it dampens fluctuations. Nevertheless, a rather good compromise between turbulence development and mean inlet velocity profile conservation is obtained around 80 m, which corresponds to the distance between the inlet and the first obstacle for the complete LES model of the MUST experiment. Beyond, the effect of the ground boundary condition has a lesser impact on the flow, as it becomes mainly driven by the significant drag forces resulting from the interaction with the containers forming the canopy.

II.6 Summary

In this chapter, we presented the LES microscale dispersion model implemented to reproduce one specific trial (2681829) of the MUST field experiment, which corresponds to neutral conditions. An important focus is placed on the definition of the inlet boundary condition, as this is what controls the atmospheric forcing, which is at the core of the general problem of this thesis of quantifying and reducing uncertainty in LES predictions induced by uncertainty in meteorological conditions. The main contribution of this chapter is the implementation of a turbulence injection approach to reproduce the microscale fluctuations of the atmospheric boundary layer. We verify the sanity of the method using free-field simulations, which is a critical verification step often neglected in the construction of LES models of ABL. The analysis of the result shows that, despite a rather simplistic turbulence injection method, the mean and fluctuation velocity statistics imposed at the domain inlet are established correctly and do not deviate significantly as they propagate within the computational domain. The method is therefore able to represent the ideal case of an atmospheric boundary layer at equilibrium.

We make extensive use of the LES model of the MUST case thus constructed throughout the thesis:

- in Chapter III, we validate the LES wind flow and tracer concentration predictions in light of the various uncertainties involved, and evaluate its main sensitivities,
- in Chapter IV, we build a reduced-order model to emulate the response surface of the LES for different wind boundary condition parameters, allowing us to drastically speed up predictions,
- in Chapter V, we attempt to improve LES predictions by reducing uncertainty on the boundary conditions through the assimilation of actual concentration measurements from the MUST experiment.

Chapter III

Robust model validation under uncertainty

The aim of this chapter is twofold: firstly, to validate the LES model introduced in the previous chapter by comparing its predictions with experimental measurements, and secondly, to quantify the uncertainties involved in order to enlighten the comparison.

Among the different uncertainties involved in LES predictions, a particular effort is placed on the uncertainty due to the microscale internal variability of the atmospheric boundary layer (ABL). To quantify this aleatory uncertainty, we adopt a bootstrap procedure based on sub-average samples. Rigorous verification of the bootstrap assumptions in our case study led us to select the stationary bootstrap algorithm from (Politis and Romano 1994) to account for time dependence between samples. We then provide didactic guidance for selecting the parameters of the algorithm. This approach appears to be particularly well suited to quantify the internal variability inherent to time-averaged statistics collected on a finite time window from LES and experimental data.

After briefly introducing the validation metrics used, the LES model is validated by comparing its prediction of flow and dispersion statistics with experimental measurements from the MUST trial 2681829. We take internal variability into account in the validation to assess model-observation discrepancies, detect potential model biases and quantify the uncertainty on the standard air quality scores obtained.

Finally, we explore the main sensitivities of the LES model to better understand how it can be improved. We analyze how the spatial discretization and the height of the domain influence the model predictions. We also compare the effects of modeling choices, for subgrid-scale models and numerical schemes, and those of large-scale boundary condition parameters. This enables us to target which LES model uncertainties are the most important to take into account in order to improve LES accuracy, in response to the general problematic of the thesis.

Chapter outline

III.1 Introduction 77
III.2 Internal microscale variability quantification 80
III.2.1 Internal variability definition 80
III.2.2 Methods to quantify internal variability 80
III.2.3 General bootstrap method principle 81
III.2.4 Application of a bootstrap approach to MUST predictions and observations 83
III.3 Model validation methodology 89
III.3.1 Wind speed and direction metrics 89
III.3.2 Tracer concentration metrics 90
III.3.3 Comparing predicted fields 91
III.4 LES model validation and microscale internal variability quantification 92
III.4.1 Validation of microscale meteorology statistics 92
III.4.2 Validation of tracer dispersion statistics 94
III.4.3 Validation of the stationary bootstrap method 103
III.5 LES model sensitivity to the main sources of uncertainty . 108
III.5.1 One-at-a-time sensitivity analysis methodology 108
III.5.2 Sensitivity to meteorological boundary conditions parameters . 109
III.5.3 Sensitivity to modeling choices 112
III.6 Conclusion 115

III.1 Introduction

The general aim of this chapter is to validate the LES model built in Chapter II of the MUST trial 2681829, while also quantifying the uncertainties involved. By taking the uncertainties into account, we can represent the predictions of the model in a probabilistic way, in order to better reflect their accuracy. This can also shed light on the origin of discrepancies between the model and observations. For a bibliographical overview of the uncertainties involved in microscale atmospheric CFD models, we refer the reader to the classification proposed in Sect. I.2.2, page 27.

Within the framework of the MUST field campaign, some sources of uncertainty typical of microscale dispersion models are very well controlled and can be expected to have a limited effect on model predictions. In particular, for each experimental trial, the location of the source, the rate and the duration of gas release are precisely identified and reported in Biltoft (2001). In addition, the urban canopy is represented in a simplified manner by an array of shipping containers which limits uncertainty related to the representation of the urban canopy. Still, Santiago et al. (2010) show that taking into account irregularities in the MUST experimental setup (container misalignment, replacement of a container by a van) leads to changes in flow pattern predictions compared with predictions obtained with simplified geometry. Nevertheless, these discrepancies are mainly limited to the building scale, and the level of detail in the urban canopy geometry has no significant effect on the spatially averaged flow properties.

In this thesis, we have decided to set aside these two forms of uncertainty and focus instead on:

i) the uncertainty related to the large-scale atmospheric boundary conditions,

ii) the model structural uncertainties, i.e. those inherent in the solver code,

iii) the aleatory uncertainty due to the microcale internal variability of the ABL.

For the large-scale atmospheric forcing, we especially focus on the wind direction, the friction velocity, and the atmospheric fluctuations. We also consider the uncertainty on the aerodynamic roughness length of the terrain which is used both in the inlet wind profile and in the wall law used for the ground boundary condition. However, we do not consider uncertainty related to the thermal stratification of the atmosphere as the considered trial is under well-established neutral conditions (Sect. II.3.3, page 60), and the LES model has not yet been adapted and validated to handle other stratification conditions. For the model structural uncertainties, we focus on the impact of the subgrid-scale model and numerical scheme choices. In the following, we take a closer look at the effect of microscale internal variability in the MUST field campaign.

Effect of the microscale internal variability in the MUST campaign

Because of the unsteadiness and turbulent state of the ABL, wind and pollutant concentration statistics computed over limited acquisition periods are subject to sampling errors and therefore inherently uncertain. This uncertainty is assumed to explain part of

the discrepancies between CFD model predictions and field measurements (Neophytou et al. 2011; Antonioni et al. 2012; García-Sanchez et al. 2018). A general presentation of this issue is given in Sect. I.2.2, page 27.

The effect of internal variability on the experimental measurements of the MUST campaign is investigated in detail by Schatzmann et al. (2010) (p.88–101). They use moving windows to show that flow and tracer concentration statistics computed over the 200-s analysis period, selected by Yee and Biltoft (2004) to limit the effect of large-scale atmospheric fluctuations, feature significant variability. This aleatory uncertainty affects the measurements used for model validation and for defining the large-scale boundary conditions of CFD models.

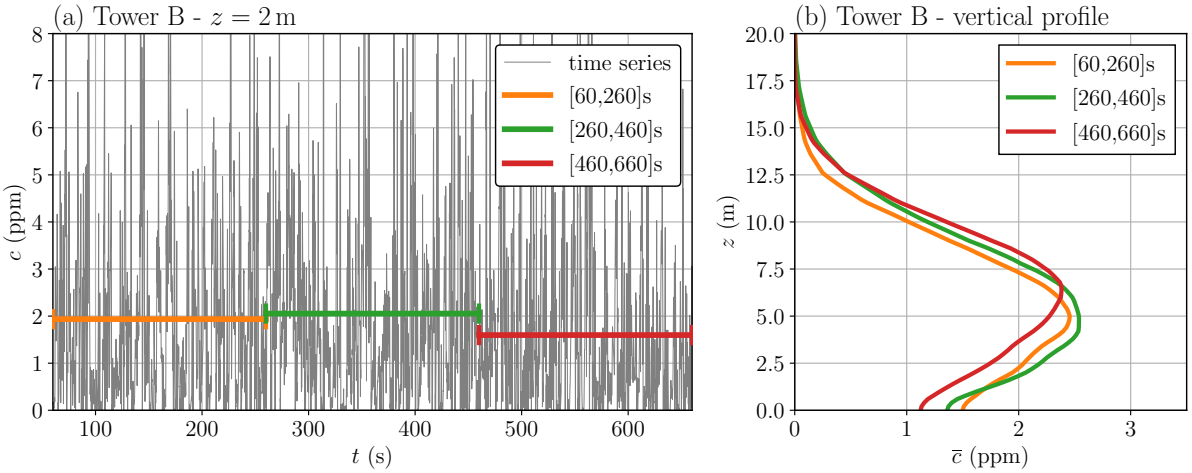


Figure III.1: *Tracer concentration simulated using LES at tower B. (a) Time series at $z = 2\,m$. (b) Time-averaged vertical profiles. Colored lines correspond to three different realizations of the time-averaged concentration obtained with different 200-s averaging windows ([60, 260 s] in orange, [260, 460 s] in green and [460, 660 s] in red).*

The microscale internal variability of the ABL also affects LES predictions, since part of the scales of turbulent fluctuations of the ABL are explicitly resolved by the LES model. As an illustration, Fig. III.1a shows that the predicted mean concentration at a height of $z = 2\,\mathrm{m}$ on tower B significantly changes between three LES estimates obtained over three consecutive 200-second time periods after the spin-up. This variability is not only present near the surface but extends to the whole vertical extent of the plume as illustrated by the vertical profile changes in Fig. III.1b (see Sect. III.4.2 for a more detailed analysis).

Note that increasing the acquisition time reduces the variability of the time-averaged quantities because our LES model has steady boundary conditions and limits the size for the largest eddies. Conversely, increasing the averaging time would increase the experimental data variability because the time scale of the turbulent structures of the ABL significantly exceeds the duration of the release experiment (Schatzmann and Leitl 2011). This large-scale variability is shown in Fig. I.8, page 30. Comparing LES simulations and field measurements for longer acquisition would therefore be inconsistent. This is why

data from the wind tunnel reproduction of the MUST experiment of Bezpalcová (2007) was chosen instead of the field measurements as validation data for the COST 732 model intercomparison exercise (Schatzmann et al. 2010; Di Sabatino et al. 2011). It is indeed possible to obtain statistically converged measurements in wind tunnels by increasing acquisition times (Schatzmann and Leitl 2011).

In this thesis, we make a different choice: use field measurements and quantify their uncertainty associated with internal variability, take advantage of the fact that the LES provides a time-resolved representation of fluctuations to reproduce internal variability in simulation estimates, and finally validate the model in light of the uncertainty involved. The key idea is to be able to define observations and model predictions as envelopes rather than deterministic point values, in order to avoid LES model validation conclusions being dependent on the choice of the 200-s period (Fig. III.1).

Part of the results presented in this chapter is the subject of an article published in the journal Boundary-Layer Meteorology (Lumet et al. (2024). *Assessing the internal variability of large-eddy simulations for microscale pollutant dispersion prediction in an idealized urban environment*).

III.2 Internal microscale variability quantification

In this section, we propose a method for quantifying the effect of the microscale internal variability of the ABL on LES model predictions and experimental measurements. First, the concept of internal variability uncertainty is formally defined in Sect. III.2.1 as sampling noise due to the lack of statistical convergence of acquisitions. We then present in Sect. III.2.2 a brief overview of methods for quantifying it, before focusing on bootstrap methods in Sect. III.2.3, which have the advantage of not requiring additional simulations to quantify internal variability. Finally, Section III.2.4 explains how to apply this approach in the context of microscale atmospheric flow, and in particular how to manage the time dependence of bootstrap samples, the choice of parameters for the chosen algorithm and its application to LES model validation.

III.2.1 Internal variability definition

Let $\overline{Y}$ be the 200-s time-averaged estimation of a given field Y, for example, the mean concentration field. It can be written as the mean of sub-samples averaged over shorter time windows $\breve{Y}_k$:

$$\begin{aligned} \overline{Y} &= \frac{1}{T_{avg}} \int_0^{T_{avg}} Y(t)\mathrm{d}t \ , \\ \overline{Y} &= \frac{1}{N_t} \sum_{k=1}^{N_t} \breve{Y}_k = \frac{1}{N_t} \sum_{k=0}^{N_t-1} \left(\frac{1}{\delta_t} \int_{k\delta_t}^{(k+1)\delta_t} Y(\tau)\mathrm{d}\tau \right), \end{aligned} \tag{III.1}$$

where δ_t is a fraction of the total time-averaging window $T_{avg} = 200\,\mathrm{s}$ such that $N_t = \lfloor T_{avg}/\delta_t \rfloor$ is the corresponding number of sub-samples. It is worth noting that extracting sub-samples over small averaging periods is feasible with an LES simulation, which provides instantaneous realizations of the turbulent phenomena, contrary to other dispersion modeling techniques such as RANS.

Written this way, the time-average $\overline{Y}$ can be seen as the sample estimator of the mean:

$$\mu(\overline{Y}) = \frac{1}{N_t} \sum_{k=1}^{N_t} \breve{Y}_k \ , \tag{III.2}$$

and internal variability corresponds to the variability of $\mu(\overline{Y})$ when the sample of sub-averages $\left\{\breve{Y}_k\right\}_{k=1}^{N_t}$ changes. In this sense, internal variability describes sampling noise error due to limited sample size N_t, i.e. limited acquisition time. The objective of this section is to estimate the variance of the sample mean estimator $\mathbb{V}(\mu)$.

III.2.2 Methods to quantify internal variability

To quantify model internal variability, the most straightforward approach is to run several independent simulations and characterize the variance of the predictions (Costes et al. 2021). However, this is very computationally intensive (each LES estimation costs

about 20 000 CPU hours), and unfeasible for observations because one cannot reproduce 200-s acquisitions with the same atmospheric conditions.

Another method is to apply the central limit theorem that provides a confidence interval for the sample mean estimator $\mu(\overline{Y})$ (Eq. III.2). However, this interval is asymptotic and a large number of realizations of $\mu(\overline{Y})$ is needed for the sample mean to converge in law to a normal distribution, which is not feasible in our case because of the model computational cost.

Alternatively, one could model the statistical distribution of the sub-average samples $\widetilde{Y_k}$ to deduce, either analytically or through Monte Carlo estimation, the distribution of the sample mean $\mu(\overline{Y})$ and hence its variance. For example, the Gamma distribution is well suited for tracer concentration modeling (Cassiani et al. 2020; Orsi et al. 2021). However, this distribution assumption is not always appropriate. For example in our case, the Kolmogorov-Smirnov test (Massey 1951) shows that it is rejected for 4 probes out of 47. More importantly, when Y is a vector, it is difficult to find a statistical distribution that properly accounts for the correlation between its components. Yet, this is essential to propagate internal variability to validation metrics without error compensation (see Sect. III.2.4.5).

To circumvent these issues, it is possible to rely on the empirical distribution of the available sub-average samples instead of assuming a priori their distribution. This is the fundamental principle of Jackknife resampling and bootstrap methods (Efron 1979), which are used in statistics for variance estimation and which are also widely used in climate science for model internal variability estimation (Huybers et al. 2014; Diffenbaugh et al. 2017; Risser et al. 2019; Chan et al. 2020). In our field of interest, Hanna (1989b) used Bootstrap to quantify confidence intervals for air quality model validation metrics; this is for instance implemented in the BOOT statistical model evaluation tool (Chang and Hanna 2005). More recently, Sood et al. (2022) used bootstrap to assess confidence intervals of ABL time-averaged estimates obtained with LES.

In what follows, we present the general bootstrap method and then explain how we adapt it to the context of this study.

III.2.3 General bootstrap method principle

The bootstrap method from Efron (1979) and its variants provide a way to infer the distribution of the sample mean (Eq. III.2) from replicates of the same size of the original sample $\left\{\widetilde{Y_k}\right\}_{k=1}^{N_t}$.

These replicates, called bootstrap samples, are obtained by resampling with replacement of the sub-series values as illustrated in Fig. III.2. The i-th bootstrap sample is formally defined as

$$\left\{\widetilde{Y_{I_k^{(i)}}} \mid I_k^{(i)} \sim \mathcal{U}\left(1, N_t\right), 1 \leq k \leq N_t\right\}, \tag{III.3}$$

where k is the time index of the sub-sample over the 200-s time period so that $I_k^{(i)}$ represents the time index of the sub-sample in the i-th bootstrap sample with i varying between 1 and B (B being the total number of bootstrap samples considered). $I_k^{(i)}$ follows

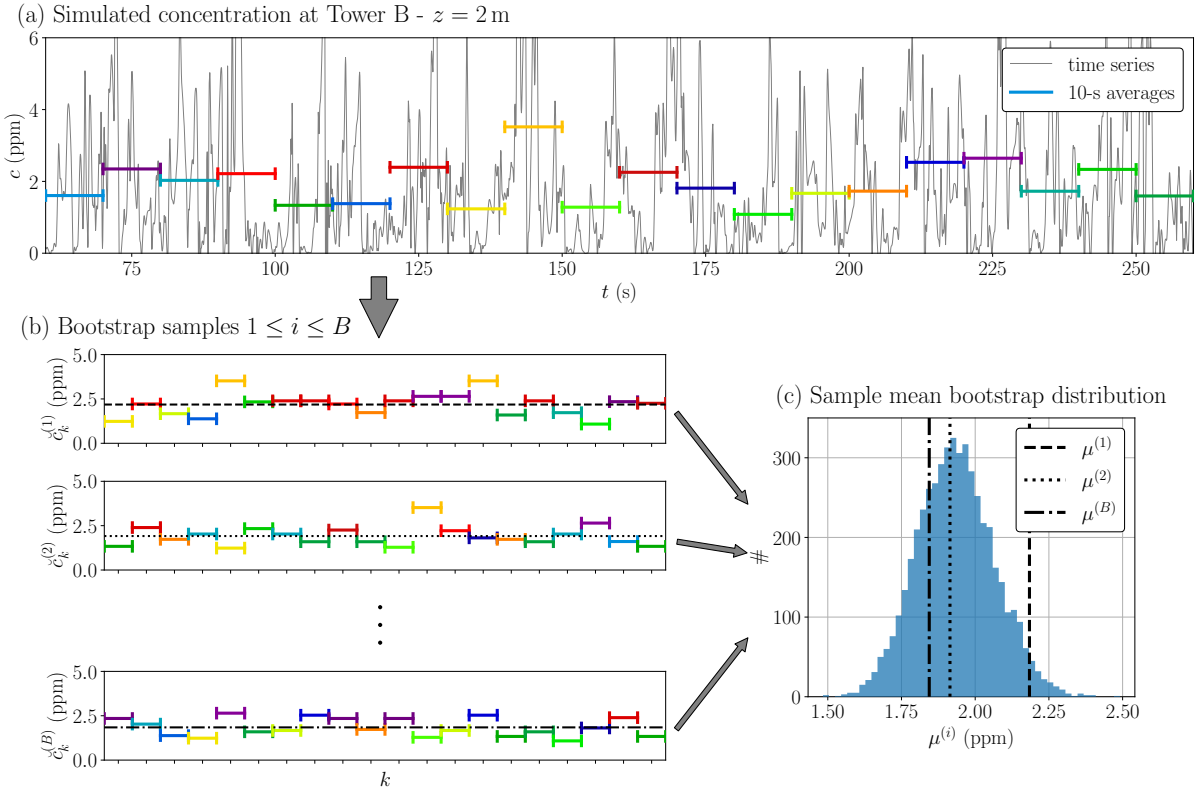


Figure III.2: *Bootstrap principle applied to LES average concentration estimation at tower B ($z = 2$ m). (a) Sub-averages $\breve{c}_k$ over 10 s (color bars) are computed over the 200-s simulated time series (grey solid line). (b) Three examples of bootstrap replicates generated by resampling with repetition from the 10-s sub-averages, their sample means $\{\mu^{(1)}, \mu^{(2)}, \mu^{(B)}\}$ (Eq. III.2) are shown as horizontal lines and illustrate the variability induced by the resampling. (c) The statistical distribution of the sample mean estimator μ is inferred from the B bootstrap replicates. The three examples of bootstrap realizations of time-averages of concentration over 200 s $\{\mu^{(1)}, \mu^{(2)}, \mu^{(B)}\}$ are also represented as vertical lines (c).*

the uniform discrete distribution of the integers between 1 and N_t denoted by $\mathcal{U}(1, N_t)$. The resulting samples can be used to generate an ensemble of sample means :

$$\mu^{(i)}(\overline{Y}) = \frac{1}{N_t} \sum_{k=1}^{N_t} \widetilde{Y_{I_k^{(i)}}}, \quad 1 \leq i \leq B. \tag{III.4}$$

Because of the occurrence of repetitions in the resampling, the bootstrap replicates of the sample mean $\left\{\mu^{(i)}(\overline{Y})\right\}_{i=1}^{B}$ show slight differences, as shown by the vertical dashed lines in Fig. III.2c. This describes the variability of $\mu(\overline{Y})$ due to sampling error, which is precisely the internal variability of the 200-s average with the decomposition in sub-averages we propose in Eq. III.1. Internal variability can thus be quantified in terms of variance as follows:

$$s^2\left(\mu(\overline{Y})\right) = \frac{1}{B-1} \sum_{i=1}^{B} \left(\mu^{(i)}(\overline{Y}) - \widehat{\mu(\overline{Y})} \right)^2, \tag{III.5}$$

with

$$\widehat{\mu(\overline{Y})} = \frac{1}{B}\sum_{i=1}^{B}\mu^{(i)}(\overline{Y}), \tag{III.6}$$

which is an alternative to the original sample mean (Eq. III.2) to estimate the population mean.

The bootstrap method does not only estimate the variance of the sample mean but also describes its complete distribution, as shown by the shaded histogram in Fig. III.2c. We can therefore infer more information about the distribution of μ, in particular, we can construct confidence intervals from the bootstrap samples. Among the different methods reported in the literature (Davison and Hinkley 1997), we use simple confidence intervals based on the percentiles of the empirical bootstrap distribution.

III.2.4 Application of a bootstrap approach to MUST predictions and observations

In this section, we explain how the bootstrap approach is implemented to quantify the effect of internal variability on the LES model predictions as well as on the field measurements. We begin by establishing that the original bootstrap method assumptions are not verified in this context, before presenting the stationary bootstrap from Politis and Romano (1994) that allows us to take into account the time dependence in the sub-averages samples. We then explain how to choose the parameters of this algorithm, and how to apply it to our quantities of interest.

III.2.4.1 On the independence of the samples

The general bootstrap procedure presented in Sect. III.2.3 is based on the assumption that the sub-average samples $\left\{\widetilde{Y_1},\ldots,\widetilde{Y_{N_t}}\right\}$ are independent and identically-distributed (Efron 1979).

In the current study, the assumption of identically distributed samples is ensured for the LES model because it is stationary by construction: first, we use a spin-up period to remove the transient state; then, the inflow boundary conditions are stationary at the scale of the total averaging period of 200 s. However, observations from field campaigns are not necessarily stationary because of mesoscale fluctuations and daily variability in weather conditions (Fig. I.8, page 30). In this regard, the 200-s analysis period for the present case was chosen to minimize the large-scale variability (Yee and Biltoft 2004), and can thus be considered quasi-stationary.

To assess the dependency between the sub-average samples $\left\{\widetilde{Y_k}\right\}_{k=1}^{N_t}$, we use the correlation length λ, defined as the maximum inter-sample distance such that the auto-correlation function $\rho_{\widetilde{Y}}$ is larger than 20%. Figure III.3a shows the auto-correlation of the sub-averages and the corresponding correlation length λ for the concentration at 2-m high at tower D for both LES predictions and observations, using a sub-averaging period of $\delta_t = 10\,\mathrm{s}$ ($N_t = 20$). It shows that observed concentration sub-averages are not independent, with a correlation length of $\lambda_{obs} = 2$. Moreover, it appears to be the case

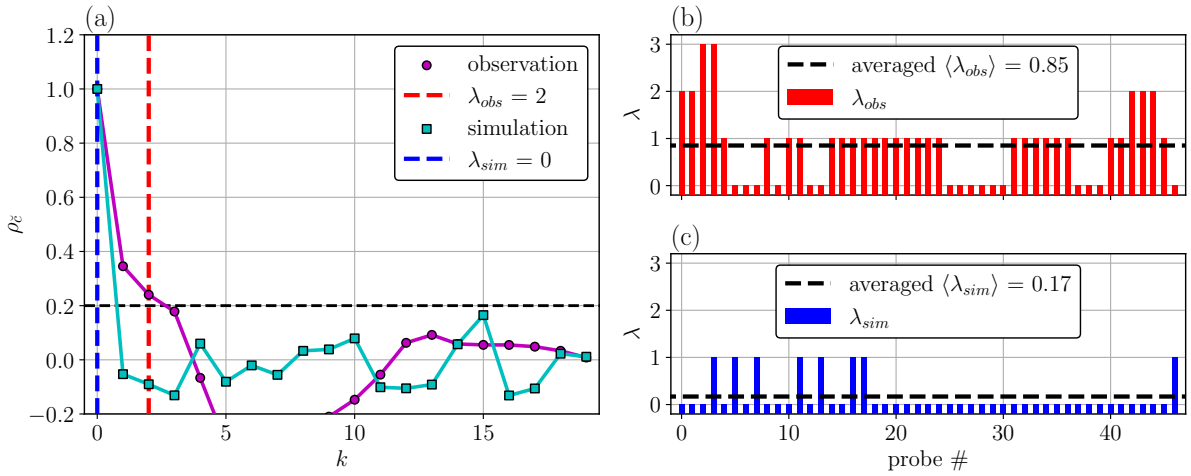


Figure III.3: *(a) Example of auto-correlation function vs. discrete time-lag k of the concentration sub-averages over 10 s, for both measured and simulated concentration at tower D at $z = 2$ m. Vertical dashed lines correspond to the correlation length (in red for measurements and in blue for simulations). (b,c) Correlation lengths computed at every probe location for measurements (in red) and simulation sub-averages (in blue). Horizontal black dashed lines correspond to the averaged correlation length over all the probes*

for the majority of the probes over the detection threshold (Fig. III.3b). Note that LES tends to underestimate the correlation of the concentration sub-averages compared to the measurements (Fig. III.3c). This is because the size of the largest eddies in the LES setup is limited by the size of the computational domain, as explained in Sect. II.4.3, page 64, thus limiting long-term correlations related to large-scale fluctuations. The fact that sample independence is not verified should not be overlooked, as it yields internal variability underestimation when using the standard bootstrap, as shown in Fig. III.5.

To deal with sample dependency, several methods for stationary weakly dependent samples are reported in the literature such as block bootstrap (Carlstein 1986), moving block bootstrap (Kunsch 1989) and stationary bootstrap (Politis and Romano 1994). In this study, we adopt the latter as i) it allows for compromise in the choice of the block length as explained in Sect. III.2.4.3, ii) it does not undersample the first and last sub-samples, and iii) it ensures that bootstrap replicates remain stationary, unlike the other methods mentioned (Politis and Romano 1994). Note that the assumption of weak dependency is verified in the current study as the correlation between sub-average samples rapidly tends to zero for both simulation and observations (Fig. III.3a). This is the case at every probe location since the estimated correlation lengths are always small compared to the number of sub-average samples N_t (Fig. III.3b, c). The stationary bootstrap method is presented in the following section.

III.2.4.2 Stationary bootstrap principle

As in block bootstrap methods, the stationary bootstrap replicates are generated by drawing blocks of samples instead of individual samples, which enables taking into

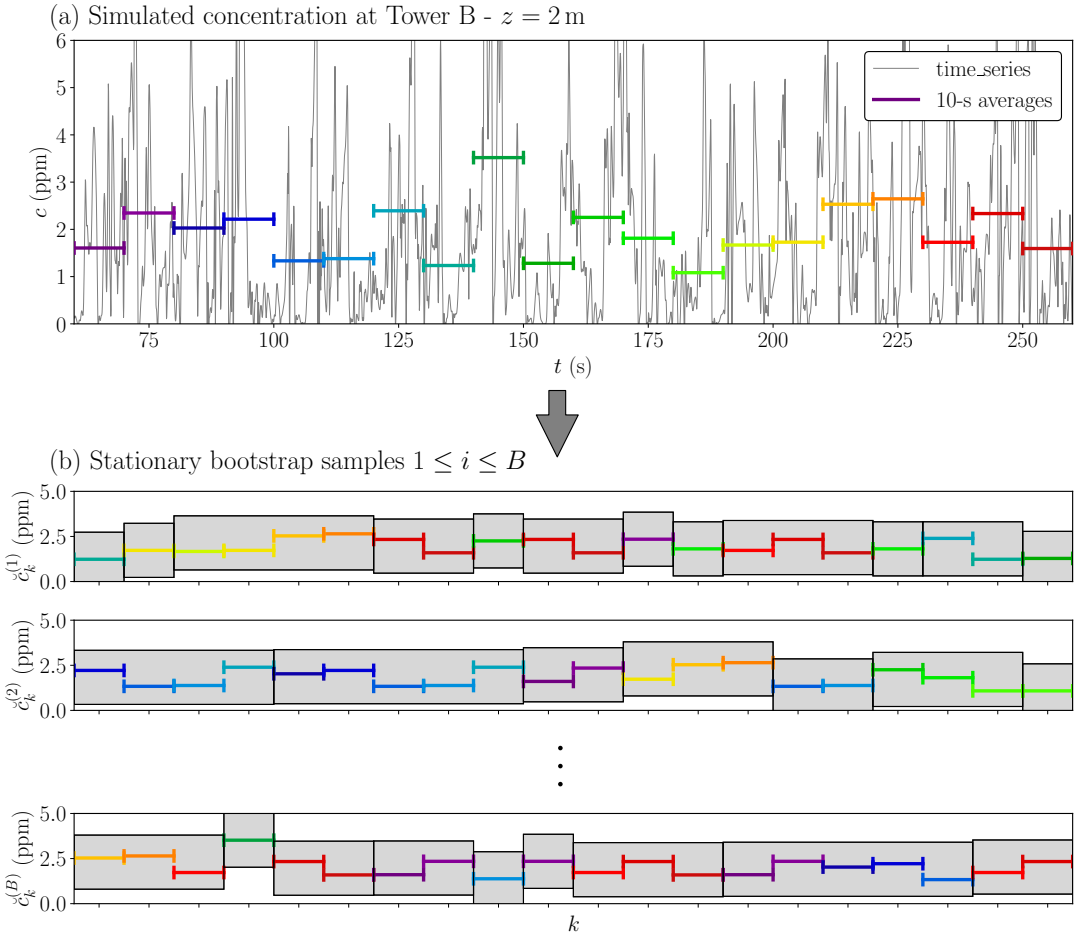


Figure III.4: *Stationary bootstrap algorithm applied to LES average concentration estimation at tower B ($z = 2\,m$). (a) Sub-averages $\check{c}_k$ over $10\,s$ (color bars) are computed over the 200-s simulated time series (grey solid line). (b) Three examples of stationary bootstrap replicates generated by resampling of blocks of 10-s sub-averages. In this example, the mean expected block length is 2.5.*

account the correlation between neighboring samples. The number of sub-averages in each block is randomly selected according to a geometrical law, implying that not all blocks are the same size, as shown in Fig. III.4b. The mean block length ℓ is an input parameter of the stationary bootstrap algorithm. The bootstrap replicates thus obtained can then be used to characterize the statistical distribution of the sample mean in the same way as with the standard bootstrap procedure (Sect. III.2.3). The disadvantage of stationary bootstrap is that it is very sensitive to the choice of the mean block length as shown in Fig. III.5. The next section explains how to choose this parameter in concrete applications.

III.2.4.3 Stationary bootstrap parameters selection

The stationary bootstrap strongly depends on three parameters: the mean block length ℓ, the number of bootstrap replicates B, and the original sample size N_t.

Mean block length selection In practice, the choice of ℓ results from a careful trade-off as using larger blocks reduces the number of samples within each bootstrap replicate: too few samples often result in internal variability underestimation (Davison and Hinkley 1997; Scheiner and Gurevitch 2001), however, using shorter blocks may also lead to internal variability underestimation as it implies neglecting sample dependency (Fig. III.5). In the limit case $\ell = 1$, stationary bootstrap is equivalent to the standard bootstrap. In this study, we define the mean block length as the averaged value of the correlation length over every probe location, i.e. $\ell = \langle\lambda\rangle + 1$, as done by Diffenbaugh et al. (2017). This approach leads to $\ell_{sim} = 1.17$ and $\ell_{obs} = 1.85$ for the mean concentration $\overline{c}$ (Table III.1).

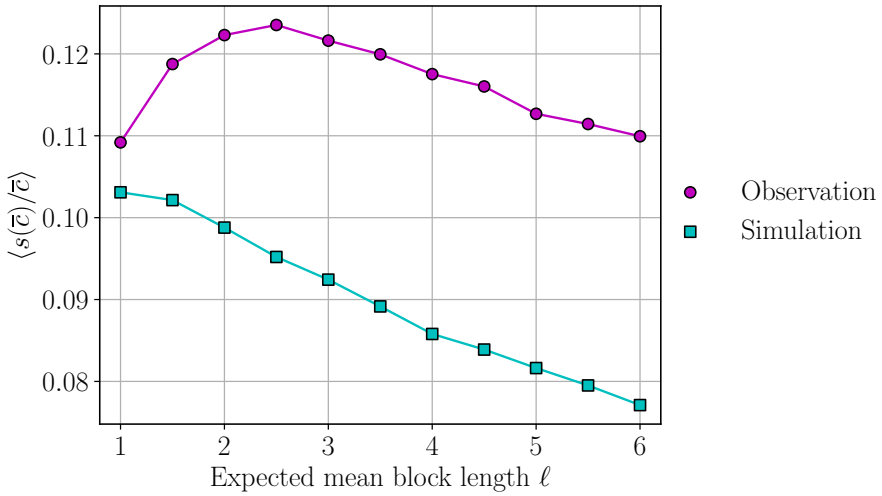


Figure III.5: *Relative standard deviations of the concentration mean over 200 s estimated with stationary bootstrap and averaged over every probe, for different mean block length ℓ. Results for observations and simulations are indicated in cyan and magenta, respectively.*

Note that a compromise is made since a single value of ℓ is used for the whole vector of concentration measurements (and its model counterpart), while the probes have different correlation lengths (Fig. III.3b, c). Indeed, to propagate the internal variability to the validation scores (see Sect. III.2.4.5), the same stationary bootstrap resampling must be used for every probe at once in order to preserve the spatial correlations between them. Otherwise, the variability of the validation metrics would be underestimated because of error compensation.

Number of bootstrap replicates It is important to use enough bootstrap replicates to avoid sampling noise in the bootstrap estimates (Eqs. III.6–III.5). Note that the minimum number of required replicates depends on the target statistical moment from the estimator μ. Here we are mainly interested in 95% confidence intervals, which require

larger B than first-order moments (Davison and Hinkley (1997) suggested $B \geq 1\,000$). A convergence test showed that $B = 5\,000$ was an appropriate value for the current study (Sect. III.4.3.1).

Number of sub-average samples The sample size N_t depends on the physical context and also on the statistical moment of the underlying distribution that we aim to infer. In this study, our objective is to characterize the mean estimator (Eq. III.2), which does not need as large N_t as the variance estimator or the median estimator (Davison and Hinkley 1997). However, a too small value for N_t would result in too short confidence intervals (Davison and Hinkley 1997; Scheiner and Gurevitch 2001) and hence internal variability underestimation. Since the original samples are sub-averages (Eq. III.1), one could simply reduce the averaging window δ_t to increase the number of sub-average samples N_t. However, it increases the dependency between samples which thus does not bring additional information on the underlying distribution as shown in Sect. III.4.3.2. In this study, based on this compromise, we retain sub-averages computed over 10 s, which yields $N_t = 20$.

It is possible to increase the number of sub-average samples by increasing the duration of the simulations to generate more samples to inform on internal variability. We have compared bootstrap estimates obtained with sub-averages from 200-s, 400-s and 600-s simulations (Sect. III.4.3.3). Results show that the acquisition of 200 s is enough to assess the variability of the time-averaged quantities over 200 s. It is therefore not necessary to run extended LES simulations to properly estimate microscale internal variability for the considered MUST trial.

III.2.4.4 Application to statistics of interest

In this study, the stationary bootstrap method is used to assess the variability of the average concentration $\overline{c}$ but also of the average wind horizontal velocity (in terms of amplitude $\overline{u_h}$ and direction $\overline{\alpha}$). As the richer description of LES provides access to higher-order statistics beyond mean predictions, the stationary bootstrap method is also applied to assess the variability of concentration fluctuation $\overline{c'^2} = \overline{(c - \overline{c})^2}$ and flow turbulent kinetic energy $k = \frac{1}{2}(\overline{u_x'^2} + \overline{u_y'^2} + \overline{u_z'^2})$. However, the fluctuations of a quantity over a given averaging period are not equal to the average of its fluctuations over shorter periods. This implies that the decomposition in Eq. III.1 does not hold for second-order statistics. To overcome this issue, one can use a bootstrap sample of both the quantity of interest and its squared value to draw the fluctuation bootstrap distribution:

$$\mu(\,\overline{c'^2}\,)^{(i)} = \mu(\,\overline{c^2}\,)^{(i)} - \left[\mu(\,\overline{c}\,)^{(i)}\right]^2, \; 1 \leq i \leq B, \tag{III.7}$$

where $\mu(\overline{Y})^{(i)}$ is the i-th bootstrap sample mean estimator of the quantity Y as defined in Eq. III.4. Note that the sub-averages used to compute $\mu(\overline{c})^{(i)}$ and $\mu(\overline{c^2})^{(i)}$ must come from the same bootstrap resampling of the original sample. The bootstrap samples of the turbulent kinetic energy estimator are computed similarly. With these samples, the variability can be described by sample variance (Eq. III.5) or percentile confidence intervals.

Table III.1: *Mean block lengths ℓ used for stationary bootstrap applied to the average concentration $\overline{c}$, root mean square concentration c_{rms}, amplitude $\overline{u_h}$ and direction $\overline{\alpha}$ of the mean horizontal wind vector and turbulent kinetic energy k. Values are determined for both simulated and observed data.*

	$\ell(\overline{c})$	$\ell(c_{rms})$	$\ell(\overline{u_h})$	$\ell(\overline{\alpha})$	$\ell(k)$
Simulation	1.17	1.17	1.15	1.15	1.15
Observation	1.85	1.85	3.38	3.38	3.38

Table III.1 summarizes the mean block length ℓ used in this work for all these quantities. Block lengths for observations are larger than for simulations because LES quantities are less temporally correlated (as shown in Fig. III.3b, c). In addition, larger block lengths are obtained for flow-related variables than for concentration, because the wind measurement samples have relatively more data acquired at high altitudes, where temporal correlations are expected to be larger.

III.2.4.5 Internal variability propagation to validation metrics

As the main objective of this study is to validate our LES modeling approach, we also apply bootstrap to infer the distribution of the validation metrics (Sect. III.3). The validation metrics will therefore be characterized by a range of variability rather than a scalar value.

We explain here how the bootstrap approach can be applied to any metric f which quantifies how close two vectors $\overline{Y}_1, \overline{Y}_2$ of time-averaged values are:

$$\begin{aligned} f : \mathbb{R}^N \times \mathbb{R}^N &\longrightarrow \mathbb{R} \\ (\overline{Y}_1, \overline{Y}_2) &\longmapsto f(\overline{Y}_1, \overline{Y}_2). \end{aligned}$$

The two compared vectors are typically prediction and observation values of a given quantity at N different probe locations or two estimates of one field provided by two different simulations. Replicates of the metric score $f(\overline{Y}_1, \overline{Y}_2)$ given the internal variability of $\overline{Y}_1$ and $\overline{Y}_2$ are computed using two independent set of bootstrap replicates (Eq. III.4) of these vectors:

$$\mu(f)^{(i)} = f\left(\mu(\overline{Y}_1)^{(i)},\ \mu(\overline{Y}_2)^{(i)}\right),\ 1 \leq i \leq B. \tag{III.8}$$

The bootstrap replicates thus obtained can then be used to infer the whole distribution of the metrics score under internal variability, as in Sect. III.2.3.

III.3 Model validation methodology

III.3.1 Wind speed and direction metrics

To assess the ability of the model to quantitatively predict the flow field within and over the canopy, three metrics based on time-averaged quantities are used to quantify the difference between model predictions and flow measurements. The hit rate (q) and the mean absolute error (MAE) evaluate discrepancies for the horizontal flow velocity u_h while the scaled averaged angle (SAA) quantifies the deviations for the horizontal direction of the flow $\overline{\alpha}$:

$$\mathrm{q} = \frac{1}{N_{obs}} \sum_{k=1}^{N_{obs}} \xi_k \text{ with } \xi_k = \begin{cases} 1 \text{ if } \left| \overline{u_{hp}}^{(k)} - \overline{u_{ho}}^{(k)} \right| \leq AD \\ 1 \text{ if } \dfrac{\left| \overline{u_{hp}}^{(k)} - \overline{u_{ho}}^{(k)} \right|}{\left| \overline{u_{ho}}^{(k)} \right|} \leq RD \\ 0 \text{ else,} \end{cases} , \tag{III.9}$$

$$\mathrm{MAE} = \left\langle \ \left| \overline{u_{hp}} - \overline{u_{ho}} \right| \ \right\rangle, \tag{III.10}$$

$$\mathrm{SAA} = \frac{\left\langle \ \overline{u_{hp}} \left| \overline{\alpha}_p - \overline{\alpha}_o \right| \ \right\rangle}{\left\langle \ \overline{u_{hp}} \ \right\rangle}, \tag{III.11}$$

where $\overline{u_{ho}}$ and $\overline{\alpha}_o$ are the observed time-averaged horizontal wind speed and direction, and $\overline{u_{hp}}$ and $\overline{\alpha}_p$ are the model colocated predictions. Each element of the N_{obs} dataset is indexed by the superscript (k) in Eq. III.9, while the angle brackets $\langle \cdot \rangle$ indicate the average over the N_{obs} elements in Eqs. III.10–III.11. To compute the hit rate (Eq. III.9), we use the same values of absolute deviation (AD) and relative deviation (RD) as Nagel et al. (2022), i.e. $AD = 1\,\mathrm{m\,s^{-1}}$ and $RD = 0$.

The hit rate and SAA metrics have been used in other MUST modeling validation studies (Santiago et al. 2010; Nagel et al. 2022) and we use in addition the MAE following recommendations by Santiago et al. (2010). The perfect scores associated with these metrics are reported in Table III.2, page 95. Note that for the SAA (Eq. III.11), the wind directions deviations are normalized by the wind magnitude because i) the wind angle evaluation becomes ill-posed when velocities are low, and ii) wind directions associated with the highest velocities are the most important for predicting tracer transport (Calhoun et al. 2004).

The metrics in Eqs. III.9–III.11 are evaluated on the full set of WDTC sonic anemometer measurements, which are located on the towers S, T and N as well as on the four masts V (Sect. II.3.2, page 57). Note that the first anemometer of the tower N downstream of the containers (located at $z = 4\,\mathrm{m}$) was excluded because of its failure during the trial. The total number of measurements for LES model validation for flow prediction is, therefore, $N_{obs} = 13$. The accuracy of the wind flow estimates is only assessed for the horizontal velocity because most of the experimental measurements were provided by 2-D anemometers.

III.3.2 Tracer concentration metrics

LES model performance for tracer concentration prediction (in ppm) is evaluated using the standard statistical metrics for air quality model evaluation (Chang and Hanna 2004), which were also used in previous MUST studies with CFD modeling approaches (Milliez and Carissimo 2007; Antonioni et al. 2012; Nagel et al. 2022). These metrics compare the simulated and observed tracer concentrations in terms of fractional bias (FB), normalized mean square error (NMSE), the fraction of predictions within a factor of two of observations (FAC2), geometric mean bias (MG), and geometric variance (VG):

$$\mathrm{FB} = \frac{\langle \overline{c}_o \rangle - \langle \overline{c}_p \rangle}{\frac{1}{2}\left(\langle \overline{c}_o \rangle + \langle \overline{c}_p \rangle \right)}, \tag{III.12}$$

$$\mathrm{NMSE} = \frac{\left\langle \left(\overline{c}_o - \overline{c}_p \right)^2 \right\rangle}{\langle \overline{c}_o \rangle \, \langle \overline{c}_p \rangle}, \tag{III.13}$$

$$\mathrm{FAC2} = \frac{1}{N_{obs}} \sum_{k=1}^{N_{obs}} \xi_k \text{ with } \xi_k = \begin{cases} 1 \text{ if } 0.5 \leq \overline{c}_p^{(k)} \,/\, \overline{c}_o^{(k)} \leq 2, \\ 1 \text{ if } \overline{c}_p^{(k)} \leq c_t \text{ and } \overline{c}_o^{(k)} \leq c_t, \\ 0 \text{ else,} \end{cases} \tag{III.14}$$

$$\mathrm{MG} = \exp\left(\langle \ln \widetilde{c}_o \rangle - \langle \ln \widetilde{c}_p \rangle \right), \tag{III.15}$$

$$\mathrm{VG} = \exp\left(\left\langle \left(\ln \widetilde{c}_o - \ln \widetilde{c}_p \right)^2 \right\rangle \right), \tag{III.16}$$

where $\overline{c}_o$ is the observed time-averaged concentration, $\overline{c}_p$ is the simulated counterpart, and c_t is the concentration sensor threshold. The tilde indicates that a threshold is applied to the concentration, i.e. $\widetilde{c} = \max(\overline{c},\, c_t)$. This concentration transformation suggested by Chang and Hanna (2004) and also by Schatzmann et al. (2010) avoids issues with the MG and VG metrics, which are not defined for a zero-value concentration and are extremely sensitive to very low values.

FAC2, MG and FB measure the systematic bias in the model predictions; NMSE and VG measure the mean relative scatter of the data. The scores that a perfect model would obtain are reported in Table III.3, page 101. FB and NMSE are more sensitive to errors on high concentrations, while MG and VG emphasize the ability of the model to predict low concentrations.

For this analysis, the observation data are made of the tracer concentration measurements at the 40 DPID sensors located at $z = 1.6$ m throughout the array of containers, at the 8 DPID sensors mounted on the tower T as well as at the 24 UVIC sensors mounted on the towers A, B, C and D (Fig. II.3, page 59). The threshold c_t used to estimate the MG and VG metrics are taken as the instrument detection threshold (0.04 ppm for DPIDs and 0.01 ppm for UVICs as detailed in Sect. II.3.2, page 57). The sensors that measure a time-averaged concentration below this threshold (over the [300; 500 s] analysis period) are excluded from the metrics estimations. This implies that only $N_{obs} = 47$ out of 72 concentration sensor measurements are used in the validation process in this study.

III.3.3 Comparing predicted fields

In this thesis, we not only seek to validate the model against observations, but also to compare different simulations, whether for sensitivity analysis (Sect III.5.1), or for reduced-order model validation (Chapter IV), or data assimilation system evaluation (Chapter V). To do so, we use the metrics previously introduced for the wind flow (Sect. III.3.1) and the tracer concentration (Sect. III.3.2). Indeed, these metrics can be used to assess the differences between two discretized fields by replacing the vectors $(\mathbf{y}_o, \mathbf{y}_p)$ of observed and predicted values by two predicted fields $(\mathbf{y}_p^{(1)}, \mathbf{y}_p^{(2)})$ in Eqs. III.9–III.16. Note that the measurement threshold used in Eqs. III.14–III.16 is not relevant for comparing two simulation predictions, and we replace it with a concentration threshold of $c_t = 10^{-4}$ ppm, which represents the minimal resolution in the numerical predictions.

Since we use an unstructured mesh to discretize the computational domain, the nodes at which the fields are expressed do not all occupy the same volume. And these volume differences can be very significant (see examples in Table A.1, page 232). For spatial averaging, we therefore weight by the dual volume of the nodes $\omega(\mathbf{x})$:

$$\langle f \rangle = \frac{\sum_{k=1}^{N} \omega(\mathbf{x}_k) f(\mathbf{x}_k)}{\sum_{k=1}^{N} \omega(\mathbf{x}_k)}. \tag{III.17}$$

Note that spatial averaging is involved in all the metrics presented so far (Eqs. III.9–III.16).

For this specific purpose of field comparison, we also use the Figure of Merit in Space (FMS) from Chang and Hanna (2004):

$$\mathrm{FMS}(c_\ell) = \frac{\Omega_\cap(c_\ell)}{\Omega_\cup(c_\ell)}, \tag{III.18}$$

where $\Omega_\cap(c_\ell)$ denotes the volume, in m^3, of the domain in which both $\overline{\mathbf{c}}_p^{(1)}$ and $\overline{\mathbf{c}}_p^{(2)}$ are over a user-specified tracer value c_ℓ. In the same way $\Omega_\cup(c_\ell)$ corresponds to the volume where $\overline{\mathbf{c}}_p^{(1)} \geq c_\ell$ or $\overline{\mathbf{c}}_p^{(2)} \geq c_\ell$. This metric quantifies how close the two plume shapes are relative to one specific concentration level. Note that we do not use it for model validation against observations because the observation network is not sufficiently dense to perform accurate shape comparisons.

III.4 LES model validation and microscale internal variability quantification

In this section, the LES model presented in Chapter II is validated against MUST field trial 2681829 measurements for both microscale wind flow statistics (Sect. III.4.1) and tracer plume-related quantities for the 200-s analysis period (Sect. III.4.2). The impact of the internal variability of the ABL on these quantities is quantified using the stationary bootstrap approach from Sect. III.2.3. The same procedure is applied to both the experimental measurements and the LES field estimates, only the mean block length used in the stationary bootstrap differs (see Table III.1). Then we demonstrate the impact of the estimated internal variability on the model validation.

The number of bootstrap replicates used is $B = 5\,000$; the samples are composed of $N_t = 20$ sub-averages over 10 s. Convergence tests and validation of the bootstrap procedure are given in Sect. III.4.3. For the stationary bootstrap, we use the algorithm implemented in the Python module Recombinator[1].

III.4.1 Validation of microscale meteorology statistics

The accuracy of the LES model is assessed in terms of prediction of mean horizontal wind velocity $\overline{u_h}$, direction $\overline{\alpha}$, and wind turbulent kinetic energy k. These quantities are key features for the prediction of the plume dispersion within and above the container canopy, as they control the tracer advection by the mean flow and the turbulent dispersion process.

III.4.1.1 Wind flow vertical profiles

Figure III.6 shows the vertical profiles of $\overline{u_h}$, $\overline{\alpha}$ and k obtained with LES at tower T (this tower location is indicated in Fig. II.3, page 59). On the one hand, results show very good agreement with the measurements of the sonic anemometers for the mean horizontal velocity and direction. The flow deceleration induced by the urban canopy compared to the inlet profile (Fig. II.5, page 64) is well reproduced. However, the model slightly overestimates the flow deviation towards the negative angles, especially at $z = 4$ m and 8 m. This might be explained by the fact that a container was replaced by a van in the vicinity of tower T during the field experiment (Biltoft 2001). On the other hand, the turbulent kinetic energy profile shows that the peak of fluctuations just above the containers is well estimated, whereas the model underestimates the turbulent kinetic energy as altitude increases. The reason for this discrepancy is twofold: i) the synthetic turbulence injection cuts off turbulence length scales larger than the domain scale and ii) the internal region of the boundary-layer flow is known to be unaffected by the finite vertical extent of the domain up to 0.2 time the height of the computational domain (Calaf et al. 2011). In this case, it corresponds to a height of 10 m; above this level, the vertical turbulent transport and other turbulent statistics start to be affected by the top boundary layer which imposes zero vertical turbulent transport. An additional test

[1] See `https://pypi.org/project/recombinator/`

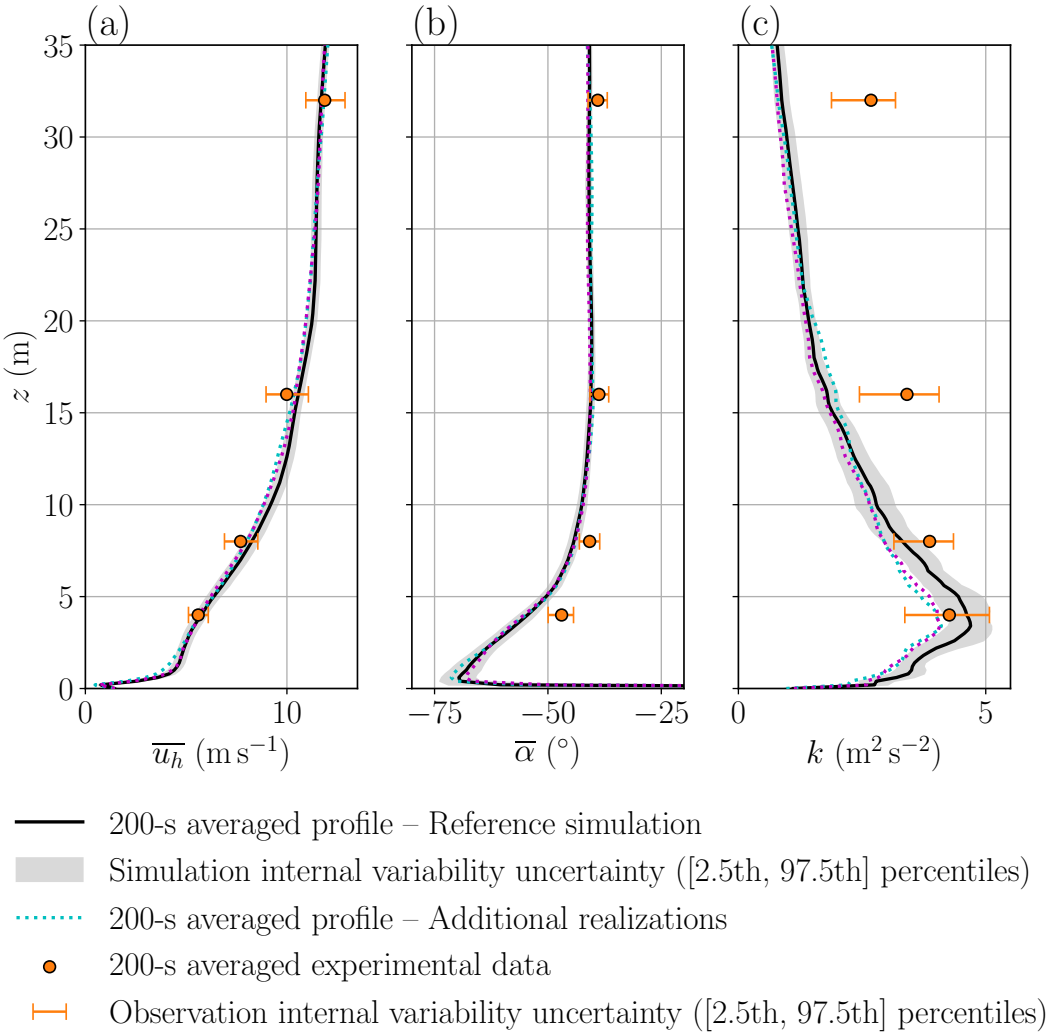


Figure III.6: *Vertical profiles of (a) mean horizontal wind velocity, (b) mean wind direction, and (c) turbulent kinetic energy k at the central tower T (Fig. II.3, page 59). Available experimental data are represented by circles, and black solid lines correspond to the LES time-averaged profiles (two additional realizations of LES estimations over 200 s are also represented as colored dotted lines). Shaded grey areas correspond to the uncertainty induced by LES internal variability and are estimated by stationary bootstrap (the counterpart for the experimental data is indicated as error bars).*

presented in Sect. A.2 demonstrates that increasing the domain height largely solves this issue.

Figure III.6 also shows the 95% confidence intervals corresponding to the bootstrap estimations of the microscale internal variability as profile envelopes for the LES and as error bars for the observations. It is found that the variability is overall quite low for the mean horizontal wind velocity and direction. It is more important for the turbulent kinetic energy but not sufficient to explain the model bias at altitude. Moreover, the LES model tends to significantly underestimate the internal variability of $\overline{u_h}$ and k compared

to that observed. This is attributed to the larger turbulent and mesoscale fluctuations which are not taken into account in the representation of the ABL by the LES model. This is consistent with the results from Nagel et al. (2022) showing that including the mesoscale processes improves the prediction of k at these locations.

Two additional independent LES estimations of time-averaged quantities over 200 s were obtained by extending the original simulation (see cyan and magenta segments in Fig. III.11, page 105). The resulting vertical profiles are shown as colored dotted lines in Fig. III.6. The deviations from the baseline estimate (black solid line) illustrate the effect of internal variability on the time averages. Overall, the estimated envelopes cover well these independent realizations, which supports the plausibility of the stationary bootstrap estimates, except for the turbulent kinetic energy just above the canopy (Fig. III.6c).

III.4.1.2 Quantification of wind flow predictions accuracy

In addition to the profiles at tower T (Fig. III.6), we also compare LES predictions and observations using the wind flow metrics (Sect. III.3.1). Table III.2 presents the scores obtained over the obstacles (towers S, T and N), within the obstacles (masts V) and for all sensors at once. For every set, the hit rate is 100%, which means that the difference between LES estimates and measurements for the wind horizontal velocity is always less than the absolute deviation $AD = 1\,\mathrm{m\,s^{-1}}$ used in Eq. III.9 by Nagel et al. (2022). Indeed, the MAE metric shows a limited level of error for the wind velocity. However, the error is larger for sensors located within the container array, as shown by the higher MAE in this region, and this is even more pronounced for the SAA metric. This is due to the proximity of the masts V to the containers (Fig. II.3, page 59), where there are strong wind direction gradients as explained by Nagel et al. (2022). Still, the overall accuracy of the LES flow estimations is satisfactory.

By computing two bootstrap samples of each measurement and colocated LES estimation, we can obtain an ensemble of metrics realizations as explained in Sect. III.2.4.5, and then quantify how uncertain the model validation scores are, given the internal variability of the system. The resulting standard deviations of the flow validation metrics are given in Table III.2. Results show that the internal variability has a limited effect on velocities, with a standard deviation of MAE of approximately $0.1\,\mathrm{m\,s^{-1}}$. Note that this variability is however larger than the sonic anemometer accuracy (between $0.01\,\mathrm{m\,s^{-1}}$ and $0.05\,\mathrm{m\,s^{-1}}$). In contrast, the variability is less important for the wind direction. Moreover, the effect of variability is rather homogeneous over the different datasets, which is coherent with the vertical distribution of the internal variability envelopes at tower T (Fig. III.6).

III.4.2 Validation of tracer dispersion statistics

III.4.2.1 Mean concentration horizontal and vertical profiles

Model performance is first analyzed in terms of mean concentration horizontal profiles within the container array in Fig. III.7a, b, c, d. At $z = 1.6\,\mathrm{m}$, the model underestimates tracer concentration along the four DPID sensor sampling lines, which could be due to a plume elevation overestimation, as discussed later. Still, the shape of the profiles is

Table III.2: *Mean horizontal wind velocity and direction: comparison between the LES model and the experimental measurements at tower S, T and N, and the sensors at the containers levels on the masts V (Fig. II.3, page 59). The differences are assessed in terms of hit rate q, mean absolute error MAE, and scaled averaged angle SAA defined in Sect. III.3.1. The stationary bootstrap method presented in Sect. III.2.4 is used to estimate standard deviations of the scores.*

	$q(\overline{u_h})$	$MAE(\overline{u_h})$ (m.s^{-1})	SAA (°)
Perfect scores	100%	0	0
Scores - above canopy	$9/9 = 100\%$	0.42	2.02
Standard deviation	0%	0.13	0.48
Scores - inside the canopy	$4/4 = 100\%$	0.65	37.60
Standard deviation	0%	0.15	1.39
Global scores	$13/13 = 100\%$	0.49	8.25
Standard deviation	0%	0.11	0.48

rather well reproduced by the model. The decrease in concentration is consistent with the observations, both in the flow direction (between each line) and in the transverse direction (on a given line). The plume deviation is also well predicted by the model as illustrated by the concentration maximum position.

Model performance is then analyzed along the vertical (above the container array) by comparing the estimated mean concentration vertical profiles with measurements from towers B, T, C, and D (Fig. III.8a, b, c, d). Overall, the LES predictions are in acceptable agreement with the observations. The model tends to overestimate the mean concentration at towers B and C. The same tendency was observed with other LES models by Camelli et al. (2005) and Nagel et al. (2022) for tower B. The model also underestimates the mean concentration at tower D due to a lack of lateral spread of the simulated plume (Fig. III.7a), tower D being far from the plume centerline (Fig. II.3, page 59). In addition, the predicted maximum concentration is located too high above the canopy, especially at tower B (if we disregard the highest sensor that is inconsistent with the others). For tower C, there are not enough sensors at high heights to conclude. This could mean that the predicted plume rises too much, which would explain the near-ground concentration underestimation (Fig. III.7a, b, c, d).

Note that there seems to be an inconsistency between the UVIC measurements from tower C (Fig. III.8c) and those from the fourth line of DPIDs (Fig. III.7d), although they are arranged on the same transverse line. Indeed, the UVIC sensor at tower C at $z = 2\,\mathrm{m}$ measures 0.21 ppm, while the two closest DPID sensors (10 m away) measure 0.54 ppm and 1.10 ppm. This may also concern other UVIC measurements, explaining why LES overestimates concentration at towers B and C.

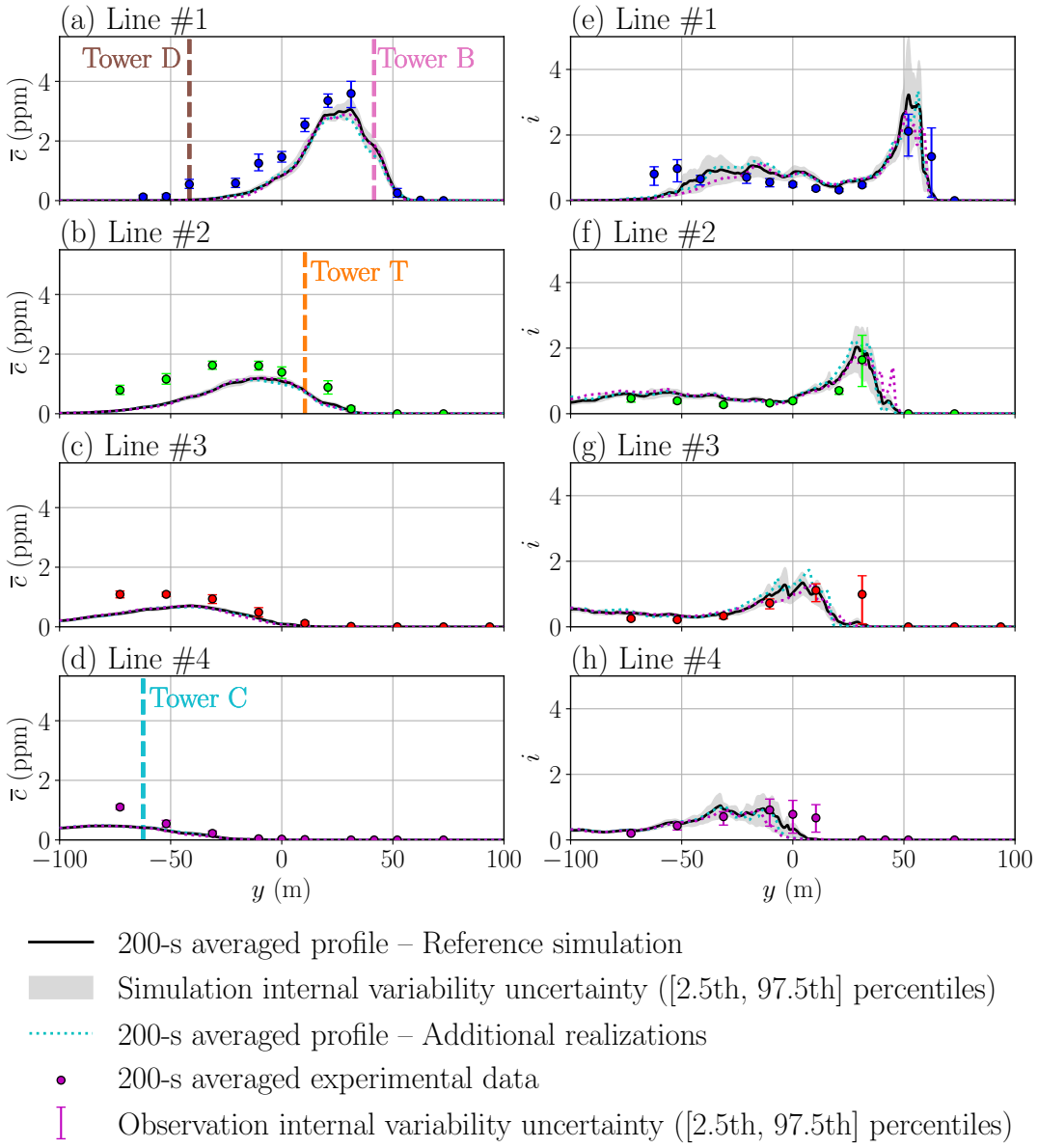


Figure III.7: *Horizontal profiles of average concentration $\bar{c}$ (a, b, c, d) and concentration fluctuation intensity i (e, f, g, h) between the containers at $z = 1.6$ m. The profiles are given for each line of DPID sensors represented with a distinct color in Fig. II.3, page 59. Circles correspond to measurements, black solid lines correspond to simulated time-averaged profiles (two additional realizations of LES estimations over 200 s are also represented as colored dotted lines). Shaded grey areas correspond to the uncertainty induced by LES internal variability and are estimated by stationary bootstrap (the counterpart for the experimental data is indicated as error bars).*

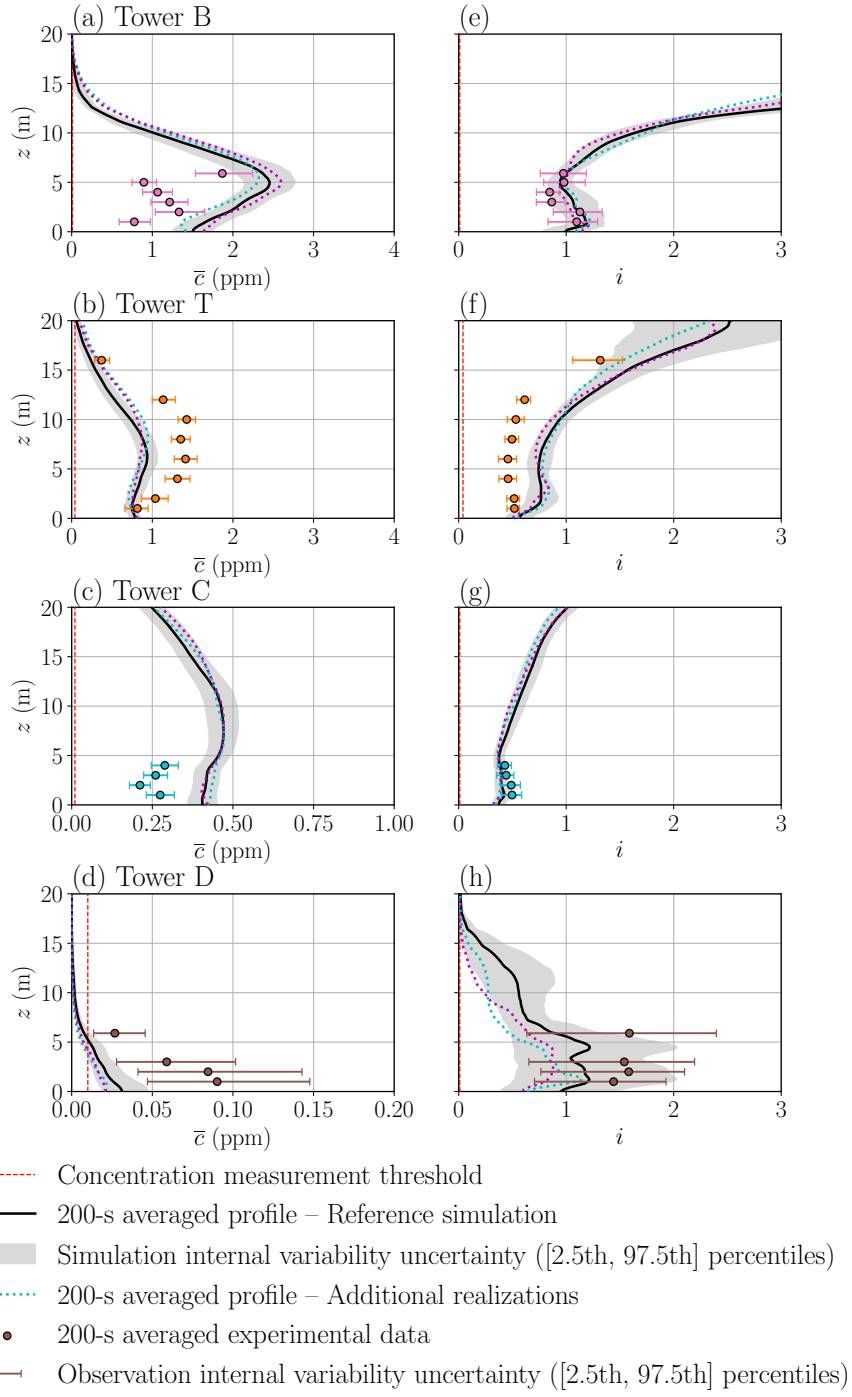


Figure III.8: *Vertical profiles of average concentration $\overline{c}$ (a, b, c, d) and concentration fluctuation intensity i (e, f, g, h) at towers B, T, C and D, respectively (Fig. II.3, page 59). Circles correspond to measurements, black solid lines correspond to simulated time-averaged profiles (two additional realizations of LES estimations over 200 s are also represented as colored dotted lines). Shaded grey areas correspond to the uncertainty induced by LES internal variability and are estimated by stationary bootstrap (the counterpart for the experimental data is indicated as error bars).*

III.4.2.2 Mean concentration internal variability

The internal variability of the mean concentration is shown in Figs. III.7–III.8 with the 95% confidence intervals estimated by stationary bootstrap. This internal variability increases with altitude (Fig. III.8a, b, c, d), because, outside the canopy, the flow statistics are more sensitive to incoming fluctuations from the ABL. LES profile envelopes are generally consistent with the observed variability (Figs. III.7–III.8). Analysis of the relative internal variability aggregated over all sensors (Fig. III.5, page 86) confirms that the LES model can reproduce the observed internal variability overall, with a slight tendency to underestimate it for instance at tower D (even if the relative variability $s(\overline{c})/\overline{c}$ is similar to that of the measurements).

Note that the stationary bootstrap estimations of the internal variability look plausible regarding the two independent LES realizations of 200-s averaged concentration. The bootstrap profile envelopes globally cover these realizations both inside and above the canopy (Figs. III.7–III.8). However, one or both realizations can be locally slightly outside the 95% confidence interval, for example at the concentration peak location (Fig. III.7a) or at high altitude at tower T (Fig. III.8b). This indicates that the internal variability is underestimated there, which is likely caused by an insufficient number of independent sub-average samples N_t, as stated by Davison and Hinkley (1997) and Scheiner and Gurevitch (2001).

Finally, although significant, the internal variability alone does not explain the mismatch between LES estimates and vertical tower measurements (Fig. III.8a, b, c, d). The lack of accuracy comes rather from another source of uncertainty. For instance, Milliez and Carissimo (2007) explain that the vertical profiles are difficult to estimate accurately because of their important sensitivity to the wind direction. This sensitivity is exacerbated in our case at tower B and to a lesser extent at tower T because both towers are located near the steepest edge of the plume where concentration gradients are very large (Fig. III.7a, b, c, d). In these areas, plume position errors have a larger impact on model accuracy than microscale internal variability. A sensitivity test to a deviation of the inlet wind direction is described in Sect. III.5.2 and confirms that it can have an important impact on the estimated vertical concentration profiles (Fig. III.14).

III.4.2.3 Concentration fluctuation intensity

In addition to time-averaged values, LES models provide an explicit temporal resolution of the flow fluctuations. In this section, we propose to further validate the model by examining its ability to predict resolved concentration fluctuations, which are directly accessible from LES data. To characterize concentration fluctuations, we use the fluctuation intensity i as Yee and Biltoft (2004). It reads $i = \sqrt{\overline{c'^2}}/(\overline{c} + c_t)$, where $\overline{c'^2} = \overline{(c - \overline{c})^2}$ is the squared resolved fluctuation of the concentration. The concentration threshold c_t, equal to the detection threshold of the sensors, i.e. 0.01ppm for UVICs or 0.04ppm for DPIDs, is added to the normalization term to avoid ill-posed values for very small concentrations.

There is a very good agreement between LES estimations and observations of fluctuation intensity. Among the containers, the LES model finely reproduces the observed

horizontal distributions (Fig. III.7e, f, g, h), including their asymmetry. However, the fluctuation peak is slightly overestimated for the first line of sensors and lacks horizontal extent for lines #3 and #4. Moreover, the LES model also appears to be very accurate for the estimation of vertical fluctuation profiles (Fig. III.8e, f, g, h), except for the T tower where the model overestimates them but still predicts a consistent profile. By normalizing the fluctuations, we show that, despite being biased for the mean concentration vertical profiles at towers B, C and D, the LES model is still able to reproduce a physically consistent estimation of the concentration second-order statistics.

The internal variability is also estimated for the concentration fluctuation intensity with the bootstrap samples from Eq. III.7. Figures III.7e, f, g and h show that internal variability is very large at the location of the peak fluctuation. Overall, the LES-predicted fluctuation envelopes are in good agreement with the observed variability and the two independent LES runs. Contrary to the average concentrations, and except for tower T, the differences between the model and the observations can be attributed to the internal variability, which is particularly visible for tower D where the fluctuations are very important.

III.4.2.4 Quantification of dispersion predictions accuracy

In the following, the accuracy of the LES model tracer transport is assessed from a more synthetic viewpoint. We also illustrate how the internal variability of the tracer concentration field should be taken into account in this model validation exercise.

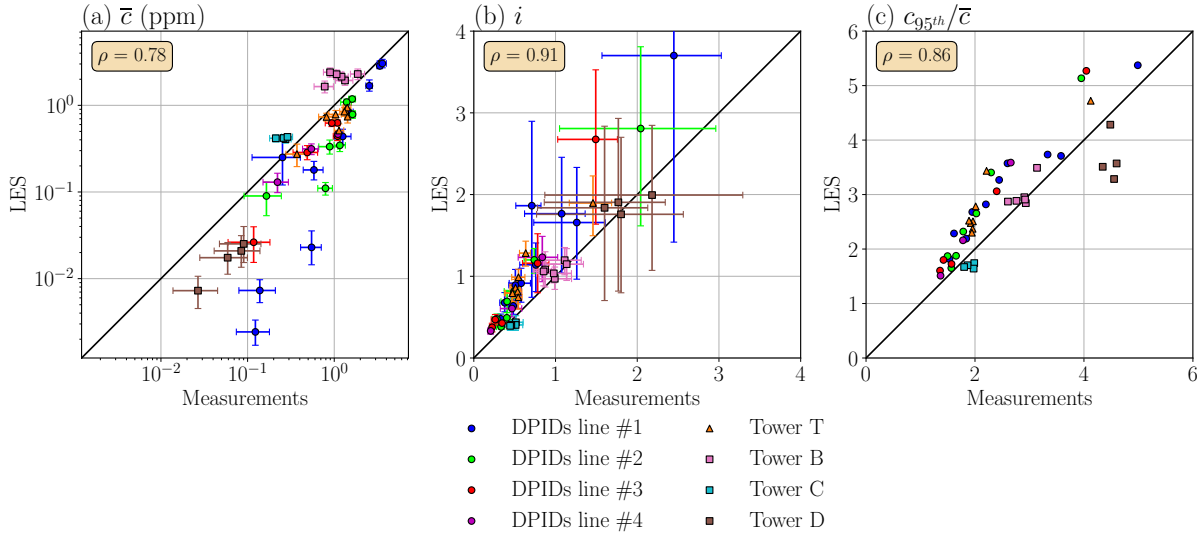


Figure III.9: *Scatter plots of simulated versus measured concentration statistics: (a) temporal mean, (b) fluctuation intensity, (c) normalized 95th percentile at each sensor over the detection threshold. Each type of sensor is represented with the same color as in Fig. II.3, page 59. The correlation coefficient ρ is indicated for each statistic. The error bars represent the 95% confidence intervals estimated with stationary bootstrap and account for the sampling error of both simulated and measured statistics (error bars are not given for the concentration maximum as the bootstrap procedure is not suitable for this statistic).*

Figure III.9 shows the scatter plots of the simulated versus measured concentration main statistics. First, for the averaged concentration, the model estimates are overall consistent with the observations (Fig. III.9a) with a correlation coefficient $\rho = 0.78$. Higher tracer concentration values (above 0.5-1 ppm) are well represented, but the LES model notably underestimates the lower concentration values. The same trend is found for the concentration fluctuations (not shown). However, if we remove the bias on the averages, the LES is able to accurately reproduce the fluctuations with a correlation coefficient $\rho = 0.91$ for the concentration fluctuation intensity (Fig. III.9b). Overall, the model tends to overestimate fluctuation intensities, and the prediction error appears to be greater for larger fluctuations.

In addition, the microscale internal variability is depicted in the scatter plots of the averaged concentration (Fig. III.9a) and fluctuation intensity (Fig. III.9b) with the 95% confidence intervals obtained with bootstrap and depicted as two error bars for each tracer concentration sensor measurement and colocated LES estimation. The internal variability is heterogeneous, with locations for which it is negligible and others for which it is very important. Note that, for most of the points the x-error and y-error bars have similar lengths, which shows that LES estimates well the variability of predicted quantities.

As suggested by Chang and Hanna (2005), we assess if the difference between simulated and observed values is significantly different from zero at the 95% confidence interval. This test is performed for each sensor, and we find that, given the internal variability, the LES model fits only 13% and 45% of the measurements, for the mean concentration and fluctuation intensity, respectively. Although internal variability is high in areas where the model lacks precision (i.e. for the low mean concentrations and high fluctuation intensities), it only explains a limited part of the misfit between simulation estimates and measurements. Therefore most of the model errors, especially for the mean concentration, must come from other sources as suggested by the sensitivity tests presented in Sect. III.5.

Besides, Figure III.9c also shows a fine agreement for the 95th percentile of concentration time series over the 200-s analysis period. The LES model appears to well predict the peak concentrations with a correlation coefficient $\rho = 0.86$. This demonstrates that the LES model is able to represent all the complexity of the dispersion phenomenon and not only mean concentration levels. The effect of internal variability on the peak concentrations is not assessed, since it is not accessible with the bootstrap procedure described in Sect. III.2.3. Nevertheless, the peak concentrations are expected to be subjected to a strong variability, as they correspond to extreme events in the LES realizations of the tracer plume. Quantifying it would therefore be an interesting prospect.

The accuracy of the LES mean concentration estimations is finally evaluated using the standard air quality metrics from Chang and Hanna (2004), following the methodology presented in Sect. III.3. As in previous works (Milliez and Carissimo 2007; Kumar et al. 2015; Nagel et al. 2022), metrics are computed separately for the DPIDs sensors on the horizontal $z = 1.6$ m plane on the one hand, and for the vertical sensors on towers A, B, C, D and T on the other hand. Metrics are then evaluated for all sensors. Results gathered in Table III.3 show an overall good agreement with observations, with only the

Table III.3: *Comparison between the LES model mean concentration prediction and experimental measurements assessed in terms of fractional bias (FB), normalized mean square error (NMSE), fraction of predictions within a factor of two of observations (FAC2), geometric mean bias (MG), and geometric variance (VG). Definitions of these metrics are given in Sect. III.3. They are computed for the horizontal sensors (i.e. the DPID sensors located at $z = 1.6\,m$), the vertical sensors (i.e. towers A, B, C, D and T), and all sensors. Sensors for which the experimental mean concentration is under the detection threshold are excluded. LES results for the same trial reported in the literature are given as an indicative basis.*

	FB	NMSE	FAC2	MG	VG
Perfect scores	0	0	100%	1	1
Acceptable[1]	[-0.3, 0.3]	<4	> 0.5	[0.7 1.3]	<1.6
Horizontal	0.51	0.37	$^{13}/_{25} = 52\%$	2.22	2.62
Vertical	-0.15	0.48	$^{14}/_{22} \approx 64\%$	1.13	1.68
Global	0.20	0.41	$^{27}/_{47} \approx 57\%$	1.62	2.13
Literature (LES)					
Nagel et al. (2022) (3 model configurations)	[0.38, 0.40]	-	[60%, 65%]	[0.86, 1.05]	[1.90, 2.33]

[1] According to Chang and Hanna (2004), note that these guidelines may not be tailored for this microscale context.

MG and VG out of the range of acceptable scores. This seems to indicate that LES models have some difficulty in capturing low tracer concentration values. Except for the MG, the scores obtained are in line with those obtained by Nagel et al. (2022) with another LES model and for the same trial. They are also comparable to the scores obtained with RANS models on a larger number of trials including the present trial 2681829 (Milliez and Carissimo 2007; Donnelly et al. 2009; Kumar et al. 2015).

The LES model appears to be less accurate on the horizontal plane within the canopy than above. In this region, the concentration is overall underestimated by the model (FB > 0 and MG > 1), which is seen in the horizontal profiles (Fig III.7a, b, c, d). Interestingly, the opposite behavior, i.e. better performances on the horizontal than on the vertical, was observed by Nagel et al. (2022). This may indicate that model structural uncertainties are important in this study case, and it would be of interest to perform a multi-model comparison to quantify it.

As explained in Sect. III.2.4.5, the internal variability of the time-averaged concentrations is propagated to the air quality metrics to quantify their uncertainty. The resulting distributions for each metric are summarized with box-and-whisker plots of the bootstrap samples (Fig. III.10). It demonstrates that the scores obtained in this model validation exercise are significantly uncertain. Moreover, it shows that the internal variability of the concentration affects each metric differently. FB and MG are less sensitive to internal variability because of error compensation, while wider spreads are found for NMSE and

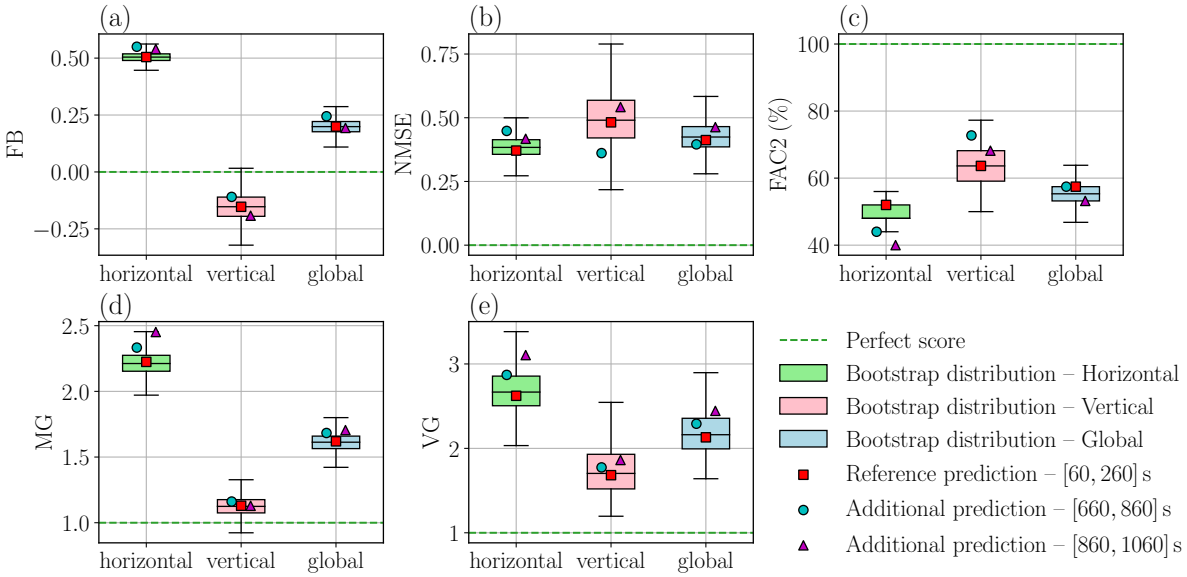


Figure III.10: *Box plots of the air quality metrics distributions obtained by taking into account the internal variability of both simulated and observed data using stationary bootstrap. Point estimations corresponding to the reference and two independent realizations of 200 s simulation are shown as red squares, cyan circles and magenta triangles. Results are given for (a) FB, (b) NMSE, (c) FAC2, (d) MG, and (e) VG metrics (Sect. III.3). These metrics are computed for all tracer concentration sensors over the detection thresholds (blue box), but also for the subsets of the horizontal probes (i.e. the DPID sensors at $z = 1.6$ m) and vertical probes (i.e. towers A, B, C, D and T), respectively represented as green and pink boxes.*

VG. This is because NMSE and VG are quadratic metrics and thus measure the dataset dispersion. Note that FAC2 also shows an important variability as it is a discrete and nonlinear metric computed over a small number of sensors. The effect of the internal variability is also higher on the vertical than on the horizontal, which is consistent with the observed envelopes in Figs. III.7–III.8.

Figure III.10 also shows the validation scores obtained for three independent LES predictions of time-averaged concentration over 200 s. Despite having exactly identical model configurations, the discrepancies between each score are not negligible. The bootstrap estimation of the variability of the metrics, obtained using only the sub-averages of the first simulation ([60, 260] s), explains quite well this variability as only two outliers are not covered by the whiskers (first and third quantiles extended by 1.5 times the inter-quartile range): one for the horizontal FAC2, and one for the horizontal MG.

In summary, we show that, for microscale dispersion experiments with small acquisition times and/or limited analysis periods, validation scores feature a high range of variability. It is thus vital to take this variability into account in a model validation exercise, but also for sensitivity analysis or multi-model comparison, to avoid drawing insignificant conclusions about the trends in the metrics. The bootstrap procedure based on sub-average samples presented in Sect. III.2.4 appears well-suited to answer this need.

III.4.3 Validation of the stationary bootstrap method

In this section, several tests are carried out to assess the convergence and validate the internal variability estimation provided by the stationary bootstrap approach (Sect. III.2.4) applied to the MUST trial.

III.4.3.1 Convergence with the number of bootstrap replicates

With bootstrap methods, such as the stationary bootstrap, bootstrap replicates of one original sample are used to compute Monte-Carlo estimates of statistics of estimators from the original sample. For instance, we estimate in this study variance (Eq. III.5) and confidence interval of the time averages or fluctuations of physical quantities of interest. As a Monte-Carlo method, the convergence of the estimated statistics is in $\mathcal{O}(1/\sqrt{B})$ with B the number of bootstrap replicates.

We assess the convergence for the 2.5th and 97.5th percentiles as it requires more bootstrap replicates than for bias or variance estimation (Davison and Hinkley 1997). Table III.4 shows the evolution of the 2.5th percentile of the mean concentration at tower B at $z = 2\,\mathrm{m}$ and of the model validation metrics evaluated according to the bootstrap procedure (Eq. III.8) for different values of B. The bootstrap estimations of the 2.5th percentiles show some variability for very low numbers of bootstrap replicates (between 100 and 500), but then quickly converge for all the considered quantities. The same analysis was carried out for the 97.5th percentile and gave similar results. We conclude that B = 5 000 bootstrap samples are more than sufficient to achieve convergence. This result is in line with the literature, which recommends between 1 000 and 10 000 replicates (Davison and Hinkley 1997; Chang and Hanna 2005).

Table III.4: *Values of 2.5th percentiles evaluated with stationary bootstrap for different numbers of replicates B. Estimations are given for one example of simulated and observed mean concentration (at tower B at $z = 2\,m$), as well as for the air quality metrics (Sect. III.3.2) and flow validation metrics (Sect. III.3.1)*

B	100	500	1 000	5 000	10 000
$\overline{c}_{obs}$ (ppm)	1.70	1.70	1.72	1.71	1.70
$\overline{c}_{sim}$ (ppm)	1.07	1.05	1.02	1.04	1.04
FB	0.13	0.14	0.14	0.14	0.13
NMSE	0.33	0.32	0.33	0.32	0.32
FAC2	0.49	0.49	0.49	0.49	0.49
MG	1.48	1.48	1.47	1.47	1.47
VG	1.73	1.76	1.77	1.77	1.77
MAE ($\mathrm{m\,s^{-1}}$)	0.37	0.35	0.35	0.34	0.35
SAA (°)	7.78	7.59	7.69	7.65	7.63

Table III.5: *Values of 2.5th percentiles evaluated with stationary bootstrap for different sub-averaging period δ_t. Estimations are given for one example of simulated and observed mean concentration (at tower B at $z = 2\,m$), as well as for the air quality metrics (Sect. III.3.2)*

2.5th percentile	$\delta_t = 10\,\mathrm{s}$	$\delta_t = 5\,\mathrm{s}$
$\overline{c}_{obs}$ (ppm)	1.71	1.63
$\overline{c}_{sim}$ (ppm)	1.04	1.04
FB	0.14	0.13
NMSE	0.32	0.32
FAC2	0.49	0.49
MG	1.47	1.48
VG	1.77	1.80

III.4.3.2 Effect of the sub-averaging period

Since both simulated and measured time series are well sampled in time, it is possible to change the sub-averaging period δ_t to adjust the number of sub-average samples N_t (Eq. III.1). To assess the effect of the sub-averaging period on the internal variability estimation, the estimated percentiles obtained with a stationary bootstrap of sub-averages over $\delta_t = 10\,s$ and $\delta_t = 5\,s$ are compared. By reducing the sub-averaging period, the samples are getting more dependent. It is therefore mandatory to adapt the mean block length parameter ℓ of the stationary bootstrap method. For $\delta_t = 5\,s$, it results to new values of $\ell_{sim} = 1.38$ and $\ell_{obs} = 2.62$ for the time-averaged concentrations. This is consistent since it means that, for more dependent data, the blocks should be larger than the ones used for $\delta_t = 10\,s$ (Table III.1, page 88). Table III.5 shows the 2.5th percentile estimates for the main quantities of interest for the two different values of δ_t. Results indicate that changing the sub-averaging period has a very limited impact on the stationary bootstrap estimations. This is because changing the sub-averaging period only changes the division of the original sample (Eq. III.1) and so does not provide any additional information on the underlying distribution of the time-averaged quantities.

III.4.3.3 Convergence with the number of sub-average samples

As mentioned in Sect. III.2.3, it is essential to have a sufficient number of sub-average samples N_t in the original sample. In particular, too few samples may result in internal variability underestimation. To increase N_t, the LES simulation acquisition time is increased from 200 s to 400 s, and then 600 s (Fig. III.11). With $\delta_t = 10\,\mathrm{s}$ the resulting number of sub-averages is 40, and 60 respectively, against 20 for the reference sample. In any case, the bootstrap replicates are obtained by resampling only 20 sub-averages over the N_t available, even if $N_t = 60$. Indeed, the objective is still to quantify the variability over the 200-s analysis period and not over 600 s. Two additional realizations of 200-s averages (in cyan and magenta in Fig. III.11) are used for validation purposes.

Figure. III.12 shows the effect of the number of sub-average samples on the 95%

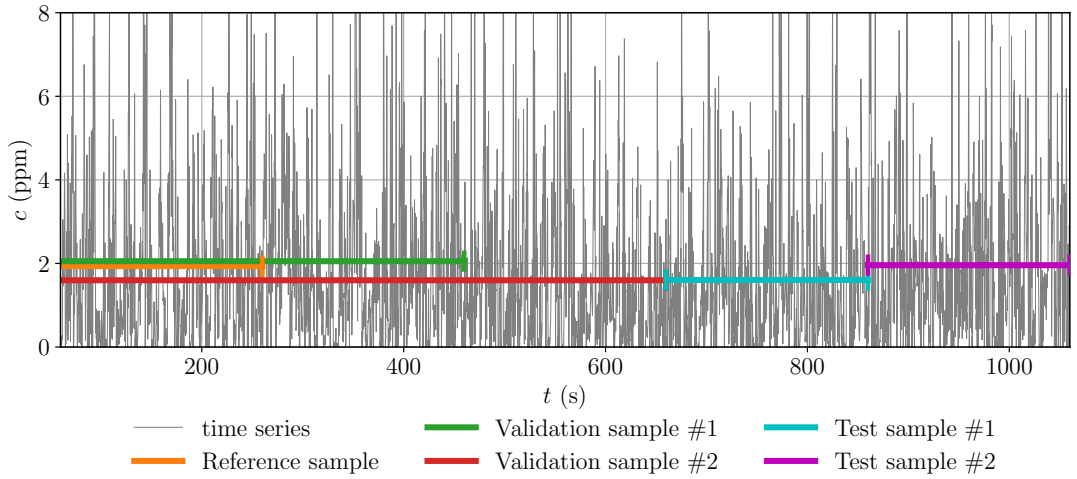


Figure III.11: *Time series of tracer concentration at tower B at $z = 2\,m$ estimated by LES. Colored segments correspond to the time windows used to validate the bootstrap estimates, namely: the reference sample ([60; 260 s]) in orange; validation samples ([60; 460 s], [60; 660 s]) in green and red; and test samples ([660; 860 s], [860; 1060 s]) in cyan and magenta.*

confidence intervals estimated by the stationary bootstrap for the vertical profiles of mean concentration at different towers. Small differences are found at tower D where confidence intervals obtained with $N_t = 20$ are larger (Fig. III.12d), and at tower T in altitude where the centers of the confidence intervals are shifted when increasing N_t which results in better coverage of the independent test realizations (Fig. III.12b). Overall, the stationary bootstrap has converged with $N_t = 20$ since providing additional sub-averages does not change the envelopes much. The convergence with the number of samples was also verified for wind velocity magnitude, direction and turbulent kinetic energy.

The distributions of the air quality metrics estimated by stationary bootstrap with resampling of 20 sub-averages among Nt =20, 40 and 60 are shown in Fig. III.13. The bootstrap ensemble averages slightly change because the time-averaged quantities over 200, 400 and 600 s are different. Nevertheless, increasing the number of samples for stationary bootstrap gives similar estimations of the metrics dispersion. For FAC2, MG and VG metrics, which are nonlinear, the tails of the distributions seem more dependent on the number of samples. As the orders of magnitude of the three estimates are overall consistent with each other, we conclude that $N_t = 20$ samples of sub-averages are sufficient for the stationary bootstrap method to converge. This implies that it is not required to run longer simulations to capture internal variability.

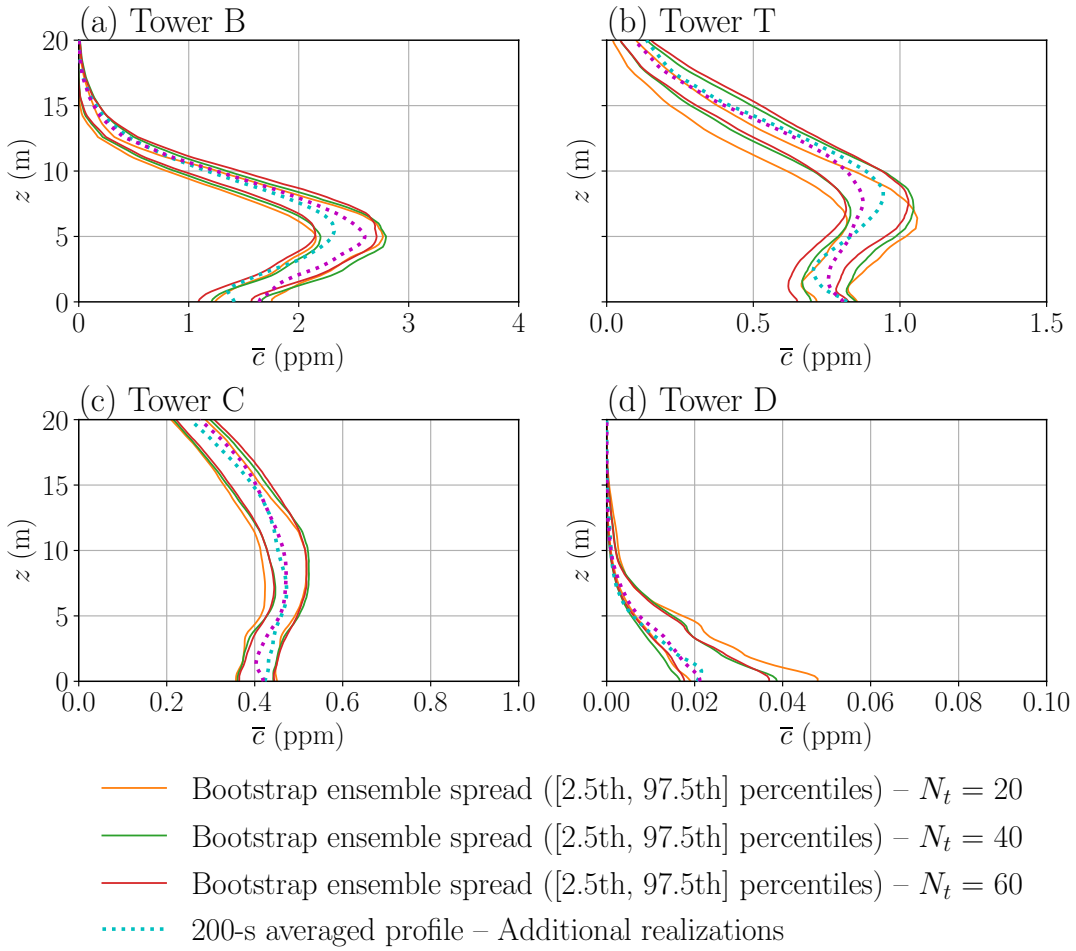


Figure III.12: *Vertical profiles of the 2.5th, and 97.5th percentiles of average concentration $\overline{c}$ (ppm) (a, b, c, d) at towers B, T, C and D, respectively (Fig. II.3, page 59). The confidence intervals are estimated with stationary bootstrap for the LES simulations with resampling of 20 sub-averages among 20, 40 and 60, in orange, green and red, respectively. Two additional realizations of LES estimations over 200 s are also represented as colored dotted lines.*

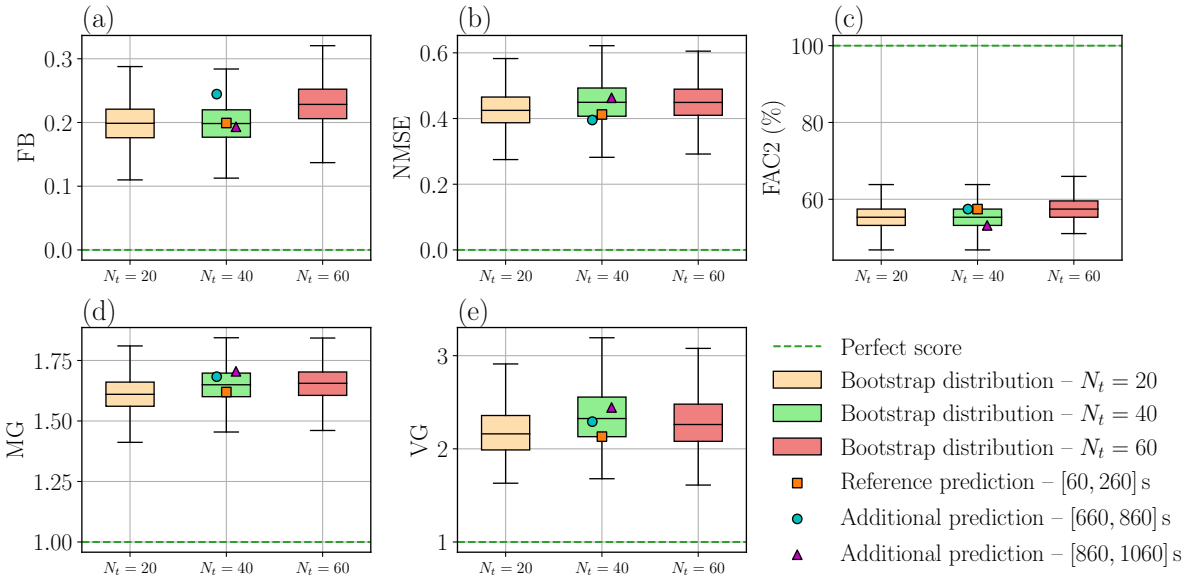


Figure III.13: *Box plots of the air quality metrics distributions obtained with stationary bootstrap with resampling of 20 sub-averages among 20, 40 and 60, in orange, green and red, respectively. Point estimations corresponding to the reference and two independent realizations of 200-s simulation are represented as orange squares, cyan circles and magenta triangles respectively. Results are given for (a) FB, (b) NMSE, (c) FAC2, (d) MG, and (e) VG metrics (Sect. III.3.2).*

III.5 LES model sensitivity to the main sources of uncertainty

In this section, we investigate how the various uncertainties associated with the LES model might influence its predictions. This is primordial to answer the general problem of this thesis of quantifying and reducing microscale LES dispersion prediction uncertainty. Moreover, it allows us to identify which uncertain parameters have the most impact on the model, in order to target them in the data assimilation system (Chapter V). Among the different uncertainties at stakes (Sect. I.2.2, page 27), we focus on:

i) the uncertainty related to the boundary condition parameters and in particular the one that accounts for the large-scale atmospheric forcing (Sect. III.5.2),
ii) the solver structural uncertainties linked to the subgrid-scale models and numerical scheme used (Sect. III.5.2).

The uncertainties related to the representation of the source and of the urban canopy are not considered in this study. Section III.5.1 presents the simple methodology used to assess the model sensitivities. These sensitivities are compared to the model internal variability, previously quantified in Sect. III.4, to conclude which sensitivities are significant.

III.5.1 One-at-a-time sensitivity analysis methodology

To find out which parameters have the most impact on the LES model mean concentration estimation, a One-At-a-Time (OAT) sensitivity analysis (Hamby 1995) is performed. This basic method consists of perturbing one-by-one each parameter of the model around its reference value while leaving the other parameters unchanged. The effect of each perturbation is assessed by comparing the estimated profiles of wind velocity and tracer concentration statistics with the baseline prediction. The discrepancies are quantified using the metrics introduced in Sect. III.3.3, in order to compare the various sensitivities. Using different metrics enables the detection of various effects on the mean concentration field: for example influence on the shape of the plume with FMS scores, or just on the high concentration with NMSE. In addition, the metrics are computed on bootstrap replicates of the fields to take into account the internal variability in the comparison, as explained in Sect. III.2.4.5.

We chose the OAT approach to limit the computational cost of the sensitivity analysis, as it requires only one model integration per parameter considered in addition to the baseline estimate. The simplicity of this method has two main drawbacks: first, the model sensitivities are evaluated locally around the reference parameters, second, this approach does not allow quantifying the combined influence of several parameters (Saltelli and Annoni 2010). Nevertheless, this approach allows us to quickly identify which uncertain parameters have no significant effect on the model predictions. It can thus be seen as a preliminary step towards a more in-depth but costly sensitivity analysis focused on the main model sensitivities. For example, a detailed sensitivity analysis of the model response to its two most impactful parameters using the Saltelli algorithm (Saltelli et al. 2010) is presented in Appendix. B.2, page 242.

III.5.2 Sensitivity to meteorological boundary conditions parameters

Microscale CFD dispersion models are known to be highly sensitive to meteorological boundary condition parameters (García-Sánchez et al. 2014; Lucas et al. 2016; Wise et al. 2018). Moreover, these parameters are often uncertain because of the internal variability of the ABL but also the limited amount of observation that can be used to determine representative boundary conditions for the model (Schatzmann et al. 2010; Pimont et al. 2017). In this study, we focus on the sensitivity of LES model prediction to four boundary condition parameters of the LES model: the mean wind direction α_{inlet}, the friction velocity u_*, the aerodynamic roughness length z_0, and the atmospheric fluctuations scaling factor $\lambda_{\mathbf{R}}$. A summary of where these parameters come into play and of the arbitrary perturbations tested in the OAT sensitivity analysis is given in Table III.6.

Table III.6: *LES model boundary condition parameters perturbations used for the OAT sensitivity analysis. The first three lines summarize the equations in which these parameters are involved.*

	α_{inlet}	u_*	z_0	$\lambda_{\mathbf{R}}$
Mean inlet wind profile (Eqs. II.17–II.18)	✓	✓	✓	✗
Turbulence injection (Eq. II.22)	✓	✓	✓	✓
Ground wall law (Eq. II.19)	✗	✗	✓	✗
Perturbation	-5°	+25%	+25%	+25%
Reference value	-41 °	$0.73\,\mathrm{m\,s^{-1}}$	0.045 m	1.0
Test value	-46 °	$0.91\,\mathrm{m\,s^{-1}}$	0.056 m	1.25

The purpose of the Reynolds stress tensor scaling parameter $\lambda_{\mathbf{R}}$ defined in Eq. II.23, page 72 is to be able to represent situations in which the incident ABL, is in a non-equilibrium state. Indeed, it can happen that the ABL is not in equilibrium due to obstacles that have not been taken into account, or due to large-scale fluctuations. Meanwhile, the precursor simulation used to estimate the injected fluctuations only represents the fully-developed ABL over the rough ground surrounding the area of interest, as explained in Sect. II.4.3, page 64. Note that since the imposed fluctuations come from a precursor simulation, they also depend on the other perturbed parameters (Table III.6).

Figure III.14 shows the effect of the -5° wind direction perturbation on the vertical profiles of mean concentration at different locations in the canopy. It appears that the inlet wind direction deviation significantly affects the mean concentration prediction by causing the plume centerline to deviate. Note that with this clockwise deviation of the wind direction, the vertical profiles at towers B, C and D are much closer to the observations. This tends to support our hypothesis that the lack of accuracy of the model could come from uncertainty in the wind direction, and could thus be improved using data assimilation to reduce the uncertainty in the wind direction. However, the -5° wind direction perturbation worsens the mean concentration prediction at tower T,

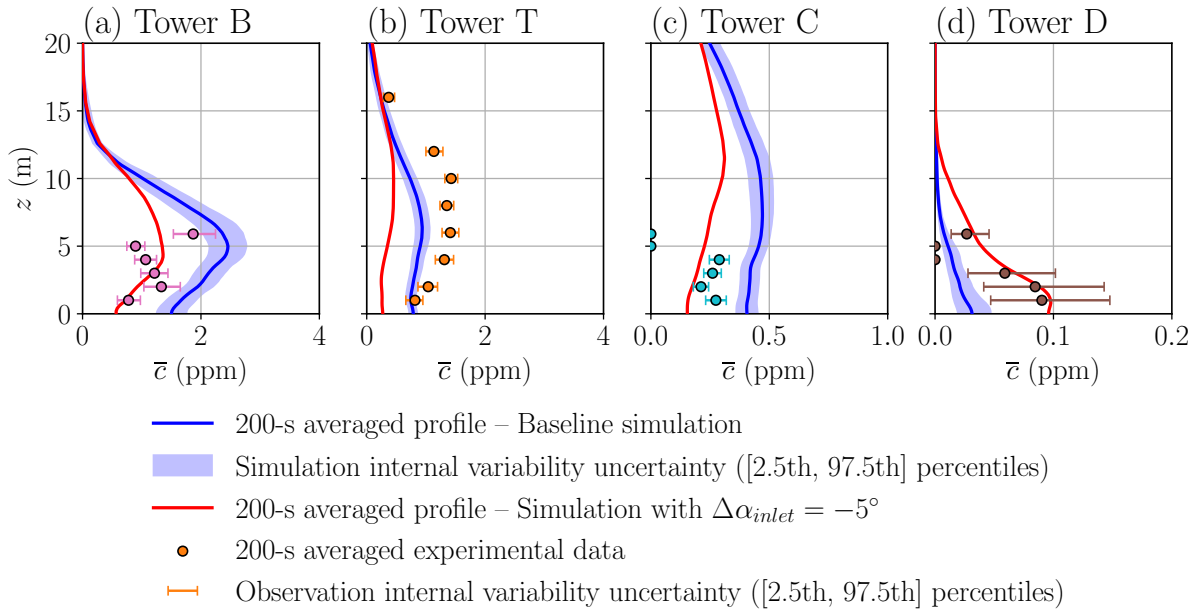


Figure III.14: *Vertical profiles of average concentration $\overline{c}$ (ppm) (a, b, c, d) at towers B, T, C and D, respectively. The baseline simulation predictions (in blue) are compared with the predictions obtained with a deviation of $\Delta\alpha_{inlet} = -5°$ in the incoming wind direction (in red). Colored circles correspond to observations. Internal variability uncertainty is represented with the [2.5th, 97.5th] percentile intervals estimated by stationary bootstrap as shaded blue areas for the baseline simulation and as error bars for the observations.*

indicating that other forms of error lead to inconsistent predicted vertical profiles.

Figure III.15 shows the effect of the three other boundary conditions perturbations (see Table III.6) on the wind velocity mean amplitude and horizontal turbulent kinetic energy at tower S. The 25% increase in friction velocity causes the same relative increase in both the mean wind magnitude and horizontal turbulent kinetic energy, which is coherent with the inlet wind profile (Eq. II.17, page 63), and the Reynolds stress tensor rescaling (Eq. IV.39). In contrast, the same relative increase of the aerodynamic roughness length z_0 just slightly slows down the wind velocity and does not affect the turbulent kinetic energy. This is because it is the logarithm of z_0 that is involved in the velocity profile (Eq. II.17, page 63). We could have tested larger perturbations on z_0 but the currently tested perturbation (0.016 m) corresponds to three times the roughness uncertainty determined by Yee and Biltoft (2004) for the MUST experimental site. For the perturbation of the fluctuations, we wind that the horizontal turbulent kinetic at tower S is significantly lower than the perturbed fluctuation profile imposed and is close to the level obtained in the baseline simulation. This is because when developing through the domain, the injected turbulent kinetic energy profile tends to return over a rather short distance to the equilibrium levels, as shown by the free-field simulation presented in Sect. II.5, page 70. Regarding tracer concentration (results not shown here), we verify that the mean concentration is inversely proportional to the friction velocity, as predicted by the similarity theory (Sect. I.1.2.4, page 19).

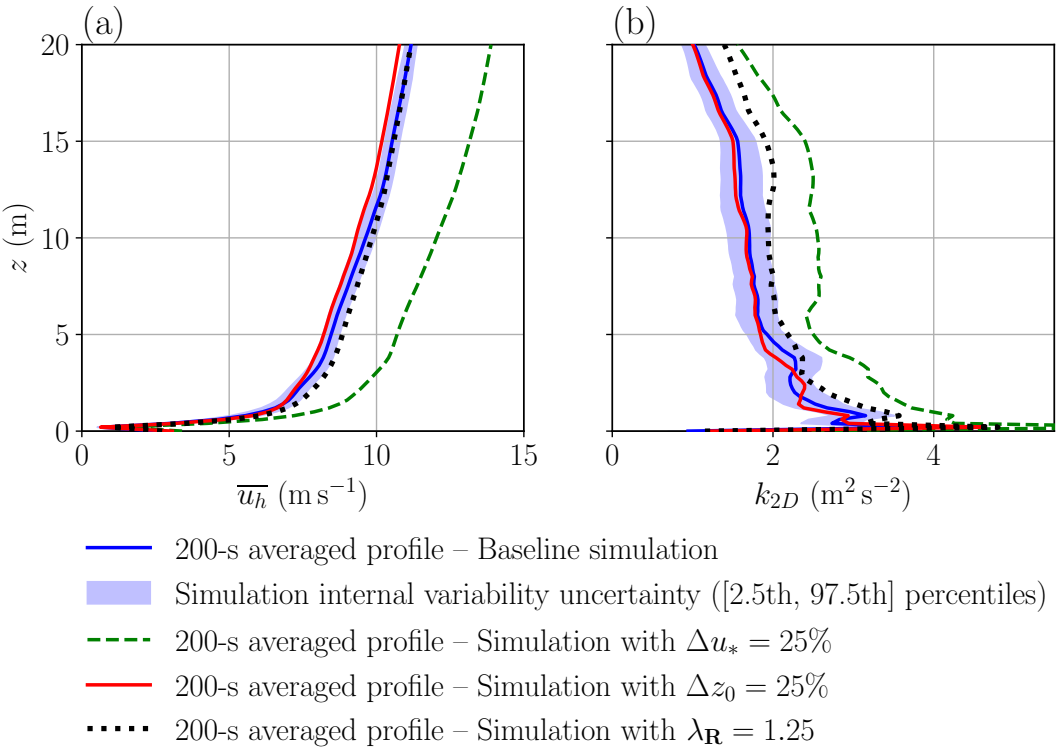


Figure III.15: *Vertical profiles of mean horizontal velocity (a) and horizontal turbulent kinetic energy (b) at tower S. Results are given for the baseline simulation (in blue), and for simulation with perturbed friction velocity (dashed green line), ground roughness (red solid line) and rescaling of the Reynolds stress tensor (black dotted line). Perturbations are given in Table III.6. Internal variability uncertainty is represented with the* [2.5*th*, 97.5*th*] *percentile intervals estimated by stationary bootstrap as shaded blue areas for the baseline simulation.*

In conclusion, the sensitivities of the mean concentration field to boundary conditions parameters are quantified and compared using the air quality metrics (Fig. III.16). Results show that the inlet wind direction α_{inlet} and the friction velocity u_* are the two most influential boundary condition parameters. The wind direction strongly influences the plume shape (Fig. III.16c, d), as well as both high and low concentrations (Fig. III.16b, f). Friction velocity introduces a mean bias to the model as concentrations decrease as friction velocity increases (Fig III.16a, e). On the other hand, we find that perturbations of the aerodynamic roughness length z_0 and of the inlet velocity fluctuations using $\lambda_{\mathbf{R}}$ have a much smaller effect on the mean concentration field. In particular, the scores obtained are not significantly larger than those estimated for the internal variability alone (Fig. III.16). In conclusion, for this study, the boundary condition parameters that are most important to take into account for the construction of the reduced model are α_{inlet} and u_*.

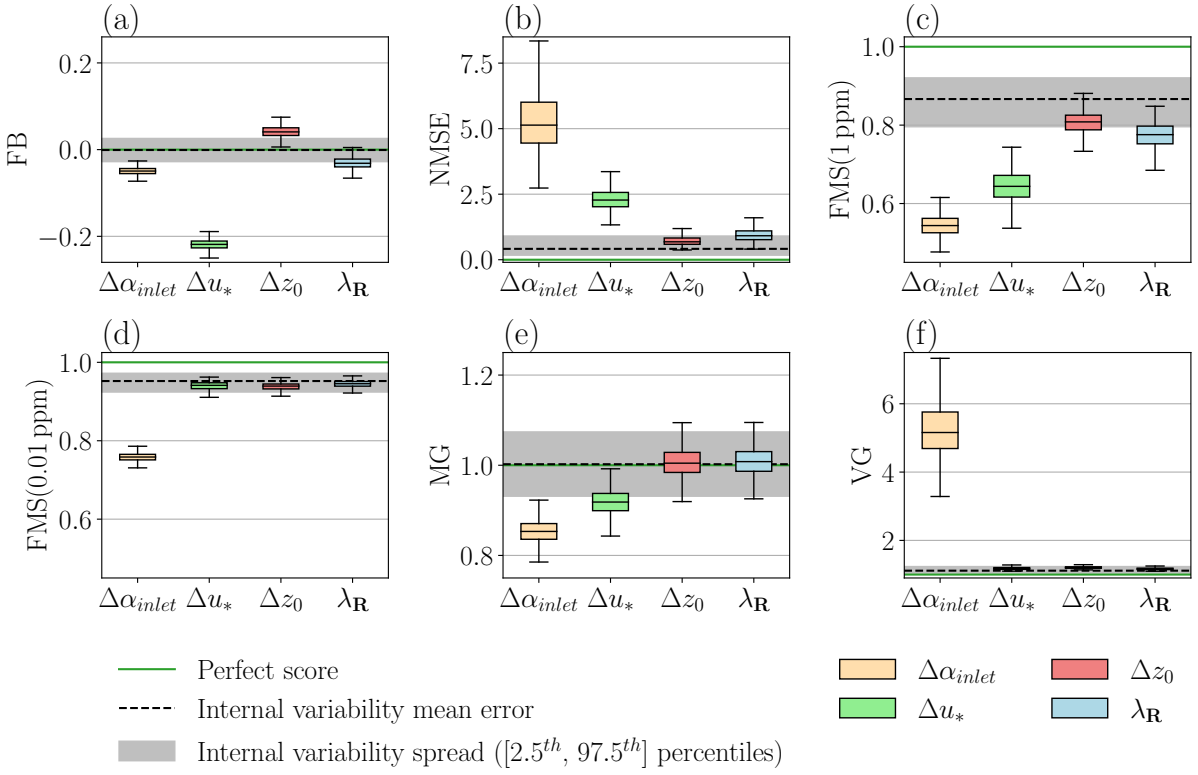


Figure III.16: *Box plots of the mean concentration field sensitivities to the boundary conditions parameters. Impacts of independent perturbations of the inlet wind direction $\Delta\alpha_{inlet}$, friction velocity Δu_*, aerodynamic roughness length z_0 and inlet Reynolds stress tensor ΔR on the mean concentration are assessed using the metrics defined in Sect. III.3.3: (a) FB, (b) NMSE, (c) FMS(1 ppm), (d) $FMS(0.01\,ppm)$, (e) MG, and (f) VG. Perturbations are given in Table III.6. The perfect scores associated with each metric are represented as horizontal green lines. The variance of the metrics induced by the microscale internal variability is estimated from bootstrap replicates of the mean concentration fields (Sect. III.2.4.5) and is represented as boxes-and-whiskers. The levels of error solely due to internal variability, also estimated using stationary bootstrap, are shown as black dashed lines, alongside the associated 95% confidence interval as grey shaded areas.*

III.5.3 Sensitivity to modeling choices

The LES predictions are also subject to structural uncertainties, i.e. uncertainties inherent to the solver code and the underlying modeling assumptions. In this section, OAT sensitivity tests are carried out to assess the effect of the choices of the subgrid-scale models and of the numerical scheme on model predictions.

We perform one sensitivity test for the numerical scheme by using the third-order in space and time, two-step Taylor Galerkin (TTGC) finite-element scheme (Colin and Rudgyard 2000) instead of the second-order Law-Wendroff (LW) finite-volume centered scheme (Schönfeld and Rudgyard 1999). Another test illustrates the differences induced by a change in the subgrid-scale model by using the dynamic Smagorinsky model (Ger-

mano et al. 1991) instead of the WALE model from Nicoud and Ducros (1999). These subgrid-scale models are defined in Sect. II.2.3, page 52. Two additional simulations assess the effect of the turbulent Schmidt number S_c^t, which controls the subgrid-scale tracer diffusive flux (Eq. II.9, page 54), with tested values close to the limits of the range of optimum value $S_c^t \in \{0.2,\ 1.3\}$ in RANS according to Tominaga and Stathopoulos (2007). Table III.7 summarizes the four sensitivity tests.

Table III.7: *OAT sensitivity tests performed to assess the structural uncertainties associated with the LES model presented in Chapter II.*

	Numerical scheme	SGS model	S_c^t
Reference	LW (2^{nd} order)	WALE	0.6
Test	TTGC (3^{rd} order)	Dynamic Smagorinsky	$\{0.2,\ 1.0\}$

Overall, these sensitivity tests show a very limited impact of the modeling choices on the main predictions of the model. The only significant effect is found for the numerical scheme as, with TTGC, the estimated mean flow velocity globally decreases which increases the predicted levels of concentration (not shown here). On the other hand, there is no clear tendency emerging from the test on the subgrid-scale model as most of the differences are comparable to the internal variability of the model. Note that more significant differences were found near the obstacles, with changes in the wind flow patterns. However, the LES model is under-resolved near the wall (see Sect. II.4.1, 61) and there are no observations in these locations to be conclusive. In addition, the turbulent Schmit number effects are totally negligible when compared to internal variability (not shown here).

Figure III.17 illustrates the effect of the modeling choices on the mean concentration field using the air quality metrics presented in Sect. III.3.3. This confirms the qualitative analysis of the concentration profiles. Indeed, the numerical scheme has a small impact on the mean concentration with an overall tendency to increase the concentration in the domain (MG < 1). Regarding the effects of the subgrid-scale model and the turbulent Schmidt number, they do not exceed the level of errors associated with internal variability alone. Note that only the comparison between $Sc_t = 1.0$ and the reference value of 0.6 is illustrated in Figure III.17, but identical conclusions are made for the test with $Sc_t = 0.2$. Figure III.17 also compares the model sensitivities to structural uncertainties with the model sensitivities to boundary condition uncertainties. It clearly shows that boundary condition uncertainties have a predominant effect on the LES model predictions: especially the inlet wind direction and the friction velocity. This is the important message of this section, and it is consistent with the literature which states that boundary condition uncertainties (and geometrical differences) dominate uncertainties related to turbulence modeling when reproducing field experiments (Neophytou et al. 2011; García-Sanchez et al. 2018; Dauxois et al. 2021).

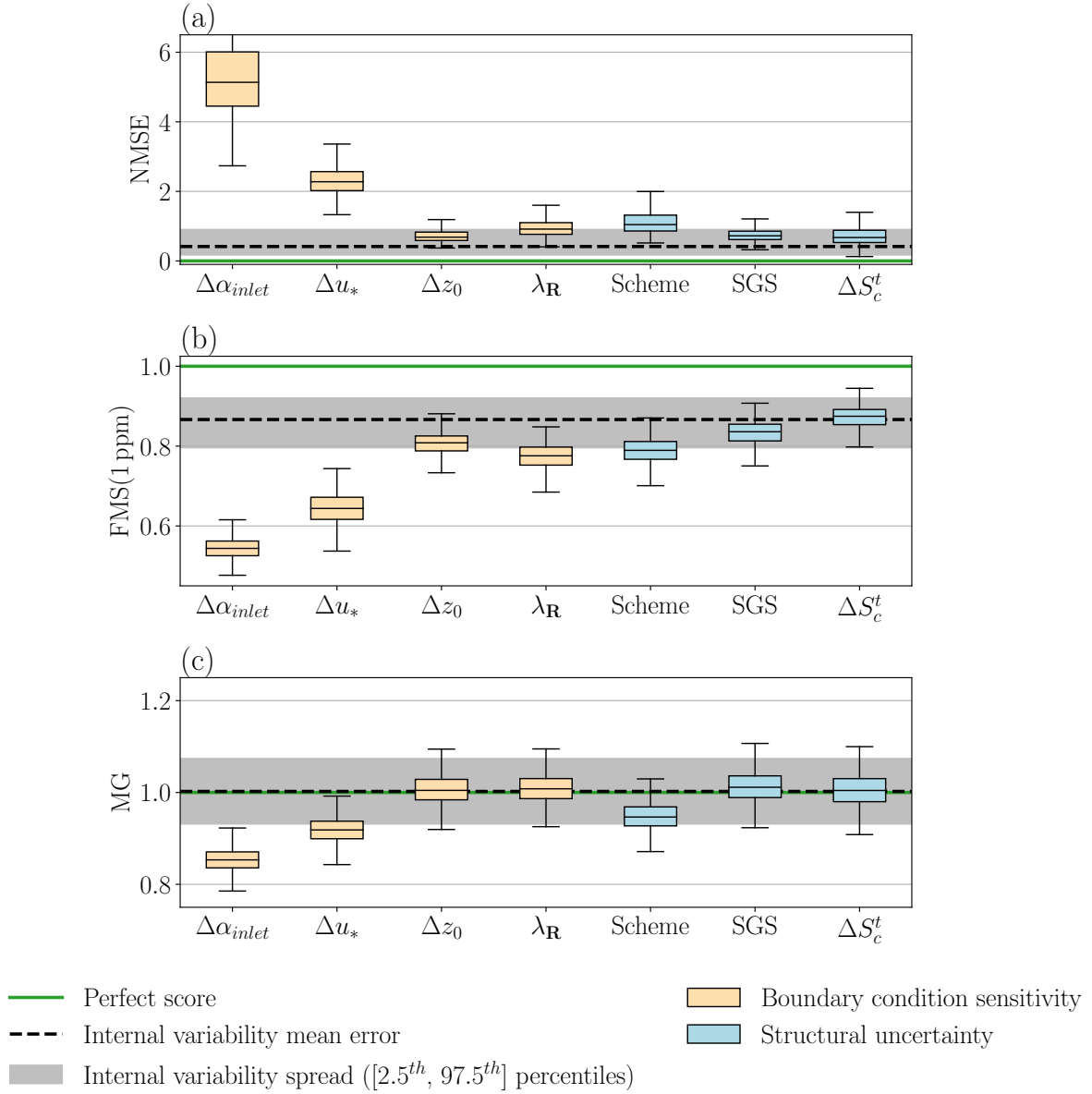


Figure III.17: *Box plots of the main sensitivities of the mean concentration field estimated by the LES model. Yellow boxes, on the left, correspond to the effects of boundary condition parameters (Table III.6). Blue boxes correspond to the effects of the structural modeling uncertainties (Table III.7). These effects are assessed using various metrics: (a) NMSE, (b) FMS(1 ppm), and (c) MG. Perfect scores associated with each metric are represented as horizontal green lines. Metric variance induced by the microscale internal variability of the ABL is estimated from bootstrap replicates of the mean concentration fields (Sect. III.2.4.5) and is represented as boxes and whiskers. The level of errors solely due to internal variability, also estimated using stationary bootstrap, are shown as black dashed lines, alongside the associated 95% confidence interval as grey zones.*

III.6 Conclusion

Summary In this chapter we have validated the LES model of the neutral MUST trial 2681829 presented in Chapter II. This validation is assessed robustly by taking into account the main sources of uncertainty involved in microscale atmospheric dispersion in urban environments. Particular attention is paid to: i) the irreducible uncertainty linked to the microscale internal variability of the ABL, ii) the effect of large-scale atmospheric forcing uncertainty; and iii) structural modeling errors. We show that these uncertainties significantly affect LES predictions and can explain part of the discrepancies between the model and observations.

LES model validation Despite these uncertainties, the LES model is in acceptable agreement with field measurements. Wind flow statistics predictions are very accurate, except near the top of the computational domain, where the LES model underestimates the turbulent kinetic energy. For the tracer mean concentration, we find some inconsistencies in the model predictions compared to the field measurements. Nevertheless, the scores obtained with the standard air quality metrics are in line with other LES studies reported in the literature. In addition, we find that the LES model is very accurate at predicting relative fluctuations and concentration peaks.

Internal variability quantification The main originality of this work is to provide a way to quantify the effect of the microscale internal variability on wind flow and pollutant concentration. For this purpose, we adopt a bootstrap approach based on sub-averages resampling and implement the stationary bootstrap algorithm from Politis and Romano (1994) to take into account the dependency between sub-averages. In the context of the studied MUST field trial, we show that: i) field measurements of wind flow and tracer dispersion are significantly affected by internal variability, ii) the LES model reproduces very well the observed variability, despite the idealized representation of the inflow (stationary flow with synthetic turbulence). As a result, internal variability leads to significant uncertainty in the scores of the metrics used to validate the model. To avoid misleading analysis, we therefore advise switching from point-wise validation to probabilistic validation to account for this internal variability. In addition, we demonstrate both the convergence and plausibility of the bootstrap estimates, which indicate that it is not necessary to perform longer simulations to capture the microscale internal variability involved. This is a significant advantage of the proposed approach, given the computational cost of the LES model.

Sensitivity analysis We find that the differences between observations and LES estimates cannot be explained by internal variability alone, implying that other forms of error must affect the LES model. We therefore carry out model sensitivity tests to assess the effect of uncertainty on boundary condition parameters (in particular large-scale atmospheric forcing and ground wall law parameters), and modeling choices (numerical scheme and subgrid-scale model). Results show that LES predictions are significantly more sensitive to meteorological boundary conditions perturbations than to structural

modeling choices, which is in line with the literature (Dauxois et al. 2021). The comparison with the error solely due to internal variability demonstrates that two parameters have a predominant effect on the model mean concentration predictions: the inlet wind direction and the friction velocity.

Link with thesis objective In this chapter, we have met one of the main objectives of the thesis by assessing the main sources of uncertainty affecting the accuracy of the model validated in this chapter. In addition, we have identified the uncertain parameters that will be most interesting to target with the data assimilation system designed in this thesis to improve the accuracy of the LES model. Finally, the internal variability quantification approach proposed in this chapter provides a valuable methodology for guiding the construction of the reduced-order model (Chapter IV) and quantifying irreducible errors within the data assimilation (Chapter V).

Chapter IV

Reduced-order modeling based on LES statistical predictions

This chapter covers the construction and validation of a reduced-order model to efficiently emulate the response surface of the LES model of the MUST field experiment presented in Chapter II. In particular, we focus on the efficient prediction of the mean concentration field for varying meteorological condition parameters.

After posing the model reduction problem in the opening section, we present the theoretical principles of the data-driven reduced-order modeling approach adopted in this thesis. This approach, called POD–GPRs, combines proper orthogonal decomposition (POD) to reduce the dimension of the fields to be emulated, and Gaussian process regressors (GPRs) to predict the reduced-basis projected components for varying meteorological boundary condition parameters. We also derive in this section the equations for estimating the uncertainty of the POD–GPRs predictions from the posterior GPR distributions.

Then, the model reduction validation methodology is briefly presented as well as another reduced-order model: the nearest neighbor. This trivial reduced-order model provides a benchmark for assessing the added value of the more sophisticated POD–GPRs approach.

Another section is devoted to the generation of the training base of 200 LES field predictions, describing how the input parameter space is sampled and how the LES model is adapted to compute such a large ensemble.

The last part of the chapter details the efforts made to apply the POD–GPRs approach to the MUST case study. In particular, we show that it requires special attention in the preprocessing of the training base to deal with the wide range of scales involved. We also propose a methodology to select the number of POD–GPRs reduced-basis modes in order to avoid overfitting noisy structures linked to atmospheric internal variability. Finally, we present a thorough validation of the POD–GPRs reduced-order model in order to assess its efficiency, accuracy and sensitivity to the size of the training database.

Chapter outline

IV.1 Introduction 119
IV.1.1 Model reduction problem statement 120
IV.1.2 Choice of the reduced-order model input parameters 120
IV.1.3 Physical similarity in reduced-order modeling 121
IV.2 Reduced-order modeling approach 122
IV.2.1 Principle 122
IV.2.2 Proper orthogonal decomposition 123
IV.2.3 Gaussian process regression 126
IV.2.4 Fields preprocessing 128
IV.2.5 Reduced-order model uncertainty prediction 131
IV.3 Reduced-order model validation methodology 133
IV.3.1 A control model: the nearest neighbor 133
IV.3.2 Model reduction error quantification 134
IV.3.3 Model cross-validation 135
IV.3.4 Accounting for internal variability in the validation 135
IV.4 LES ensemble generation 137
IV.4.1 Definition of parameter ranges from micro-climatology 137
IV.4.2 Parameter space sampling 138
IV.4.3 Adaptation of the LES model to run ensembles 138
IV.4.4 Computation of the LES ensemble 142
IV.5 Setting up the POD–GPRs model 144
IV.5.1 Statistical approach to selecting the number of modes 144
IV.5.2 Field preprocessing effect on proper orthogonal decomposition 148
IV.5.3 Ability of Gaussian process regressors to represent internal variability 152
IV.6 Validation of the POD–GPRs model 155
IV.6.1 Analysis of the model reduction error 155
IV.6.2 Behavior of the reduced-order models for restricted train set . . 160
IV.6.3 Sensitivity to the choice of the training data 162
IV.6.4 Dispersion of errors over test samples 162
IV.6.5 Why is it primordial to restrict the number of modes? 164
IV.7 Conclusion 167

IV.1 Introduction

One of the main problems with LES dispersion models is their high computational cost as outlined in Sect. I.2.1, page 26. This burden makes LES unsuitable for applications requiring a large ensemble of fast predictions and thereby represents a major obstacle to overcome in order to answer the problematic of the thesis of quantifying and reducing uncertainty in microscale LES dispersion models. The general objective of this chapter is to overcome this computational cost limitation by building a data-driven reduced-order model of the LES model of the MUST experiment presented in Chapter II. In particular, we aim at efficiently emulating the LES mean concentration response surface for varying boundary condition parameters, in order to use this reduced-order model as a surrogate for the LES model in the data assimilation system designed in Sect. I.3.3 and implemented in Chapter V.

Data-driven reduced-order modeling techniques have emerged as a promising solution to overcome the computational cost limitation of CFD models, as shown in the literature overview presented in Sect. I.3.1. In an off-line phase, this type of reduced-order model learns the response surface of the costly model for varying input parameters from a database of pre-computed simulations. Once trained, reduced-order models can make new online predictions at almost no computational cost and with very little loss of accuracy.

Among the various data-driven model reduction approaches presented in Sect. I.3.1, we adopt an approach based on the combination of proper orthogonal decomposition (POD) to reduce the field dimension, and Gaussian process regressors (GPRs) to estimate the reduced-basis projected components for new input parameters. This reduced-order modeling approach is referred to as POD–GPRs in the following. It has been used by Nony (2023) to emulate LES predictions of a canonical dispersion case (i.e. two-dimensional boundary-layer flow interacting with a surface-mounted obstacle) with a very limited error and for a large range of uncertain input parameters including the operating wind condition. One of the main challenges of this chapter is to adapt the POD–GPRs to the challenges arising for the more realistic dispersion case which is the MUST field experiment. These challenges notably include the considerable increase in dimension, the more complex interactions between flow and obstacles, and the wider ranges of concentration scales involved (the test case studied by Nony (2023) can be seen as a focus on the near-source region).

In the following, we define more precisely the model reduction problem as well as the input and output spaces considered.

IV.1.1 Model reduction problem statement

The goal of this chapter is to build an efficient reduced-order model that emulates as closely as possible the LES model response surface relative to the input parameters[1] $\boldsymbol{\theta}$, this means finding a function:

$$\begin{aligned} \mathcal{M}_{\mathrm{ROM}} : \Omega_{\boldsymbol{\theta}} &\longrightarrow \mathbb{R}^N, \\ \boldsymbol{\theta} &\longmapsto \mathbf{y}_{\mathrm{ROM}}, \end{aligned} \tag{IV.1}$$

that minimizes $\int_{\Omega_{\boldsymbol{\theta}}} \|\mathbf{y}_{\mathrm{ROM}}(\boldsymbol{\theta}) - \mathbf{y}_{\mathrm{LES}}(\boldsymbol{\theta})\| \, \mathrm{d}\boldsymbol{\theta}$, where $\mathbf{y}_{\mathrm{LES}} \in \mathbb{R}^N$ is the field to emulate, discretized on a grid of N nodes, $\mathbf{y}_{\mathrm{ROM}}$ is its counterpart predicted by the reduced-order model, and $\Omega_{\boldsymbol{\theta}} \subset \mathbb{R}^d$ is the input parameter space of dimension d.

In the data-driven method retained for this work, this function is obtained by learning from a database of precomputed LES predictions $\left\{\mathbf{y}_{\mathrm{LES}}(\boldsymbol{\theta}^{(i)})\right\}_{i=1}^{N_{train}}$ called the train set. A fundamental good practice is to set aside part of the database, the test set, and not use it to build the reduced-order models. The generalization error of the reduced-order model can be then evaluated on these test samples.

In this chapter, we focus on the emulation of the mean tracer concentration field, meaning that $\mathbf{y} = \overline{\mathbf{c}}$. This choice is motivated by the fact that mean concentration is: i) the main quantity of interest in dispersion studies; and ii) the observational data we want to use in the data assimilation system built in Chapter V. However, the chosen reduced-order modeling approach can be used to emulate any field predicted by the LES model, as shown in Appendix. B. The choice of emulating mean statistics represents a first step in the proof-of-concept of the reduced-cost assimilation system designed in this thesis. Taking into account the temporal dimension is an important prospect, but greatly complicates the task of model reduction and is outside the scope of this thesis.

As explained in Chapter III, to prevent time averages from being affected by the internal variability of the ABL, a very long simulation time should be used. Nevertheless, we decided to limit simulation duration to 200-s of steady state, i.e. the same duration as the analysis period defined by Yee and Biltoft (2004). This limited simulation time allows us to greatly reduce the computational cost of the LES ensemble. Moreover, this choice is more in line with the perspective of using the reduced-order model in a data assimilation context where measurements are assimilated sequentially and therefore defined over restricted time windows. Taking into account the aleatory uncertainty related to the internal variability of the ABL in the construction of the reduced-order model and its predictions represents one of the main challenges of this chapter.

IV.1.2 Choice of the reduced-order model input parameters

To be used in a data assimilation system, the reduced-order model must accurately reproduce the LES response surface to the uncertain physical parameters involved. An

[1]Note that the input parameters vector $\boldsymbol{\theta}$ should not be confused with potential temperature, which is denoted θ in Chapter I by convention in meteorology, but is not used in the rest of the thesis.

overview of the main sources of uncertainty affecting microscale CFD dispersion models is given in Sect. I.2.2, page 27. Concerning pollutant source parameters, Nony (2023) demonstrates that addressing uncertainties in source placement presents a strong challenge for model reduction, which requires a large training dataset. Because of the high computational cost of the LES model used in this thesis (20 000 core hours per prediction), we decided to focus on meteorological boundary conditions and not consider source parameters.

As shown by the one-at-a-time sensitivity analysis presented in Sect. III.5.2, page 109, the meteorological boundary condition parameters to which our model is most sensitive are the inlet wind direction α_{inlet} and friction velocity u_*. In this study, we therefore consider only two parameters:

$$\boldsymbol{\theta} = (\alpha_{inlet}, u_*)^{\mathrm{T}}, \tag{IV.2}$$

to define the input space of the reduced-order model.

We remind that this thesis is limited to the study of pollutant dispersion under neutral thermal stratification conditions. The extension of the reduced-order modeling approach developed in this chapter to varying thermal stratification conditions is a direct and important prospect since i) the pollutant plume is significantly affected by the thermal stratification of the atmosphere as shown in Fig. I.6, page 20; ii) the exact thermal stratification state of the atmosphere might be uncertain in operational contexts.

IV.1.3 Physical similarity in reduced-order modeling

As shown in Sect. I.1.2.4, page 19, the velocity and species concentration fields are linearly dependent on the friction velocity u_* under neutral atmospheric conditions. This information can be used to simplify the model reduction problem by predicting quantities normalized by friction velocity:

$$\tilde{c} = \frac{c\, u_* H^2}{Q_{source}}, \tag{IV.3a}$$

$$\tilde{\mathbf{u}} = \mathbf{u}/u_*. \tag{IV.3b}$$

with $H = 2.54\,\mathrm{m}$ the obstacle height and Q_{source} the volumetric release rate in the studied MUST trial (Table II.2, page 60). This strategy was for example used in the reduced-order model used by Sousa et al. (2018) to predict the wind flow on Stanford's campus.

Using the similarity of the governing equations, it is no longer necessary to consider friction velocity when constructing the LES set to train the reduced-order model. Nevertheless, we decided to still sample the full space of input parameters (Eq. IV.2), as we aim to i) demonstrate the ability of the chosen reduction method to handle more than one dimension; ii) explore the effect of internal variability for different friction velocities; iii) use similarity theory as a theoretical basis to validate the reduced-order model; iv) compare the reduced-order model accuracy with and without using similarity.

IV.2 Reduced-order modeling approach

In this section, we define in more detail the POD–GPRs reduced-order modeling approach adopted in this chapter. In particular, we present the theoretical basis of the proper orthogonal decomposition (Sect. IV.2.2) and Gaussian process regression (Sect. IV.2.3), which are at the core of the POD–GPRs approach. We also introduce in Sect. IV.2.4 the various methods we use for preprocessing the fields used to train the reduced-order model. Finally, we derive an analytical expression of the uncertainty of the POD–GPRs estimates from the Gaussian process posterior distribution (Sect. IV.2.5).

IV.2.1 Principle

The fundamental principle of the POD–GPRs reduced-order modeling approach is to combine a reduction dimension step and a regression step (Chinesta et al. 2011).

The reduction step consists in finding a reduced basis $\{\boldsymbol{\psi}_\ell\}_{\ell=1}^{L}$ of dimension L on which to project the LES modeled field $\mathbf{y}(\boldsymbol{\theta})$ to be emulated:

$$\mathcal{T}\left(\mathbf{y}(\boldsymbol{\theta})\right) \approx \sum_{\ell=1}^{L} k_\ell(\boldsymbol{\theta})\boldsymbol{\psi}_\ell, \tag{IV.4}$$

with $\mathcal{T}$ a field preprocessing treatment, $\boldsymbol{\psi}_\ell \in \mathbb{R}^N$ the ℓth reduced basis vector, commonly referred to as the ℓth mode, and $k_\ell(\boldsymbol{\theta}) \in \mathbb{R}$ the unique coefficient associated with the ℓth mode in the projection of the given field $\mathbf{y}(\boldsymbol{\theta})$ on the latent space span $\left(\{\boldsymbol{\psi}_\ell\}_{\ell=1}^{L}\right)$. The key idea of the reduction step is to find the best-reduced basis to minimize the approximation error in Eq. IV.4, while also minimizing the number of modes L to characterize the variability of the quantity $\mathbf{y}$ in the parameter space. By finding a reduced basis of dimension $L \ll N$, the reduction step drastically decreases the computational burden linked to the dimension of the outputs.

The regression step then consists in predicting the decomposition $\{k_\ell(\boldsymbol{\theta})\}_{\ell=1}^{L}$ of the field $\mathbf{y}(\boldsymbol{\theta})$ in the latent space given the train set $\left\{\mathbf{y}_{\text{LES}}(\boldsymbol{\theta}^{(i)})\right\}_{i=1}^{N_{train}}$.

One important thing to note is that this reduction-regression approach separates the parametric dependence of the field $\mathbf{y}(\boldsymbol{\theta})$ from the spatial variability. Indeed, the spatial structure of the fields is now carried by the modes $\boldsymbol{\psi}_\ell$, while the coefficients k_ℓ reflect the field parametric dependency.

As its name suggests, the POD–GPRs method relies on Proper Orthogonal Decomposition (POD) for the reduction step, and Gaussian Process Regression (GPR) for the regression step. The POD–GPRs method follows the standard approach of statistical learning methods, with an initial training phase that consists of i) preprocessing the LES fields, ii) building the POD reduced basis based on the train set, and iii) optimizing the GPRs in the latent space (Fig. IV.1a). This training phase is performed offline and only once. Then the prediction phase takes place as follows, given any wind conditions $\boldsymbol{\theta}$, i) the associated POD reduced coefficients are predicted by the fitted GPRs, ii) the inverse POD projection and inverse fields scaling are applied to these coefficients to recover the physical field $\mathbf{y}_{\text{ROM}}$ (Fig. IV.1b). The following sections introduce the theoretical basis of POD and GPR techniques.

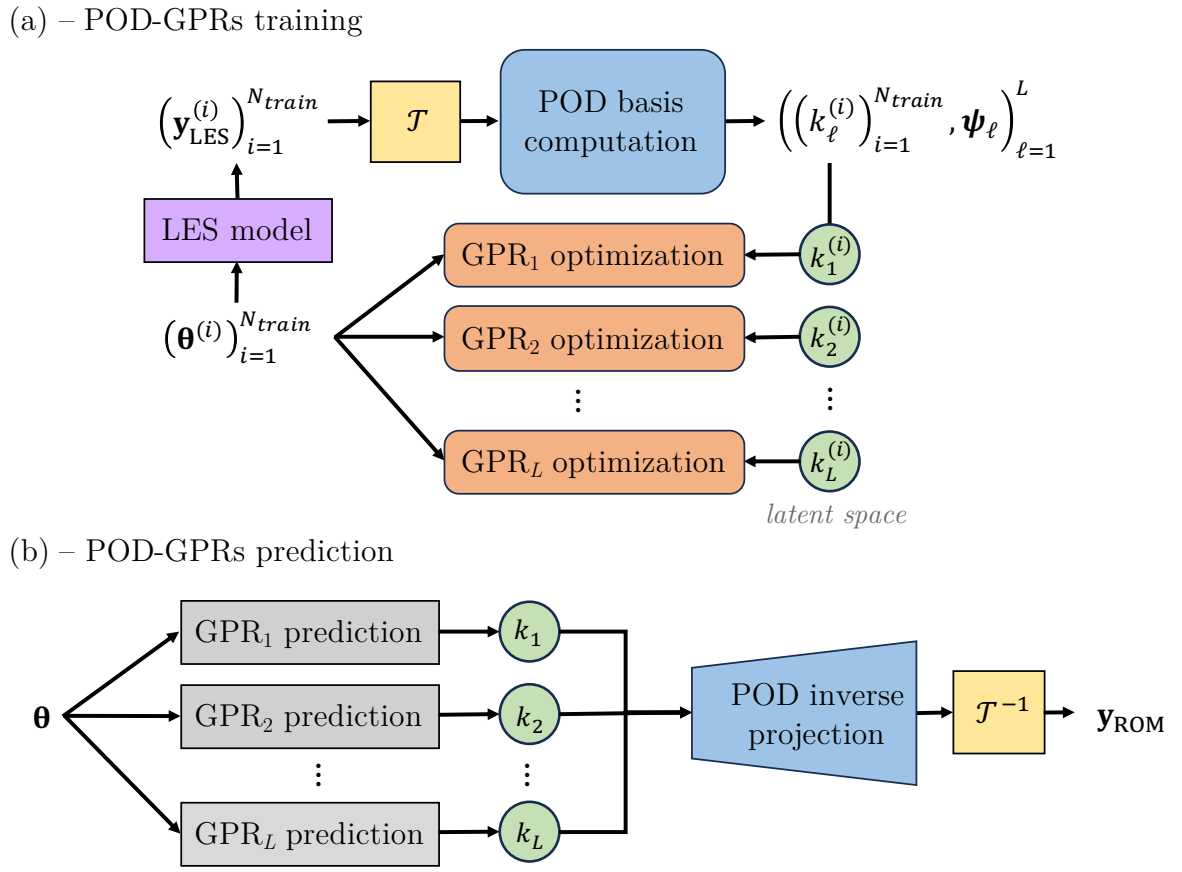


Figure IV.1: *Schematic of the POD–GPRs reduced-order model. Its operation is divided into two stages: the training phase (a), and the prediction phase (b). First, a preprocessing $\mathcal{T}$ is applied to the LES predicted field, and the POD reduced basis is built on the scaled train set, then L independent GPR models are optimized to emulate the L reduced coefficients $(k_1, ..., k_L)$ for the input parameters $\boldsymbol{\theta}$, with L the number of POD modes. For the prediction phase, the fitted GPRs predict the POD reduced coefficients associated with the set of wind conditions $\boldsymbol{\theta}$ of interest, then the inverse POD projection and inverse scaling $\mathcal{T}^{-1}$ are applied to recover the associated physical field.*

IV.2.2 Proper orthogonal decomposition

Proper Orthogonal Decomposition (POD) is a dimension reduction method historically used in the context of fluid dynamics for turbulence analysis (Sirovich 1987; Berkooz et al. 1993). It allows for defining a reduced basis for the modal decomposition of a physical field given an ensemble of snapshots. POD is also commonly referred to as Principal Component Analysis (PCA) in geoscience and statistics, and Singular Value Decomposition (SVD) in linear algebra.

Because of its simplicity and its useful properties (see Eq. IV.9), POD is a very popular tool in model reduction (Chinesta et al. 2011; Vinuesa and Brunton 2022). Note that other methods exist, in particular, neural-network autoencoders (Bourlard and Kamp 1988; Hinton and Zemel 1993) can provide greater accuracy by better managing the

disparity of scales in the concentration field, as shown by Nony (2023). However, POD has a substantial advantage over this approach, since it hierarchizes the information carried by the modes in the reduced space. This property motivates our choice to use POD as it allows filtering out the noise induced by the internal variability of the atmospheric boundary layer (see Sect. IV.5.1). Concerning the problem of scale disparity, we address it by preprocessing the fields before building the POD basis (see Sect. IV.2.4).

The POD is usually built thanks to a set of realizations, called *snapshots*, of one field, for example at different acquisition times (Sirovich 1987; Berkooz et al. 1993). In this work, the realizations correspond to the train set of time-averaged concentration fields $\left\{\mathbf{y}_{\mathrm{LES}}(\boldsymbol{\theta}^{(i)})\right\}_{i=1}^{N_{train}}$, obtained for different input parameters $\boldsymbol{\theta}^{(i)}$. In this way, the POD spatial mode vectors describe the coherent spatial structures emerging from variations of the wind conditions $\boldsymbol{\theta} = (\alpha_{inlet},\ u_*)^{\mathrm{T}}$.

The POD basis is obtained by diagonalizing the covariance matrix of the snapshot ensemble:

$$\mathbf{C} = (\mathrm{Cov}(\mathbf{y}_{\mathrm{LES}}(\mathbf{x}_i),\ \mathbf{y}_{\mathrm{LES}}(\mathbf{x}_j))_{1\leq i,j\leq N} = \frac{1}{N_{train}-1}\mathbf{S}\mathbf{S}^T, \tag{IV.5}$$

with $\mathbf{S} = \left(\mathcal{T}(\mathbf{y}_{\mathrm{LES}}^{(1)}), \ldots, \mathcal{T}(\mathbf{y}_{\mathrm{LES}}^{(N_{train})})\right) \in \mathbb{R}^{N\times N_{train}}$ the snapshot matrix, and $\mathcal{T}$ a preprocessing transformation applied to all snapshots which include centering (see Sect. IV.2.4). As a real, symmetric matrix, $\mathbf{C}$ is diagonalizable, such as:

$$\mathbf{C} = \boldsymbol{\Psi}\boldsymbol{\Sigma}^2\boldsymbol{\Psi}^{\mathrm{T}}, \tag{IV.6}$$

with $\boldsymbol{\Psi} = \{\boldsymbol{\psi}_\ell\}_{\ell=1}^N$ an orthonormal basis of $\mathbb{R}^N$, and $\boldsymbol{\Sigma} = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_N)$ the associated eigenvalues of $\mathbf{C}$. In this study, the POD modes $\boldsymbol{\Psi}$ are computed using the scikit-learn[2] implementation of the stochastic algorithm from Halko et al. (2009).

The POD reduced basis is then defined by truncating the basis $\boldsymbol{\Psi}$ to the L vectors associated with the first L eigenvalues sorted in decreasing order. This is motivated by the fact that the fraction of total ensemble variance carried by the ℓth mode is quantified by Λ_ℓ, and a limited number of modes $L \ll N$ is often sufficient to represent accurately the main features of the field (Cordier and Bergmann 2006; Nony 2023). In this thesis, we introduce a new methodology to make an informed selection of the optimal number of modes by taking into account snapshot noise through a bootstrap approach (see Sect. IV.5.1).

The projection on the POD reduced basis is then simply defined, by the linear transformation:

$$\begin{aligned} \mathcal{P}_{\mathrm{POD}} : \mathbb{R}^N &\longrightarrow \mathbb{R}^L, \\ \mathbf{y} &\longmapsto \mathbf{a} = \widetilde{\boldsymbol{\Psi}}^{\mathrm{T}}\mathcal{T}(\mathbf{y}), \end{aligned} \tag{IV.7}$$

with $\mathbf{a} = \{a_1, \ldots, a_L\}$ the coefficients in the projection of $\mathbf{y}$ on the truncated POD mode vectors basis $\widetilde{\boldsymbol{\Psi}} = \{\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_L\}$. This allows approximating any field $\mathbf{y}$ as linear combination of the POD mode vectors as in Eq. IV.4 with the POD inverse projection

[2]See `https://scikit-learn.org/`

reading:

$$\mathcal{P}_{\text{POD}}^{-1} : \mathbb{R}^L \longrightarrow \mathbb{R}^N, \tag{IV.8}$$
$$\mathbf{a} \longmapsto \mathbf{y} = \mathcal{T}^{-1}\left(\sum_{\ell=1}^{L} a_\ell \boldsymbol{\psi}_\ell\right).$$

Mode orthogonality for the scalar product gives rise to some very useful properties (Berkooz et al. 1993; Cordier and Bergmann 2006):

$$\textit{(decorrelation)} \quad \mathbb{E}(a_i a_j) = \begin{cases} \Lambda_i, \text{ if } i = j, \\ 0, \text{ else,} \end{cases} \qquad \forall i, j \in 1,\ \mathbb{N},\ 1 \leq i, j \leq L, \tag{IV.9a}$$

$$\textit{(optimality)} \qquad \sum_{\ell=1}^{n} \mathbb{E}(a_\ell a_\ell) = \sum_{\ell=1}^{n} \Lambda_\ell \leq \sum_{\ell=1}^{n} \mathbb{E}(b_\ell b_\ell), \quad \forall n < N, \tag{IV.9b}$$

with $\mathbb{E}$ the expectation operator which takes place in the ensemble space in our case, i.e. $\mathbb{E}(f) = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} f^{(i)}$. The optimality property (Eq. IV.9a), specifies that, whatever the truncation, the POD decomposition is the linear combination that reproduces the most variance of the original set. This property motivates the use of POD in many studies, in particular, it can be linked to kinetic energy maximization when POD is built for an ensemble of snapshots of instantaneous velocity fluctuations fields (Berkooz et al. 1993).

In the general POD–GPRs framework, we want to predict the POD coefficients of any field $\mathbf{y}$ using GPRs (Fig. IV.1). It is therefore interesting to rescale the latent space so that the regression problem can be better posed. First, note that the POD coefficients are already centered in average since the train fields are centered by the preprocessing operator $\mathcal{T}$ (see Sect. IV.2.4) and the POD projection is linear (Eq. IV.7). Then, we introduce the POD reduced coefficients $k_\ell = a_\ell / \sqrt{\Lambda_\ell}$. This technique, sometimes called *whitening* (Kessy et al. 2018), ensures that the POD reduced coefficients have unit component-wise variances:

$$\mathbb{E}(k_\ell) = 0, \quad \forall \ell \in \mathbb{N},\ 1 \leq \ell \leq L, \tag{IV.10a}$$

$$\text{Cov}(k_i, k_j) = \begin{cases} 1, \text{ if } i = j, \\ 0, \text{ else,} \end{cases} \qquad \forall i, j \in 1,\ \mathbb{N},\ 1 \leq i, j \leq L. \tag{IV.10b}$$

The field decomposition now reads:

$$\mathcal{T}(\mathbf{y}) = \sum_{\ell=1}^{L} \sqrt{\Lambda_\ell} k_\ell \boldsymbol{\psi}_\ell, \tag{IV.11}$$

Note that the properties (Eq. IV.9–IV.10) only hold for the coefficients associated with the projection of the fields of the train set $\left\{\mathbf{y}_{\text{LES}}(\boldsymbol{\theta}^{(i)})\right\}_{i=1}^{N_{train}}$. Nevertheless, given the density of the Halton sequence (Fig IV.5) and since we expect the response surface of the model to be fairly regular locally, we assume that these properties are verified in a broader sense, i.e. hold for any set of fields $\mathbf{y}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$. Numerical tests show that, in our case, the properties IV.10 are approximately verified over the test set.

IV.2.3 Gaussian process regression

Once the POD reduced basis is computed, we seek to predict the POD reduced coefficients $\{k_\ell(\boldsymbol{\theta})\}_{\ell=1}^{L}$ for any new wind conditions $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$, knowing the POD reduced coefficients of the train set. Examples of these training point clouds for different mode orders ℓ are given as a function of the inlet wind direction α_{inlet} in Fig. IV.10. This vector regression problem is then simplified by splitting it into L scalar regression subproblems, as illustrated in Fig. IV.1. This is motivated by the fact that the POD reduced coefficients are assumed to be decorrelated as explained in Sect. IV.2.2. Then, we choose to use Gaussian Process Regressors (GPRs) to solve these regression problems because GPRs predict complete probability distributions and not only point-wise estimates which is particularly in line with the approach of this thesis. It allows us to embed the internal variability uncertainty in GPRs predictions (as shown in Sect. IV.5.3) and then to account for this uncertainty in the data assimilation system built in Chapter V. In addition, the comparison with other classical statistical methods such as polynomial chaos expansion and decision tree shows that GPRs are well-suited for this LES dispersion model reduction context (Nony 2023).

Gaussian processes (Rasmussen et al. 2006; Ebden et al. 2008), also often referred to as kriging in geostatistics (Stein 1999), is a popular and rather universal statistical learning method that consists is assuming that data distribution can be described by a Gaussian stochastic process. In our context, it means:

$$k_\ell = f_\ell(\boldsymbol{\theta}) + \epsilon_\ell \tag{IV.12}$$

with:

$$\begin{cases} f_\ell(\boldsymbol{\theta}) \sim \mathcal{GP}(m_\ell(\boldsymbol{\theta}), r_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}^*)),\ \forall (\boldsymbol{\theta}, \boldsymbol{\theta}^*) \in \Omega_{\boldsymbol{\theta}}, & \text{(IV.13a)} \\ \epsilon_\ell \sim \mathcal{N}(0, s_\ell^2), & \text{(IV.13b)} \end{cases}$$

where m_ℓ and r_ℓ are the mean and covariance functions of the Gaussian process, and ϵ_ℓ is an additive Gaussian noise of variance s_ℓ^2. It accounts for the fact that the data is noisy which is typically the case in this study, as shown in Fig. IV.7 and IV.9 in physical and latent spaces, respectively. In addition, we do the common assumption (Ebden et al. 2008) that the prior distribution of each GP f_ℓ is null everywhere in average, i.e. $\forall \boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}, m_\ell(\boldsymbol{\theta}) = 0$. In our case, this is supported because the POD reduced coefficients are centered in average (Eq. IV.10a).

From the GP modeling assumption (Eq. IV.13a) follows the key property that any finite subset $\mathbf{K}$ of realizations of the output data k_l follows a multivariate Gaussian distribution. In particular, if we denote $(\boldsymbol{\theta}^{train}, \mathbf{K}_\ell^{train})$ the train set of the POD ℓth-reduced coefficients and associated inputs parameters, and $\boldsymbol{\theta}^*$ a new set of inputs for which we want to predict the POD reduced coefficients $\mathbf{K}_\ell^*$, we have the prior joint distribution:

$$\begin{bmatrix} \mathbf{K}_\ell^{train} \\ \mathbf{K}_\ell^* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} r_\ell(\boldsymbol{\theta}^{train}, \boldsymbol{\theta}^{train}) + \sigma_\ell^2 & r_\ell(\boldsymbol{\theta}^{train}, \boldsymbol{\theta}^*) \\ r_\ell(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{train}) & r_\ell(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \end{bmatrix}\right). \tag{IV.14}$$

Then, the posterior distribution is obtained by conditioning this prior joint distribution to the observations, which gives (Rasmussen et al. 2006):

$$\mathbf{K}_\ell^* \mid \boldsymbol{\theta}^*, \mathbf{K}_\ell^{train}, \boldsymbol{\theta}^{train} \sim \mathcal{N}(\boldsymbol{\mu}_\ell,\ \mathrm{Cov}(\mathbf{K}_\ell^*)), \tag{IV.15}$$

with:

$$\begin{cases} \boldsymbol{\mu}_\ell = r_l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{train}) \left[r_l(\boldsymbol{\theta}^{train}, \boldsymbol{\theta}^{train}) + s_\ell^2 \mathbf{I} \right]^{-1} \mathbf{K}_\ell^{train}, & \text{(IV.16a)} \\ \mathrm{Cov}(\mathbf{K}_\ell^*) = r_l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - r_l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{train}) \left[r_l(\boldsymbol{\theta}^{train}, \boldsymbol{\theta}^{train}) + s_\ell^2 \mathbf{I} \right]^{-1} r_l(\boldsymbol{\theta}^{train}, \boldsymbol{\theta}^*). & \text{(IV.16b)} \end{cases}$$

These equations are the predictive equations for GP regression. In the following, GPR estimates are implicitly defined by the mean of the posterior distribution (Eq. IV.16a). In addition, Eq. IV.16b provides an estimation of the uncertainty associated with the GPRs mean estimates, and random realizations of the outputs can be sampled from the posterior distribution (Eq. IV.15). To compute this distribution, we use the scikit-learn[3] implementation which is based on the algorithm proposed by Rasmussen et al. (2006) in which the matrix $[r_l(\boldsymbol{\theta}^{train}, \boldsymbol{\theta}^{train}) + s_\ell^2 \mathbf{I}]$ is inverted using Cholesky decomposition.

Specifying the covariance function r_ℓ, mostly referred to as the kernel function in the GP literature, is crucial as it is what defines the GPRs estimates (Eq. IV.16a). Kernel functions typically express the covariance in Eq. IV.13a as a function of the distance in the input space. One standard kernel choice is the Radial Basis Function (RBF) which reads:

$$r_{\mathrm{RBF}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \rho \exp\left(-\frac{1}{2} d(\boldsymbol{\theta}, \boldsymbol{\theta}')^2 \right), \tag{IV.17}$$

with ρ the maximum allowable covariance, which should be high for processes that have to cover a wide range of output (Ebden et al. 2008), and d a distance in the input space, $\mathbb{R}^d$ in our case. Following the POD–GPRs design proposed by Nony (2023), we also consider using a Matérn kernel (Stein 1999):

$$r_{\text{Matérn}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \rho \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\ d(\boldsymbol{\theta}, \boldsymbol{\theta}') \right)^\nu \mathcal{B}_\nu \left(\sqrt{2\nu}\ d(\boldsymbol{\theta}, \boldsymbol{\theta}') \right), \tag{IV.18}$$

where Γ and $\mathcal{B}_\nu$ are the Gamma and modified Bessel functions, and ν is a smoothness parameter. This parameter is generally assigned a half-integer $\nu = p + \frac{1}{2},\ p \in \mathbb{N}$. The larger ν is, the smoother the kernel, so in the limit $\nu \to \infty$ the kernel expression (Eq. IV.18) tends to the one of the RBF (Eq. IV.17). For both kernels, we use an anisotropic ℓ_2 distance in the input space:

$$d(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sqrt{(\boldsymbol{\theta} - \boldsymbol{\theta}')^{\mathrm{T}} \boldsymbol{\lambda} (\boldsymbol{\theta} - \boldsymbol{\theta}')}, \tag{IV.19}$$

where $\boldsymbol{\lambda} = \mathrm{diag}\left(1/\lambda_{\alpha_{inlet}}^2, 1/\lambda_{u_*}^2 \right)$ is the length-scale matrix assuming independent and distinct length scales for each parameter dimension. In a preliminary study (not shown here), we compared the RBF and Matérn kernel functions and found that in our case the

[3] See footnote 2 page 124.

choice of the kernel has a very limited impact on the GPRs accuracy, with slightly better performance for the Matérn kernel with $\nu = 5/2$. This is in line with the results obtained by Nony (2023). And we use this kernel function in the rest of the study.

Each of the GPRs is defined by four hyperparameters: the noise variance, the maximum allowable covariance, and the parameters length scales. We note

$$\boldsymbol{\gamma}_\ell = (s_\ell,\ \rho_\ell,\ \lambda_{\alpha_{inlet},\ell},\ \lambda_{u_*,\ell})^{\mathrm{T}} \in \mathbb{R}^{+4},$$

the vector of the hyperparameters. We use Maximimum Log-Likelihood (MLL) estimation (Hastie et al. 2009) to determine the optimal set of hyperparameters using Bayes' theorem:

$$\boldsymbol{\gamma}_\ell^* = \underset{\boldsymbol{\gamma}_\ell}{\operatorname{argmax}} \left\{ \ln p(\mathbf{K}_\ell^{train} | \boldsymbol{\theta}^{train}, \boldsymbol{\gamma}_\ell) + \ln p(\boldsymbol{\gamma}_\ell) \right\}, \tag{IV.20}$$

with $p(\mathbf{K}_\ell^{train} | \boldsymbol{\theta}^{train}, \boldsymbol{\gamma}_\ell)$ the conditional distribution of $\mathbf{K}_\ell^{train}$, knowing $(\boldsymbol{\theta}^{train}, \boldsymbol{\gamma}_\ell)$, and $p(\boldsymbol{\gamma}_\ell)$ the hyperparameters prior distribution which is assumed to be uniform in MLL estimation. In this study, this optimization problem is solved using the L-BFGS-B algorithm of Liu and Nocedal (1989). Note that it is possible to provide a more informative hyperparameter prior distribution $p(\boldsymbol{\epsilon_\ell})$ to guide the optimization (maximum a posteriori estimation). But we do not use this alternative because Nony et al. (2023a) shows that, with the same model architecture applied to a simplified case, it gives very similar results as MLL.

An example of Gaussian process regression on a simple one-dimensional case is illustrated in Fig. IV.2. The prior distribution of the Gaussian process has zero mean and constant variance over the input space (Fig. IV.2a). On the other hand, the mean of the posterior distribution conditioned by the training data effectively and smoothly interpolates the training data (Fig. IV.2b). As expected, the 95% confidence interval around the mean, which is equal to plus or minus twice the Gaussian process standard deviation by construction, is narrowed at learning points and looser as the distance increases. Note that the variance at the training sample location does not cancel. This shows that the Gaussian process noise variance s^2 found by MLL estimation is not null, hence preventing GPR from over-fitting noisy training data.

We would like to stress that, in the current model reduction context, GPRs are not used "spatially" as in kriging. Indeed, we do not use GPRs to describe a physical field based on local observations. Instead, we use them to predict the POD reduced coefficient k_ℓ for different wind conditions $\boldsymbol{\theta} = (\alpha_{inlet}, u_*)^{\mathrm{T}}$, as shown in Fig. IV.10.

IV.2.4 Fields preprocessing

In this section, we present the field preprocessing applied before building the POD reduced-basis (Fig. IV.1a). Preprocessing is used here in a broad sense to mean any field transformation such as weighting, rescaling, thresholding, etc. We encapsulate all applied transformations in the preprocessing operator $\mathcal{T}$. Finding an adequate preprocessing is critical as it changes the meaning of the optimality and orthogonality properties of the POD modes (Schmidt and Colonius 2020), and therefore conditions the ability of the POD to efficiently represent fields in a smaller dimension. Note that the inverse transformation

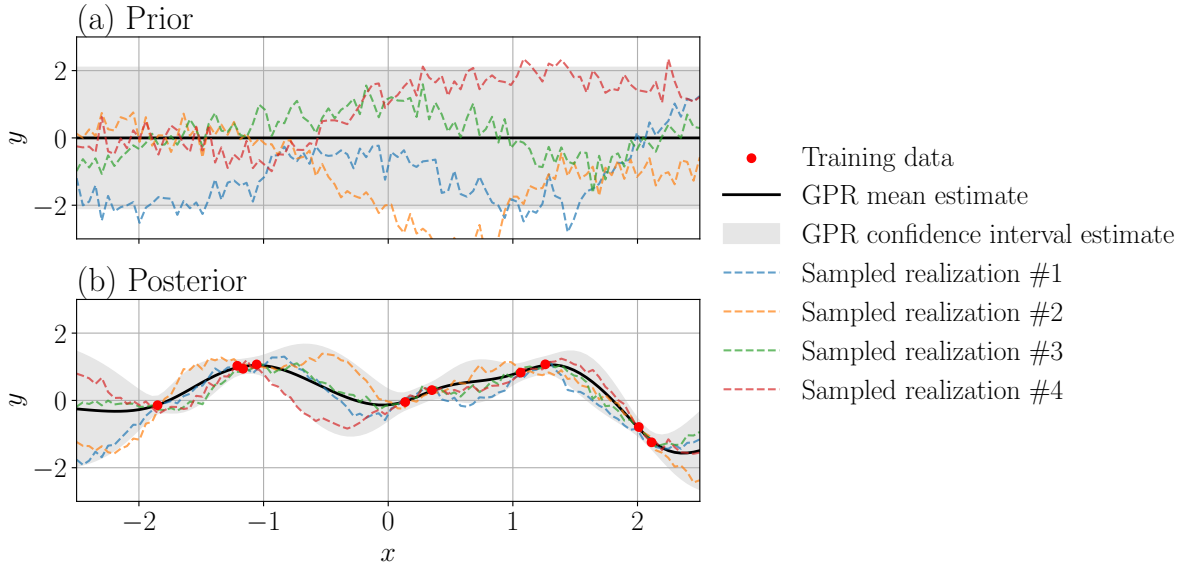


Figure IV.2: *Gaussian Process Regression principle illustrated on a one-dimensional example from scikit-learn[1]. The black line corresponds to the mean of the Gaussian process prior distribution (a), and posterior distribution (b). The training data used to build the posterior distribution are represented as red circles. They are generated at 10 random locations using the function* $x \mapsto \sin(x^2) + \epsilon$ *with* $\epsilon \sim \mathcal{N}(0, 0.15)$ *an additional Gaussian noise. Four realizations sampled from the Gaussian Process distribution are illustrated as colored dashed lines. The shaded areas correspond to the 95% confidence interval estimated by the Gaussian Process. The Gaussian process covariance function used is a Matérn kernel with* $\nu = 5/2$ *(Eq. IV.18), and the Gaussian Process hyperparameters* $\boldsymbol{\gamma} = (s, \rho, \lambda_x)^{\mathrm{T}}$ *are obtained by maximum log-likelihood estimation.*

[1]See scikit-learn documentation.

$\mathcal{T}^{-1}$ has to be applied at the end of the prediction chain (Fig. IV.1b). Different strategies are presented here, and we show in Sect. IV.5.2 that choosing the right preprocessing is essential to minimize the POD projection error.

The standard preprocessing $\mathcal{T}$ consists of centering the LES train set (Berkooz et al. 1993). Here we use in addition nodal volume weighting:

$$\begin{aligned} \mathcal{T} : \mathbb{R}^N \quad &\longrightarrow \mathbb{R}^N, \\ \mathbf{y}(\mathbf{x}_k) &\longmapsto \sqrt{\frac{\omega(\mathbf{x}_k)}{\Omega}} \left[\mathbf{y}(\mathbf{x}_k) - \langle \mathbf{y}_{\mathrm{LES}}(\mathbf{x}_k) \rangle\right], \ 1 \leq k \leq N, \end{aligned} \tag{IV.21}$$

where $\langle \mathbf{y}_{\mathrm{LES}} \rangle$ is the field averaged over the train set, $\omega(\mathbf{x}_k)$ is the nodal volume[4] of the node $\mathbf{x}_k$, and $\Omega = \sum_{k=1}^{N} \omega(\mathbf{x}_k)$ is the total volume of the domain. It is motivated by the fact that, in this study, the fields $\mathbf{y}$ are expressed on an analysis mesh which is not uniform (see Sect. II.4.1, page 61). In particular, the mesh resolution is higher in some

[4]See footnote 3, page 52.

areas, for instance near the obstacles. To avoid artificially putting too much weight on these areas, we weight the field node values by the dual volume of the node, as suggested by Schmidt and Colonius (2020). Early tests demonstrated that node volume weighting had a limited impact on the overall performances of the complete POD–GPRs models (not shown here).

Additionally, it would be possible to scale each grid node value $\mathbf{y}(\mathbf{x}_k)$ by its standard deviation over the train set. This avoids putting too much weight on areas of the domain where the variance is high but in return treats equally the variability induced by the wind conditions $\boldsymbol{\theta}$ that we want to emulate and the noise induced by internal variability. In particular, we find that rescaling the mean concentration puts too much weight on areas where the concentration is negligible and the variance is only due to numerical noise. Therefore, we choose not to scale the fields by the train-set standard deviation.

We also consider using a log-transformation in addition to the standard normalization:

$$\begin{aligned} \mathcal{T}_{log} : \mathbb{R}^N &\longrightarrow \mathbb{R}^N, \\ \mathbf{y}(\mathbf{x}_k) &\longmapsto \sqrt{\frac{\omega(\mathbf{x}_k)}{\Omega}} \left[\ln(\mathbf{y}(\mathbf{x}_k) + y_t) - \langle \ln(\mathbf{y}_{\mathrm{LES}}(\mathbf{x}_k) + y_t) \rangle\right], \ 1 \leq k \leq N, \end{aligned} \tag{IV.22}$$

with y_t a threshold to avoid issues with values close to zero. This transformation is particularly in line with the mean concentration field, which decreases exponentially with distance from the source, thus varying over several orders of magnitude (as shown in by the standard Gaussian plume model defined in Eq. I.15, page 22). With this log-transformation, the overall field reduction is no longer linear, and the field reconstruction (Fig IV.1b) relies on the bijection of $\mathcal{T}_{log}^{-1}$:

$$\begin{aligned} \mathcal{T}_{log}^{-1} : \mathbb{R}^N &\longrightarrow \mathbb{R}^N, \\ \widetilde{\mathbf{y}}(\mathbf{x}_k) &\longmapsto \exp\left(\sqrt{\frac{\Omega}{\omega(\mathbf{x}_k)}}\widetilde{\mathbf{y}}(\mathbf{x}_k) + \langle \ln(\mathbf{y}_{\mathrm{LES}}(\mathbf{x}_k) + y_t) \rangle\right) - y_t, \end{aligned} \tag{IV.23}$$

with $\tilde{\mathbf{y}}$ the log-transformed field.

In addition, the reduction of the model inputs dimension thanks to friction velocity similarity presented in Sect. IV.1.3 can be encapsulated in the preprocessing operator. For a field inversely proportional to the friction velocity u_*, such as the mean concentration, it gives:

$$\begin{aligned} \mathcal{T}_{\mathrm{1D}} : \mathbb{R}^N \times \mathbb{R} &\longrightarrow \mathbb{R}^N, \\ (\mathbf{y} \ u_*) &\longmapsto \mathcal{T}(\mathbf{y} \times u_*), \end{aligned} \tag{IV.24}$$

$$\begin{aligned} \mathcal{T}_{\mathrm{1D}}^{-1} : \mathbb{R}^N \times \mathbb{R} &\longrightarrow \mathbb{R}^N, \\ (\widetilde{\mathbf{y}}, \ u_*) &\longmapsto \mathcal{T}^{-1}(\widetilde{\mathbf{y}})/u_*. \end{aligned} \tag{IV.25}$$

The effect of different field preprocessing treatments on the POD projection error is assessed in detail in Sect. IV.5.2.

IV.2.5 Reduced-order model uncertainty prediction

In this section, we further develop the POD–GPR approach adopted by Nony et al. (2023a) by explaining how to propagate the probability distributions estimated by the GPRs in the latent space to the physical space (Fig. IV.1). From an uncertainty quantification perspective, knowing these distributions is very useful as it gives confidence intervals around the predictions of the reduced-order model. We use it extensively for the validation of the POD–GPRs in Sect. IV.6, where we also demonstrate that the estimated uncertainty reproduces well the aleatory uncertainty related to internal variability. It is also particularly valuable for data assimilation, as it can be used to prevent putting too much confidence in the model during the analysis (see Chapter V).

First, we recall that POD–GPRs predictions $\mathcal{T}(\mathbf{y}(\boldsymbol{\theta}))$ are written as linear combinations of the POD reduced coefficients $k_\ell(\boldsymbol{\theta})$ (Eq. IV.11). But, by model design assumption, $k_\ell(\boldsymbol{\theta})$ are decorrelated (Eq. IV.10b) and normally distributed since they are predicted by GPRs (Eq. IV.16). At each grid node $\mathbf{x}_k$, the POD–GPRs prediction $\mathcal{T}(\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k))$ therefore follows a normal distribution of expectation:

$$\mathbb{E}\left(\mathcal{T}(\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k))\right) = \sum_{\ell=1}^{L} \sqrt{\Lambda_\ell}\mathbb{E}(k_\ell(\boldsymbol{\theta}))\boldsymbol{\psi}_\ell(\mathbf{x}_k), \tag{IV.26}$$

and variance:

$$\mathbb{V}\left(\mathcal{T}(\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k))\right) = \sum_{\ell=1}^{L} \Lambda_\ell \mathbb{V}(k_\ell(\boldsymbol{\theta}))\boldsymbol{\psi}_\ell(\mathbf{x}_k)^2, \tag{IV.27}$$

with $\mathbb{E}(k_\ell(\boldsymbol{\theta}))$ and $\mathbb{V}(k_\ell(\boldsymbol{\theta}))$ the mean and variance the posterior distribution of the ℓth GPR (Eq. IV.16b). It now remains to determine the distribution of the bijection $\mathcal{T}^{-1}$ of $\mathcal{T}(\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k))$ to express the distribution and hence the uncertainty of the physical field.

- If using the standard preprocessing $\mathcal{T}$ (Eq. IV.21), it is straightforward as $\mathcal{T}^{-1}$ is a linear operator, and the rescaled predictions are therefore also normally distributed:

$$\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k) \sim \mathcal{N}(\mu(\boldsymbol{\theta}, \mathbf{x}_k), \sigma(\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k))^2),\ 1 \leq k \leq N, \tag{IV.28}$$

  with

$$\begin{cases} \mu(\boldsymbol{\theta}, \mathbf{x}_k) = \sqrt{\dfrac{\Omega}{\omega(\mathbf{x}_k)}} \displaystyle\sum_{\ell=1}^{L} \sqrt{\Lambda_\ell} k_\ell(\boldsymbol{\theta})\boldsymbol{\psi}_\ell(\mathbf{x}_k) + \langle \mathbf{y}_{\mathrm{LES}}(\mathbf{x}_k) \rangle, & \text{(IV.29a)} \\ \sigma(\boldsymbol{\theta}, \mathbf{x}_k)^2 = \left(\dfrac{\Omega}{\omega(\mathbf{x}_k)}\right) \displaystyle\sum_{\ell=1}^{L} \Lambda_\ell \mathbb{V}(k_\ell(\boldsymbol{\theta}))\boldsymbol{\psi}_\ell(\mathbf{x}_k)^2. & \text{(IV.29b)} \end{cases}$$

  where $\langle \mathbf{y}_{\mathrm{LES}} \rangle$ is the field averaged over the train set.

- If using the log-transformation $\mathcal{T}_{log}$ (Eq. IV.22), we find that $\ln(\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k) + y_t)$ follows a normal distribution, implying that rescaled predictions are log-normally distributed:

$$\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k) + y_t \sim \text{log-}\mathcal{N}(\mu(\boldsymbol{\theta}, \mathbf{x}_k), \sigma(\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k))^2),\ 1 \leq k \leq N, \tag{IV.30}$$

with

$$\mu(\boldsymbol{\theta}, \mathbf{x}_k) = \sqrt{\frac{\Omega}{\omega(\mathbf{x}_k)}} \sum_{\ell=1}^{L} \sqrt{\Lambda_\ell} k_\ell \boldsymbol{\psi}_\ell(\mathbf{x}_k) + \langle \ln(\mathbf{y}_{\mathrm{LES}} + y_t) \rangle, \tag{IV.31}$$

and with $\sigma(\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k))^2$ defined as in Eq. IV.29b. In this case, the variance of the field can be expressed as follows:

$$\mathbb{V}(\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k)) = \mathbb{V}(\mathbf{y}(\boldsymbol{\theta}, \mathbf{x}_k) + y_t) = [e^{\sigma(\boldsymbol{\theta}, \mathbf{x}_k)^2} - 1] e^{(2\mu(\boldsymbol{\theta}, \mathbf{x}_k) + \sigma(\boldsymbol{\theta}, \mathbf{x}_k)^2)}. \tag{IV.32}$$

Note that when normalizing the fields by the friction velocity with $\mathcal{T}_{\mathrm{1D}}$ (Eq. IV.24), the distributions of the physical fields predicted by the POD–GPR are obtained by rescaling the distributions (Eq. IV.28 or Eq. IV.30) by $\left(\frac{Q_{source}}{u_* H^2}\right)$ for the expected value and $\left(\frac{Q_{source}}{u_* H^2}\right)^2$ for the variance.

IV.3 Reduced-order model validation methodology

In this section, we present the methodology used to validate the POD–GPRs reduced-order model. First, we introduce another trivial model reduction approach, the nearest neighbor, which provides a benchmark for assessing the added value of the more sophisticated POD–GPRs approach (Sect. IV.3.1). Then we introduce the metrics employed to quantify the model reduction over the test set (Sect. IV.3.2), and the cross-validation procedure which allows us to assess the robustness of the POD–GPRs to the composition of the train set (Sect. IV.3.3). Finally, we present how the internal variability error is quantified to provide an estimate of an upper limit of the best accuracy achievable by the POD–GPRs.

IV.3.1 A control model: the nearest neighbor

To provide a reference against which to compare the POD–GPRs model reduction approach, we use the nearest neighbor model (1–NN) as a trivial control model. This model is a simplified version of the classical regression k-nearest N neighbor (k–NN) model (Hastie et al. 2009) with only one neighbor ($k = 1$). This means that the 1–NN surrogate simply predicts the field of interest $\mathbf{y}$ as equal to the nearest LES train field in the parameter space:

$$\mathbf{y}_{\mathrm{ROM}}(\boldsymbol{\theta}) = \mathbf{y}_{\mathrm{LES}}^{train}(\boldsymbol{\theta}^*), \text{ with } \boldsymbol{\theta}^* = \min_{1 \leq i \leq N_{train}} d(\boldsymbol{\theta}_i^{train}, \boldsymbol{\theta}), \tag{IV.33}$$

where d is a distance in $\mathbb{R}^d$. In this work, we use the standard Euclidean norm but with specific parameters rescaling:

$$\widetilde{\boldsymbol{\theta}} = (\widetilde{\alpha}_{inlet}, \widetilde{u}_*)^{\mathrm{T}} = \left(\frac{\alpha_{inlet} - \alpha_{inlet}^{min}}{\alpha_{inlet}^{max} - \alpha_{inlet}^{min}}, \ \zeta \left(\frac{u_* - u_*^{min}}{u_*^{max} - u_*^{min}} \right) \right)^{\mathrm{T}}, \tag{IV.34}$$

with ζ a rescaling factor that distorts the distances in the parameter space.

This hyperparameter allows giving more or less weight to one parameter when looking for which LES field is the closest in Eq. IV.33. A cross-validation procedure (see Sect. IV.3.3), with 8-fold resampling of the train set, is used to calibrate this hyperparameter. For the mean concentration field, we find an optimal value of $\zeta = 0.275$ which gives the best compromise between the RMSE, VG and FMS(1 ppm) scores. This value $\zeta < 1$ reduces distances along the axis of friction velocity, which means that the 1–NN is freer to select samples with friction velocities that are far apart, and thus tends to seek out the nearest neighbor with very close wind directions. This is consistent with our general observation that concentration is mainly controlled by the wind direction (see Fig. IV.7, and Sect. III.5.2, page 109).

The 1–NN is a pertinent control model because it represents the generalization error obtained by simply querying the available simulation database. Thus, if a more sophisticated model does not reduce the NN error then it is not worth using.

IV.3.2 Model reduction error quantification

In statistical learning, the fundamental principle of model validation is to evaluate the model accuracy on a set of test samples independent of the train set used to build the model (Hastie et al. 2009). This is essential to assess the ability of the reduced-order model to transpose the information learned from the train set to other operating conditions, in our case different wind conditions $\boldsymbol{\theta} = (\alpha_{inlet},\ u_*)^{\mathrm{T}}$, and to detect possible overfitting of the train set. In this study, we use 80% of the LES samples as train samples ($N_{train} = 160$) and the remaining 20% as test samples. The resulting train and test sets are depicted in Fig. IV.5.

To evaluate the reduced-order model accuracy after training, we then compare its mean concentration field prediction $\mathbf{y}_{\mathrm{ROM}}$ with the one from the LES model $\mathbf{y}_{\mathrm{LES}}$ for each of the N_{test} test samples. The comparison can be assessed using any of the standard air quality metrics of Chang and Hanna (2004) presented in detail in Sect. III.3.2. The overall accuracy of the reduced-order model is then assessed by averaging the scores obtained with the chosen metric f over the whole test set:

$$\langle f \rangle = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} f(\mathbf{y}_{\mathrm{ROM}}(\boldsymbol{\theta}_i^{test}),\ \mathbf{y}_{\mathrm{LES}}(\boldsymbol{\theta}_i^{test})). \tag{IV.35}$$

Instead of comparing spatially the predicted fields and then averaging the metrics scores over the test set (Eq. IV.35), it is also possible to do the opposite, i.e. calculate metrics over the test set and then average it spatially. The two approaches are complementary, as the first describes the reduced-order model accuracy per sample while its counterpart evaluates the accuracy for each node of the domain. The second approach focuses more on the ability of the model to reproduce the test set statistics, for example, the Q2 metric, extensively used by Nony (2023), quantifies the variance explained by the reduced-order model:

$$Q2(\mathbf{x}) = 1 - \frac{\|\mathbf{y}_{\mathrm{LES}}(\mathbf{x}) - \mathbf{y}_{\mathrm{ROM}}(\mathbf{x})\|^2}{\|\mathbf{y}_{\mathrm{LES}}(\mathbf{x}) - \langle \mathbf{y}_{\mathrm{LES}}(\mathbf{x}) \rangle\|^2}, \tag{IV.36}$$

with $\|\cdot\|$ the Euclidean norm, and $\langle \cdot \rangle$ the mean value over the test set. Q2 variation range is $]-\infty,\ 1]$ with $Q2 = 1$ the best possible score. A global score can then be calculated by taking the local variance-weighted average over the whole domain:

$$Q2 = \frac{\sum_{k=1}^{N} \mathbb{V}(\mathbf{y}_{\mathrm{LES}}(\mathbf{x}_k)) Q2(\mathbf{x}_k)}{\sum_{n=k}^{N} \mathbb{V}(\mathbf{y}_{\mathrm{LES}}(\mathbf{x}_k))}. \tag{IV.37}$$

The choice of metrics used for validation is crucial and mostly depends on the user's needs. For example, if one is interested in predicting concentrations that exceed a toxicity threshold, the FMS (Eq. III.18) would probably be the most important metric as it evaluates the accuracy of the predicted exposure maps predicted relative to this threshold. In this thesis, we decide not to favor any particular metric, which makes it possible to further assess the strengths and weaknesses of the reduced-order model. The counterpart of multi-criteria validation is that optimality may be impossible to define as

improving one metric often deteriorates the others. These considerations are not only key for model validation but also model optimization, however multi-objective optimization makes the task much more difficult and is outside the scope of this applied study.

IV.3.3 Model cross-validation

When developing and validating the reduced-order model, we realized that reshuffling the train/test samples could occasionally have a significant impact on the final performances evaluated on the test set. This sensitivity can be explained by the presence of outliers in the original data set and the limited size of the learning base.

To quantify this variability of the reduced-order model prediction errors, we adopt the K-fold cross-validation procedure classically used in statistical learning (Hastie et al. 2009). Its principle is to: i) divide the samples set into K blocs; ii) calibrate the reduced-order model on a train set of $K-1$ blocs, and use the remaining bloc as a test set to evaluate the accuracy of the surrogate; iii) repeat the last step K times so that the test set is different each time.

Most of the time, cross-validation is used for model tuning or model selection. The original train set is split K times into a new train subset and a validation (or calibration) subset. Then the reduced-order model hyperparameters are selected based on the averaged performances over the K validation subsets. Cross-validation allows to ensure robust model reduction tuning by selecting averaged hyperparameters and limiting the generalization error. This method is used to calibrate the Nearest Neighbor model presented in Sect. IV.3.1.

In addition, we use cross-validation on the complete database (train and set) for the final evaluation of the mean reduced-order model prediction error. This allows for providing in addition an order of magnitude of the variance of this error. This also enables the assessment of the model performance for each sample of the complete database. Associated results are given in Sect. IV.6.3.

The choice of the number of blocks is a matter of compromise because increasing K improves the quality of the estimation but also increases the cost of the procedure (Hastie et al. 2009). In this study, we use a relatively small value of $K = 5$ which results in a train set of 160 samples. Figure IV.5 represents the train set of 160 simulations and the test set of 40 simulations, as blue squares and green circles respectively, for the first validation fold.

IV.3.4 Accounting for internal variability in the validation

As previously explained in Chapter III, LES predictions are subject to an irreducible aleatory uncertainty due to the internal variability of the ABL. It would therefore be pointless to try to build a reduced-order model whose accuracy exceeds this uncertainty. To quantify the error solely due to internal variability, we use the stationary bootstrap method presented in Sect. III.2.4, page 83, to generate two independent sets of bootstrap replicates of the same LES field and then compare these replicates as shown in Fig. IV.3. These comparisons can be carried out for any metric f, including the air quality metrics

presented in Sect. III.3.2, page 90. In this way, we obtain a set of realizations of f from which we can estimate statistics, and in particular, its mean value which characterizes the mean error made when comparing two LES predictions obtained with exactly the same inputs, taking into account the internal variability involved. This mean error corresponds, for a given metric and a given field, to an upper limit of estimation accuracy.

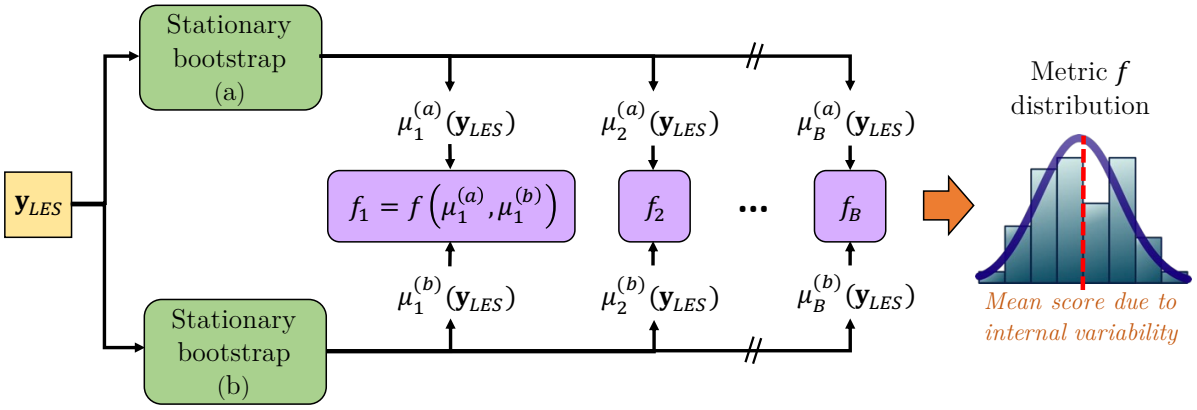


Figure IV.3: *Schematic diagram of internal variability error quantification. Two independent sets of replications of a mean field $\mathbf{y}_{LES}$ predicted by the LES model are calculated from sub-averages samples using the stationary bootstrap algorithm presented in Sect. III.2.4, page 83. Each replication is then compared two by two using a metric f to obtain a set of B scores quantifying the difference between two fields due to internal variability. Statistics can then be calculated on the distribution thus obtained, in particular the mean error induced by internal variability.*

The internal variability error quantification method shown in Fig. IV.3, is applied for each LES prediction sample in the train and test sets. Note that the mean block length used by the stationary bootstrap is re-evaluated for each sample using the approach explained in Sect. III.2.4.3, page 86. Once averaged over the test set, the internal variability errors give an upper bound estimate of the best overall accuracy achievable for each metric when validating the POD–GPRs reduced-order model. An analysis of the distribution of the internal variability errors as a function of input parameters $\boldsymbol{\theta} = (\alpha_{inlet}, u_*)^{\mathrm{T}}$ is given in Sect. IV.6.4.

IV.4 LES ensemble generation

This section describes the construction of the ensemble of LES predictions used to train and test the POD–GPRs. First, relevant wind boundary condition parameter ranges are estimated from a mirco-climatology study (Sect. IV.4.1). The parameter space thus defined is sampled using the low-discrepancy Halton sequence (Sect. IV.4.2). The LES model is then integrated for each input parameter corresponding to the first 200 samples of the Halton sequence. All the LES model adaptations that were made to compute such a large ensemble of simulations are presented in Sect. IV.4.3. Finally, we provide technical insights on the computation of the ensemble in Sect. IV.4.4, as well as a brief overview of the LES response surface.

IV.4.1 Definition of parameter ranges from micro-climatology

The reduced-order model has to cover a wide, but plausible and feasible, range of variation of the input parameters $\boldsymbol{\theta} = (\alpha_{inlet}, u_*)^{\mathrm{T}}$. In what follows, we propose to use micro-climatology, i.e. a compilation of near-surface wind measurements over a long period of time, to determine the lower and upper limits of α_{inlet} and u_*.

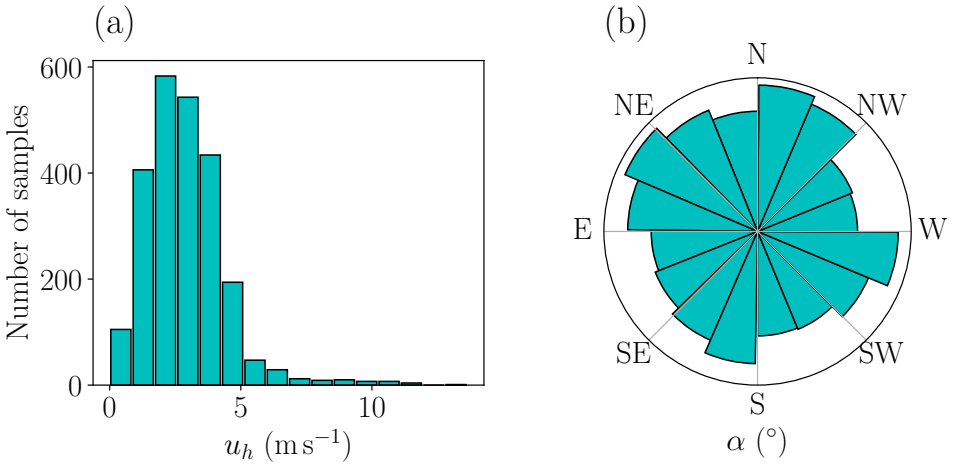


Figure IV.4: *Distributions of the (a) time-averaged horizontal wind velocity and (b) time-averaged wind direction measured at the SAMS meteorological station #8 at $z = 10\,m$ and over 12 days during the MUST field campaign.*

The micro-climatology is built using all available data from the SAMS meteorological station #8 located approximately one kilometer from the MUST site (Fig. II.2, page 58). A total of 2391 wind measurements at $z = 10\,\mathrm{m}$, averaged over 15-minute periods, were collected over 12 days during the MUST field campaign, the resulting distributions are depicted in Fig. IV.4. It appears that all wind directions are likely to appear and that more than 99% of the horizontal wind speed measurements are under $12\,\mathrm{m\,s^{-1}}$ which corresponds to friction velocity of $0.89\,\mathrm{m\,s^{-1}}$ (Eq. II.17 page 63).

To reduce the number of LES computations required to cover the input space, the range of variation for the inlet wind direction α_{inlet} was narrowed to $[-90\,^\circ,\ 30\,^\circ]$, so

the corresponding plume always remains mostly in the canopy and therefore at the level of existing sensors. The extension of this range of variation would not be a problem, provided that a sufficient simulation budget is available. We also limit the minimal friction velocity to $0.07\,\mathrm{m\,s^{-1}}$ which corresponds to a wind speed of about $1\,\mathrm{m\,s^{-1}}$ at $10\,\mathrm{m}$ altitude as we are interested in windy conditions. In the end, the input parameter space reads:

$$\Omega_{\boldsymbol{\theta}} = [-90\,^\circ,\ 30\,^\circ] \times [0.07\,\mathrm{m\,s^{-1}},\ 0.89\,\mathrm{m\,s^{-1}}]. \tag{IV.38}$$

IV.4.2 Parameter space sampling

To sample the input parameter space (Eq. IV.38) we use Halton's sequence (1964). As a low-discrepancy sequence, it samples the space uniformly and covers it more efficiently than a purely random sequence by avoiding sampling the same area several times. Since samples tend to be equidistributed, no a priori assumptions are made about the distribution of input parameters for the construction of the reduced-order model. In this way, we expect it to perform equally well in any situation. Figure IV.5 shows the first 200 draws from the Halton sequence.

In the Halton sequence, each parameter vector sample has different values of α_{inlet} and u_*. When normalizing the fields by the friction velocity to build a one-dimensional reduced-order model (see Sect. IV.1.3), this property prevents creating multiple antecedents for a single value of α_{inlet} when collapsing the parameter space into one dimension as shown in Fig. IV.5b. This is one of the strengths of the Halton sequence: if we were to use regular sampling of the parameter space instead, we would lose $N_{samples} - \sqrt{N_{samples}}$ out of the $N_{samples}$ when reducing the problem dimension.

We then compute an LES prediction for each input parameter sample to provide the database used to build the reduced-order model. To capture the nonlinear behavior of the LES model and guarantee the accuracy of the reduced-order model, we have set ourselves a budget of $N_{samples} = 200$ simulations corresponding to the first draws from the Halton sequence. As shown in Fig. IV.5, the resulting sampling is quite dense, particularly for the reduced-dimension space. In Sect. IV.6.2, we investigate the effect of reducing the size of the learning base on the performance of reduced models.

Ultimately, the samples obtained are separated into a train set and a test composed respectively of the first $N_{train} = 160$ samples and the next $N_{test} = 40$ samples. The samples of the test set are not used during training to ensure proper validation of the reduced models (see Sect. IV.3).

IV.4.3 Adaptation of the LES model to run ensembles

To explore the model answer to its inflow boundary conditions uncertainties, several modifications have been made so that the LES model presented Sect. II.4 can predict the fields of interest for every parameter $\boldsymbol{\theta} = (\alpha_{inlet}, u_*)$ possible. The model is then integrated for each parameter of the Halton sequence presented in Sect. IV.4 to generate the LES ensemble used to build the reduced-order model.

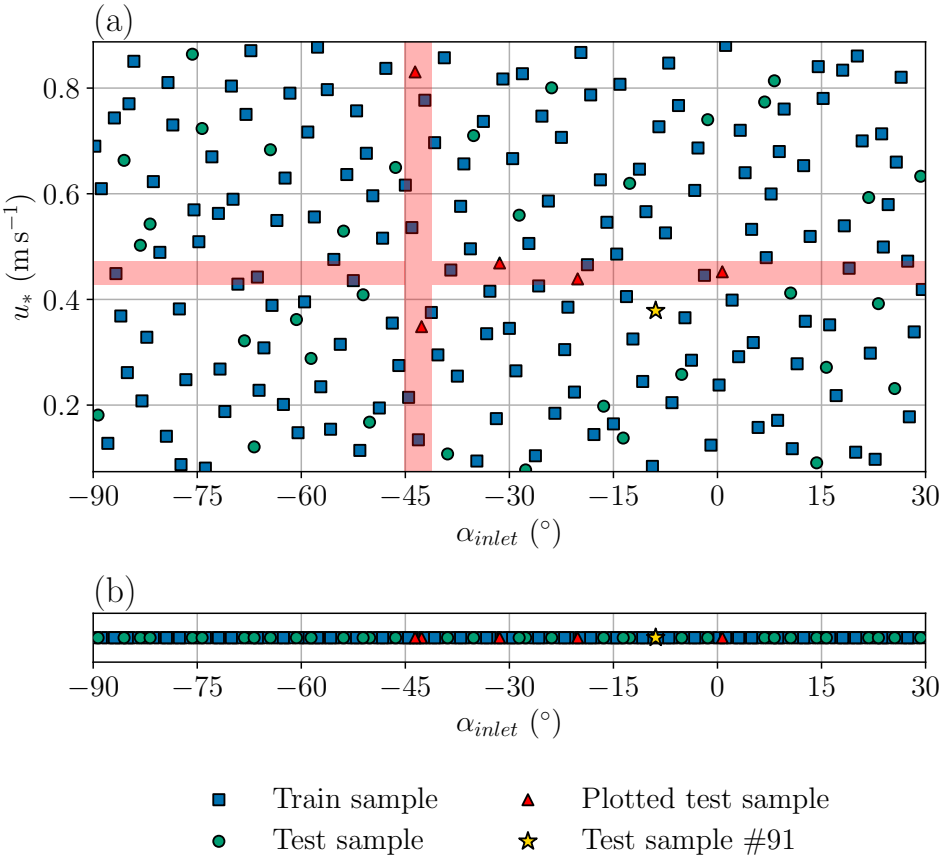


Figure IV.5: *Input parameter space sampling obtained with Halton's low-discrepancy sequence (a): each point is a pair of parameters for which we perform an LES prediction. The training and test sets are represented as blue squares and green circles respectively. The horizontal red shaded area corresponds to the parameter space sub-section scanned by taking a margin of* ± 5 *% around the constant friction speed* $u_*^{plot} = 0.45\,m\,s^{-1}$*. The vertical shaded area is similarly defined around the constant inlet wind direction* $\alpha_{inlet}^{plot} = -43°$ *with a margin of* $\pm 2°$*. The test samples that are within these ranges are depicted as red triangles, they are used in Sect. IV.6.1 to evaluate the response surfaces of the reduced-order model. Another specific test sample (#91) used to validate the models is identified by a yellow star. The projection of the parameter space into one dimension obtained by normalizing the output fields by the friction velocity is represented in the second panel (b).*

IV.4.3.1 Computational domain adaptation to the mean wind direction

In the reference LES model, if the mean flow direction deviates too much from the reference wind direction value, it induces lateral confinement and numerical instabilities because of the shear-free boundary conditions on the domain sides. This problem is solved by rotating the computational domain so the sides always remain parallel to the mean flow direction. To efficiently implement this feature, the domain is split into two subdomains as shown in Fig. IV.6: the peripheral domain D2 that is rotated to align with α_{inlet}, and the internal domain D1 which is fixed. The Navier-Stokes equations are solved

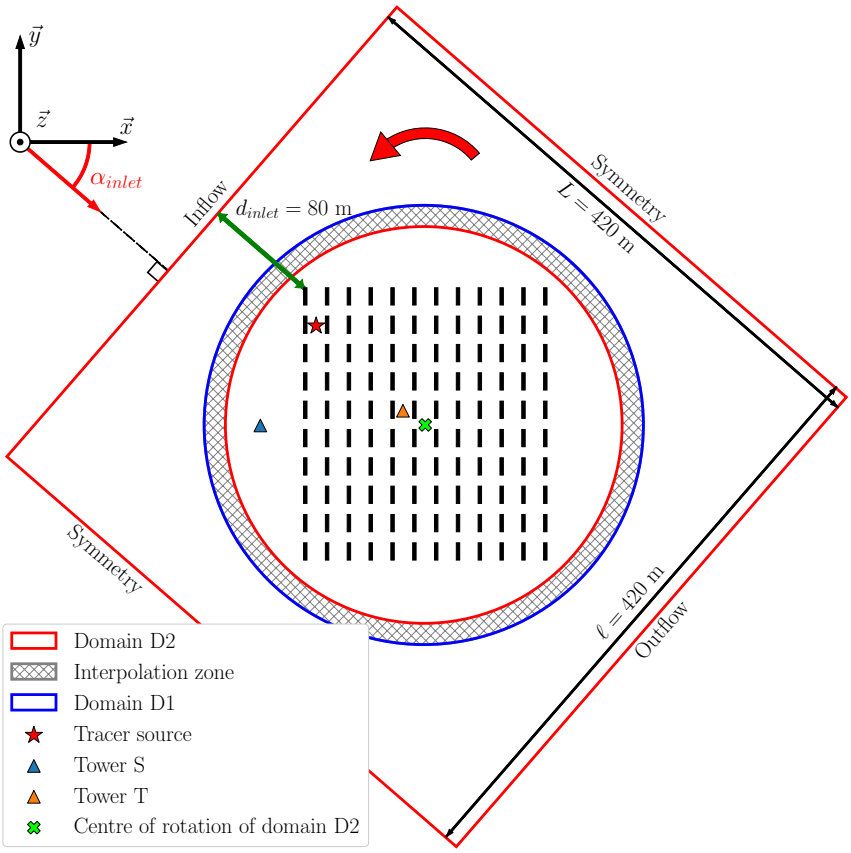


Figure IV.6: *Schematic of a horizontal cut of the MUST computational domain at the height of the container, the domain is divided into two subdomains: the fixed domain D1 in blue and the rotating peripheral domain D2 in red. The interpolation overlapping area between D1 and D2 used for the coupling is hatched. Lateral boundary conditions are indicated. The tracer source location in the MUST trial 2681829 is identified as the orange star. The triangle symbols represent the towers at which observations were acquired during the trial: the upstream tower S is indicated in green, while the internal tower T is indicated in blue.*

on each domain by parallel AVBP instances (Wang et al. 2014; Duchaine et al. 2015) that are coupled using the CWIPI library developed by ONERA[5] (Refloch et al. 2011). This domain decomposition facilitates the generation of a large ensemble of simulations because it does not require generating a new mesh for each new wind direction. In addition, having a static internal domain avoids the use of interpolation to compare LES estimates obtained with different wind conditions. The interpolation between the two domains is computed over an overlapping region (shown cross-hatched in Fig. IV.6). This region should contain at least 10 cells between the sub-domains boundaries and in each direction, which leads to a 13% increase in the number of cells of the computational mesh. Hopefully, it has less impact on the computational time since most of these cells are large.

[5]Office national d'études et de recherches aérospatiales.

IV.4.3.2 Reynolds stress tensor rescaling

To avoid running a precursor simulation to estimate the Reynolds stress tensor prescribed in the turbulence injection method (see Sect. II.4.3) for each boundary condition parameter $\boldsymbol{\theta}$ in Halton sequence, the corresponding Reynolds stress tensor $\mathbf{R}_{ij}(\boldsymbol{\theta})$ is obtained by rotating and then rescaling the reference Reynolds stress tensor $\mathbf{R}_{ij}^{(ref)}$:

$$\mathbf{R}_{ij}(\boldsymbol{\theta}) = \left(\frac{u_*}{u_*^{(ref)}}\right)^2 \times \mathbf{M}_\alpha \mathbf{R}_{ij}^{(ref)} \mathbf{M}_\alpha^{\mathrm{T}}, \tag{IV.39}$$

$$\text{with } \mathbf{M}_\alpha = \begin{pmatrix} \cos(\alpha_{inlet}) & -\sin(\alpha_{inlet}) & 0 \\ \sin(\alpha_{inlet}) & \cos(\alpha_{inlet}) & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This transformation is based on the assumption that turbulent velocities are proportional to the friction velocity u_* in the surface layer (Sect. I.1.2.4, page 19).

IV.4.3.3 Adaptation of the spin-up time

The spin-up time of the LES model has to be adapted to the friction velocity u_* since the time required to reach a steady state depends on the bulk velocity of the flow which is directly related to the friction velocity:

$$U_{bulk} = \int_0^H U(z)\mathrm{d}z = \frac{u_*}{\kappa H}\left[(H+z_0)\ln\left(\frac{H+z_0}{z_0}\right) - H\right], \tag{IV.40}$$

with H the height of the computation domain. In particular, if the friction velocity decreases, then the convective timescale increases and the injected eddies take longer to traverse the domain.

The spin-up time is therefore set by re-scaling the reference spin-up time by the friction velocity as follows:

$$t_{spin-up} = t_{spin-up}^{(ref)} \times \frac{u_*^{(ref)}}{u_*}. \tag{IV.41}$$

Given the variation range of the friction velocity considered in this study (Eq. IV.38), the corresponding spin-up times, therefore, vary between 50 and 580 s.

IV.4.3.4 Field projection onto an analysis mesh for greater efficiency

To reduce the computational burden associated with the dimension of the fields $N \approx 10^7$ (see description of the mesh used by the LES model in Sect. II.4.1, page 61), all the results presented in this section are obtained using an analysis mesh with fewer nodes. The resolution of this mesh is twice coarser than that of the mesh used by the LES model. In addition, we limit the analysis to the inner domain D1 in Fig. IV.6 and to a height of 20 m as most of the pollutant is located in this area. This leads to an analysis mesh of $N = 1.88 \times 10^6$ nodes, with characteristic cell sizes ranging from 0.6 m to 4 m, that efficiently facilitates the model reduction. In addition, it was verified that using a

lower-resolution grid has negligible impact on the accuracy of the reduced-order models. This is because little information is lost as the original resolution is significantly lower than the correlation lengths of the fields of interest.

IV.4.4 Computation of the LES ensemble

The adapted LES model is then integrated for each of the 200 parameter samples of the Halton sequence (Fig. IV.5).

The computation of the LES ensemble was performed on different supercomputers: Nemo and Kraken from CERFACS, Météo-France's Belenos, and Joliot-Curie, also known as Irene, from TTGC[6]. Access to Irene's resources was granted by GENCI[7] as part of the DARI project A0062A10822, 2020-2022. Technical characteristics of the different supercomputers are summarized in Table IV.1. The scaling of the LES model was tested for each cluster, resulting in different optimal numbers of cores (see Table IV.1). In total, the 200 simulations of the LES ensemble have cost 5.7 million core hours. An estimation of the associated carbon footprint is presented in Appendix C.

To limit the volume of data saved, instantaneous fields were not saved, apart from those required to restart simulations, and mean fields were saved every 10 s of physical time in order to enable internal variability quantification using the bootstrap approach developed earlier (Fig. III.4, page 85). In total, the conserved LES fields occupy a volume of approximately 40 To.

Table IV.1: *Main characteristics of the supercomputers used to compute the LES ensemble. N_{CPU} corresponds to the number of cores on which the LES computations were parallelized. Total consumption on each machine is given, in millions of core hours ($M\,h_{CPU}$).*

Supercomputer	Nemo[1]	Kraken[1]	Joliot-Curie[2]		Belenos
Host	CERFACS	CERFACS	TGCC		Météo-France
Partitions	Haswell	Skylake	Skylake	Rome	Rome
Processors	Intel E5-2680v3	Intel 6140	Intel 8168	AMD Epyc 7H12	AMD Epyc 7742
N_{CPU}	600 – 900	540 – 900	1344	1024	1536
$\mathrm{M\,h_{CPU}}$	0.699	0.335	1.57	1.15	1.95

[1] https://cerfacs.fr/les-calculateurs-du-cerfacs
[2] https://www-hpc.cea.fr/fr/Joliot-Curie.html

Figure IV.7 shows the topology of the resulting LES ensemble, with the example of the mean concentration at tower B at $z = 2\,\mathrm{m}$. The mean concentration reaches a maximum for $\alpha_{inlet} \approx 30\,^\circ$ and then decays towards 0 ppm in both directions as the plume no longer crosses tower B.

[6] *Très Grand Centre de Calcul*, a computing center operated by CEA (*Commissariat à l'Energie Atomique*), and which hosts one of France's three supercomputers available to the scientific community.

[7] *Grand Équipement National de Calcul Intensif*, an organization that steers the national HPC equipment strategy for the benefit of French scientific research.

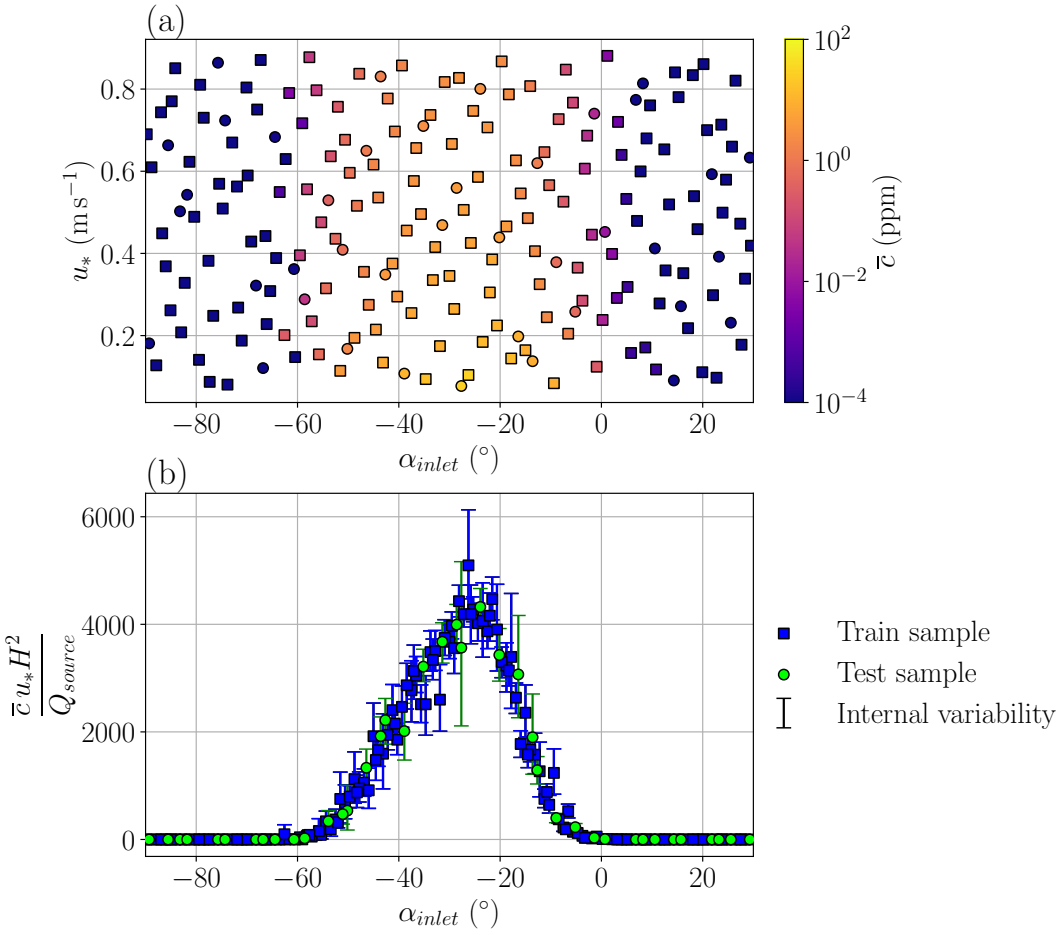


Figure IV.7: *LES prediction of the mean concentration at tower B at $z = 2\,m$ for each sample of parameters $\boldsymbol{\theta} = (\alpha_{inlet}, u_*)^{\mathrm{T}}$ from Halton sequence (Fig. IV.5). Results are given for both the non-normalized (a) and normalized mean concentration (Eq. IV.3a) (b). Train and test samples are represented as squares and circles respectively. The internal variability uncertainty is represented as error bars corresponding to the 95% confidence intervals estimated by the bootstrap procedure presented in Sect. III.2.4, page 83.*

When collapsing the LES ensemble to one dimension by normalizing the concentration (Eq. IV.3a), the mean concentration similarity hypothesis appears to not be perfectly verified: normalized concentration samples close to the same inlet wind directions α_{inlet} exhibit a significant scatter (Fig. IV.7b). The mean fields are indeed subject to internal variability as they are not statistically converged over the 200-s analysis period, as explained in Sect. III.2, page 80. The internal variability of each sample is quantified using the stationary bootstrap procedure presented in Sect. III.2.4 page 83, with mean block lengths for the stationary bootstrap re-estimated for each sample. The obtained 95% confidence intervals, shown as error bars in Fig. IV.7b, effectively explain the variability between neighboring samples.

IV.5 Setting up the POD–GPRs model

This section discusses the implementation of the POD–GPRs reduced-order model and its adaptation to the specific constraints arising from the emulation of the mean concentration field predicted by LES. In a first step, we propose an approach to select the number of reduced-basis POD modes in light of the aleatory uncertainty in the train fields due to the internal variability of the ABL (Sect. IV.5.1). This important contribution prevents the GPRs from overfitting noise during their optimization and thereby optimizes the POD–GPRs accuracy. In a second step, we investigate how the field preprocessing choice affects the POD projection error, which enables us to define best practices to deal with the wide range of scales involved in the concentration fields (Sect. IV.5.2). Finally, we validate the GPRs and their ability to infer and represent the internal variability in the training LES ensemble (Sect. IV.5.3).

IV.5.1 Statistical approach to selecting the number of modes

In this section, we propose a method to make an informed selection of the number of modes. This method builds on the bootstrap methodology developed in Chapter III to quantify the effect of internal variability aleatory uncertainty on the POD modes. Results are given for the concentration field, but the method can be generalized to any field.

The choice of the number of modes is problem-dependent and often comes down to user choice. By construction, the more the number of modes, the more variance of the original ensemble is captured in the POD reduced basis. Nony (2023) shows that using a large number of modes ($L = 100$ for a training database of 450 LES snapshots) is necessary to well represent local spatial concentration structures. However, the train samples used to build the POD are noisy, as shown in Fig. IV.7. This noise is expected to be captured by high-order POD modes (Forkman et al. 2019). In our context where the decomposition is used to build predictive reduced-order models, it is desirable to discard modes that encode mostly noise, since noise is inherently ungeneralizable.

To quantify the noise carried by each mode, we assume that, given its importance, internal variability is responsible for most of the noise in the train set and we quantify it using the bootstrap procedure proposed in Chapter III. Then, we propagate the internal variability of the fields to the POD-reduced coefficients as illustrated in Fig. IV.8. First, we build one POD reduced-order basis from the original time-averaged train samples. Then we produce B replicates of each field $\mathbf{y}^{(i)}$ of the Halton sequence, using a stationary bootstrap based on sub-average samples as detailed in Sect. III.2, page 80. Note that the mean block length used in the stationary bootstrap is re-calculated for each sample. This is important to take into account the fact that sub-averages correlation in time may vary in the parameter space, especially with the friction velocity (see Sect. IV.6.4). The field replicates are then projected on the POD basis to obtain B realizations $\left\{\mu_b(k_\ell^{(i)})\right\}_{b=1}^{B}$ of the POD reduced coefficients associated with each mode $1 \leq \ell \leq L$.

This approach is applied to the first $L = 50$ modes. And we use $B = 1\,000$ replicates, against $5\,000$ for the model validation in Chapter II, because the current procedure is quite expensive. Indeed, it requires resampling of the complete field and it is applied to

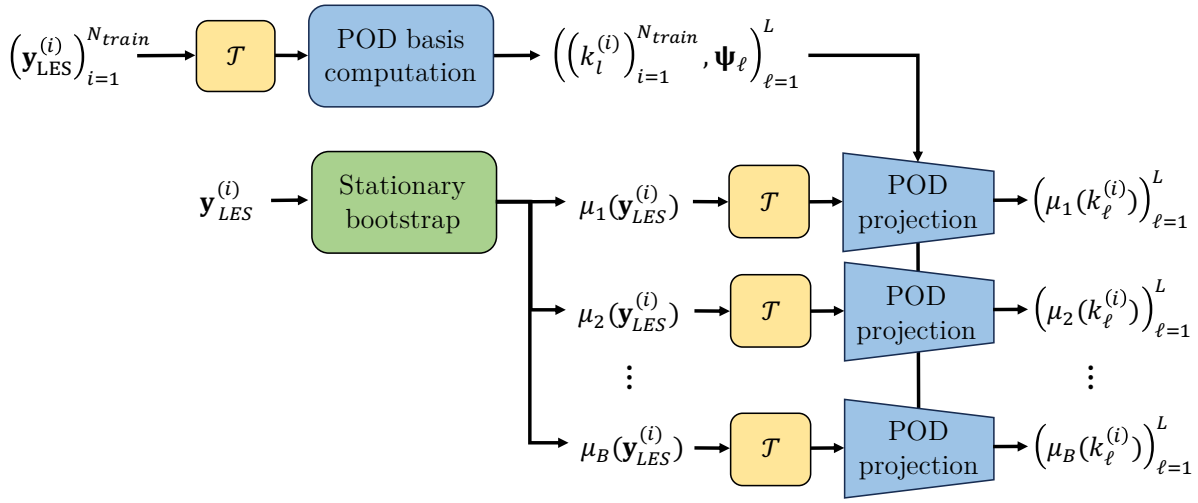


Figure IV.8: *Schematic diagram of internal variability propagation to POD reduced coefficients. First, the POD basis is computed based on the LES train set. Then bootstrap replicates $\mu_b(\mathbf{y}_{LES}^{(i)})$ of one given field $\mathbf{y}_{LES}^{(i)}$ are computed using the stationary bootstrap algorithm (Sect. III.2.4.2, page 84). Finally, each replicate is projected on the POD basis to get a set of B realizations of the POD reduced coefficients of the field $\mathbf{y}_{LES}^{(i)}$.*

each of the 200 LES samples of the Halton sequence. However, we show in Sect. III.4.3, page 103, that bootstrap estimates of the 2.5th and 97.5th percentiles are already well converged with $B = 1\,000$ replicates.

Figure IV.9a shows the difference between the L original reduced coefficients of the mean concentration train sample #11, and the first three bootstrap replicates of these coefficients as colored dashed lines. It also represents the 95% confidence interval of this difference, based on the 2.5th and 97.5th percentiles of the bootstrap ensemble. It is striking that the latter coefficients are way more sensitive to internal variability than the first ones. For this example in particular, the bootstrap replicates of the first three reduced coefficients are nearly constant, meaning that these modes correspond to systematic patterns associated with the wind conditions. Then the amount of internal variability projected on the POD reduced coefficients increases and starts to explode near $L = 10$ before reaching a plateau. This means that the physical noise that lies in the data is carried by the higher modes. And since the noise is mostly independent of the input parameters, the distribution of the reduced coefficients in the input parameter becomes increasingly chaotic as the number of modes increases, as shown in Fig. IV.10.

On the other hand, Figures IV.9b and c show that the more the number of modes the less the POD projection error. However, the gains in total variance become increasingly negligible as L increases. This supports the idea that using a small number of modes is often sufficient (Cordier and Bergmann 2006). Indeed, since the modes are sorted by decreasing eigenvalues, the largest part of the total ensemble variance is projected on the first POD modes, they are therefore the ones that represent interesting systematic patterns in the fields given the parameters $\boldsymbol{\theta}$ variability.

Note that the number of modes also affects the computational cost of the POD–GPRs

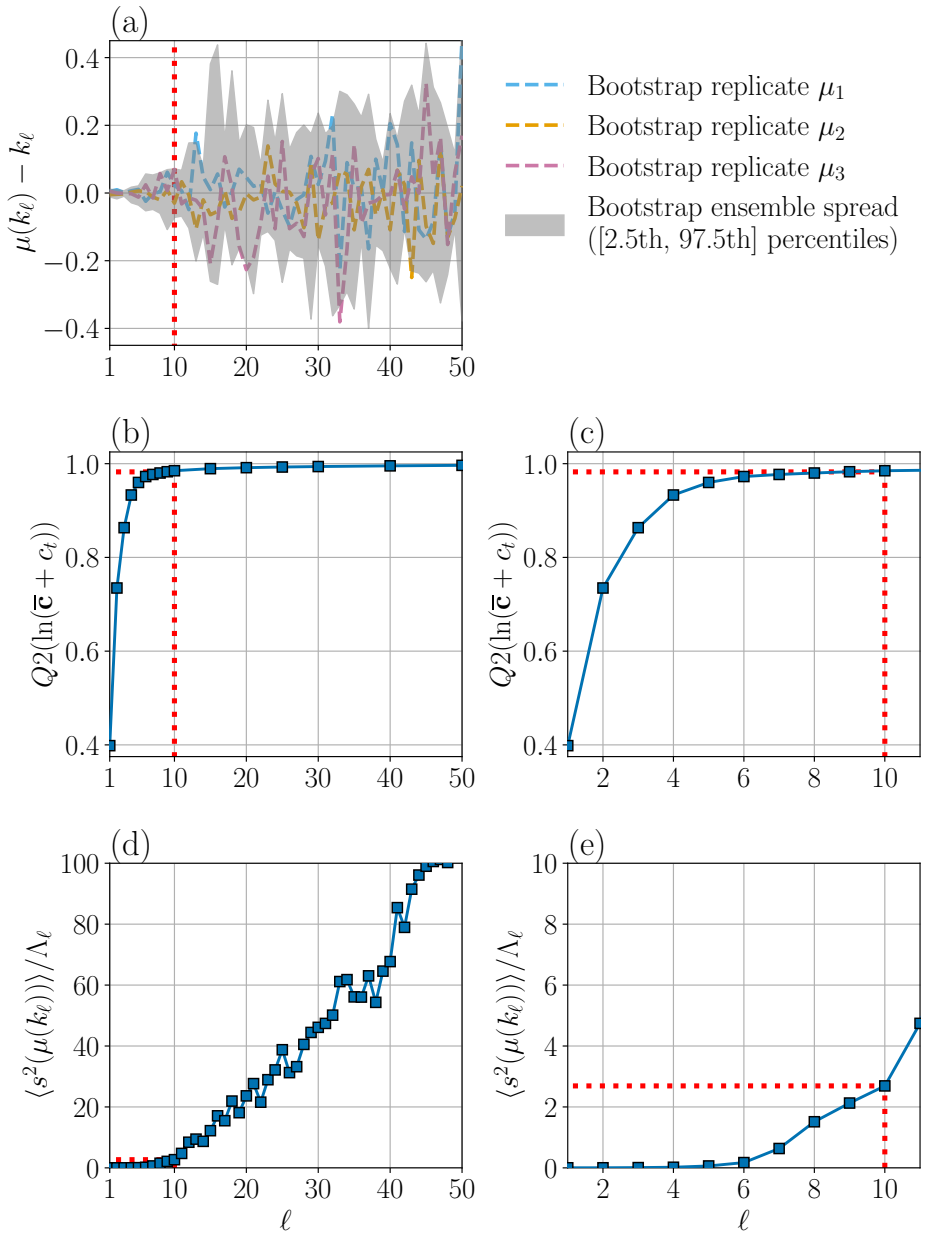


Figure IV.9: *(a) Difference between the reduced coefficients k_ℓ of the POD projection of the mean concentration train sample #11 with $(\alpha_{inlet}, u_*) = (-11\,°, 0.65\,m\,s^{-1})$, and its bootstrap replicates $\mu(k_\ell)$ obtained following the procedure illustrated Fig. IV.8. The differences with three bootstrap replicates are shown as colored dashed lines, and the grey-shaded area corresponds to the bootstrap ensemble spread of the difference (defined by the 2.5th and 97.5th percentiles). (b, c) POD projection error evaluated over the train set with the Q2 metric (Eq. IV.36) as a function of the number of modes. (d, e) The ratio between the averaged variance of the reduced coefficient bootstrap replicates $\langle s^2(\mu(k_\ell))\rangle$ and the POD eigenvalue Λ_ℓ associated with each mode ℓ. The red dotted lines indicate the number of modes selected for this study. Panels c and e are close-ups of panels b and d.*

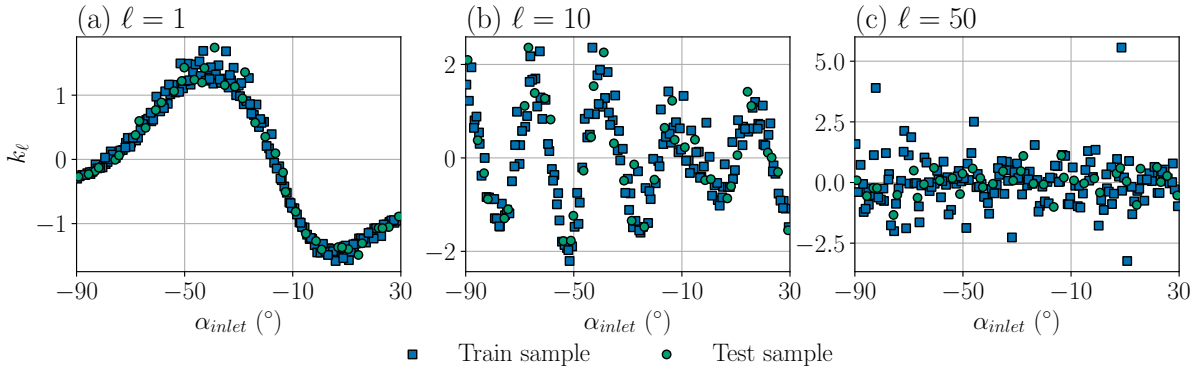


Figure IV.10: *Distribution of the reduced coefficients k_ℓ associated to the ℓth POD mode, as a function of the inlet wind direction α_{inlet} for $\ell = 1$ (a), $\ell = 10$ (b), and $\ell = 50$ (c). Squares and circles correspond to train and test samples respectively.*

reduced-order model. We find that both training and prediction scales linearly with L. This is coherent since the POD reduced-basis is computed using the randomized method of Halko et al. (2009) which only estimates the first L modes, and since L independent GPRs are optimized and used for prediction (Fig. IV.1). Nevertheless, cost is not an issue, as it is very limited, as illustrated for $L = 10$ in the Table IV.4 on page 160.

In the end, the choice of the number of modes is a matter of compromise between POD projection error minimization (Fig. IV.9b, c) and noise minimization (Fig. IV.9a). We propose to assess this compromise by calculating the ratio of the noise related to internal variability carried by the ℓth mode and the part of the total ensemble variance represented by this mode:

$$\frac{\langle s^2(\mu(k_\ell))\rangle}{\Lambda_\ell}, \tag{IV.42}$$

with $\langle s^2(\mu(k_\ell))\rangle$ the empirical variance of the bootstrap replicates of the POD reduced coefficients (see Eq. III.5, page 82) averaged over the train set, and Λ_ℓ the ℓth highest eigenvalue of the POD decomposition. Figures IV.9d and e, show that this ratio is close to zero for the six first modes and then increases sharply with the order of the mode up to approximately 100. This demonstrates that the internal variability of the high-order POD reduced coefficients dominates the part of explained variance carried by the associated modes.

Based on this observation, we choose to use $L = 10$ modes to reduce the mean concentration field in this study. The performances associated with this choice are highlighted by the dotted red lines in Fig. IV.9, with an explained variance of $Q2 = 0.98$ and a maximum ratio of between noise and explained variance reaches approximately 2.5 for the tenth mode.

We highlight that the selection method we propose has the major advantage of being entirely a priori: neither the test set nor the GPRs need to be used to choose L. However, it is a rather qualitative method and hence it only gives an approximate value for L. Defining an optimal criterion would be an interesting prospect but is outside the scope of this study. The same analysis with the $\mathcal{T}_{log\text{-}1D}$ preprocessing, which normalizes the data

by the friction speed, yields results very similar to those shown in Fig IV.9. We use $L = 10$ for both $\mathcal{T}_{log}$ and $\mathcal{T}_{log\text{-1D}}$ and throughout the rest of this study, except in Sect IV.6.5 where we compare a posteriori the results obtained by the complete POD–GPRs with 10 or 50 modes.

IV.5.2 Field preprocessing effect on proper orthogonal decomposition

We compare the POD projection error for three different field preprocessing: the standard centering (Eq. IV.21), the logtransformation (Eq. IV.22), and the log-transformation with friction velocity normalization (Eq. IV.24). To quantify the POD projection error, we first build the POD reduced basis using the train set as explained in Sect. IV.2.2. Then, each preprocessed field in the test set is projected on the POD reduced basis (Eq. IV.7). Finally, the fields are reconstructed using the POD inverse projection (Eq. IV.8) and the inverse preprocessing operator $\mathcal{T}^{-1}$. This procedure determines the loss of information associated with the POD compression-decompression.

For the log-transformation in $\mathcal{T}_{log}$ and $\mathcal{T}_{log\text{-1D}}$, we use a concentration threshold of $c_t = 10^{-4}$ ppm (Eq. IV.22). This choice provides an effective compromise that does not over-cut low concentrations and does not over-emphasize very low variances which are mainly numerical noise. It represents 100 times the minimal concentration threshold of the sensors used during the MUST field experiment.

Table IV.2: *POD projection error with $L = 10$ modes and estimated with the different metrics chosen in Sect. IV.3.2 computed over the test set. Perfect scores of the different metrics are recalled in the second row. Results are given for three different field preprocessing: the standard centering $\mathcal{T}$ (Eq. IV.21), the log-transformation $\mathcal{T}_{log}$ (Eq. IV.22), and the log-transformation with friction velocity normalization $\mathcal{T}_{log\text{-}1D}$ (Eq. IV.24).*

	Q2($\overline{\mathbf{c}}$)	Q2(ln($\overline{\mathbf{c}}$))	FB	NMSE	FAC2	MG	VG	FMS (1ppm)	FMS (0.01ppm)
Perfect score	1	1	0	0	1	1	1	1	1
$\mathcal{T}$	0.93	0.05	-0.10	3.27	0.56	0.22	2.75×10^6	0.67	0.43
$\mathcal{T}_{log}$	-0.05	0.98	-0.03	20.4	0.91	1.00	1.33	0.75	0.93
$\mathcal{T}_{log\text{-1D}}$	0.84	0.98	-0.02	4.96	0.90	0.96	1.38	0.79	0.94

We first evaluate the POD projection error using the Q2 metric (Eq. IV.36) (see Sect. IV.3.2). It is a relevant sanity check, as the POD is precisely designed to maximize the reconstructed variance. Table IV.2 shows the Q2 scores averaged over the test set and using $L = 10$ POD modes, which is the number of modes chosen in Sect. IV.5.1. The POD works as expected, with Q2 scores close to the optimal value for the Q2 on concentration with $\mathcal{T}$ and for the Q2 on log-concentration with $\mathcal{T}_{log}$ and $\mathcal{T}_{log\text{-1D}}$. Indeed, with logarithmic rescaling the POD optimizes the variance of the log-concentration, which implies that it does not necessarily reproduce well the variance of the concentration as

shown with the rather poor $Q2(\overline{\mathbf{c}})$ score obtained with $\mathcal{T}_{log}$. This is explained because the area of the domain that corresponds to high variance is not the same for concentration and log concentration. In particular, a large part of the total concentration variance is located near the source, whereas it is more on the plume edges for the log concentration. Interestingly, using the $\mathcal{T}_{log\text{-}1\mathrm{D}}$ preprocessing gives good results for both $Q2(\overline{\mathbf{c}})$ and $Q2(\ln(\overline{\mathbf{c}}))$.

In addition to the Q2 metric, the POD projection error is also evaluated using the air quality metrics defined in Sect. III.3.2, page 90. This second round of validation is essential because a good Q2 score does not necessarily imply good scores for all air quality metrics, as shown in Table IV.2. In particular, despite its good $Q2(\overline{\mathbf{c}})$ score, the standard preprocessing $\mathcal{T}$ gives poor air quality metrics scores, except NMSE. This is because, in this case, the POD puts too much weight on areas of high concentration and represents poorly the areas where the variance is lower in absolute value. Conversely, with log-transformation, the weights are distributed more evenly, which improves the overall POD accuracy. Indeed, $\mathcal{T}_{log}$ and $\mathcal{T}_{log\text{-}1\mathrm{D}}$ obtain air quality metrics scores close to the reference internal variability error, defined in Sect. IV.3.4. However, the area near the source is less well represented with this preprocessing as shown by the NMSE deterioration. This deterioration is less important with $\mathcal{T}_{log\text{-}1\mathrm{D}}$, which is thus the preprocessing that leads to the best compromise in air quality metrics performances.

A qualitative checking of the reconstructed field obtained by POD compression-decompression is also performed. Figure IV.11 shows horizontal cut at $z = 1.6\,\mathrm{m}$ of the mean concentration for the three preprocessing considered and for the original LES sample #91. This sample corresponds to the wind conditions $(\alpha_{inlet}^{(91)}, u_*^{(91)}) = (-8.9\,^\circ, 0.38\,\mathrm{m\,s^{-1}})$ and is represented as a yellow star in Fig. IV.5. This test sample is chosen because the errors of the complete POD–GPRs model evaluated with the different air quality metrics for this sample are close to the average errors over the entire test base (see Fig. IV.19, page 163). The horizontal cut (Fig IV.11b) illustrates why POD with standard processing $\mathcal{T}$ achieves very poor air quality metrics scores (Table IV.2). In this example, the reconstructed concentration field is far from the reference as high concentrations cover a larger area, artificial structures appear and the edges of the plume are sharper. On the other hand, the fields obtained with $\mathcal{T}_{log}$ and $\mathcal{T}_{log\text{-}1\mathrm{D}}$ seem to fit much better, but the 5-ppm isoline is more elongated than in the reference field, and the zone close to the source is not perfectly reconstructed (Fig IV.11c, d). This is consistent with the averaged NMSE scores obtained (Table IV.2).

Figure IV.12 illustrates the spatial structures carried by different modes $\boldsymbol{\psi}_\ell$ of the POD computed using $\mathcal{T}_{log}$ preprocessing. The modes are normalized by the square root of the ratio between the variance Λ_ℓ explained by the ℓth mode, and the total variance $\mathbb{V}(\overline{\mathbf{c}}_{\mathrm{LES}}^{train})$ of the field at each node of the domain and computed over the train set. In this way, the maps in Fig. IV.12 represent spatially the correlation between the original train set and the POD modes (Nony et al. 2023a). Overall, the POD modes have a fan-shaped structure with cone shapes of opposite correlation signs that start from the source and sweep across the domain (IV.12a, c, and e). When the number of modes increases the number of cones increases, while the correlations decrease which is because the first modes are the ones that explain most of the original sample variance. This organization

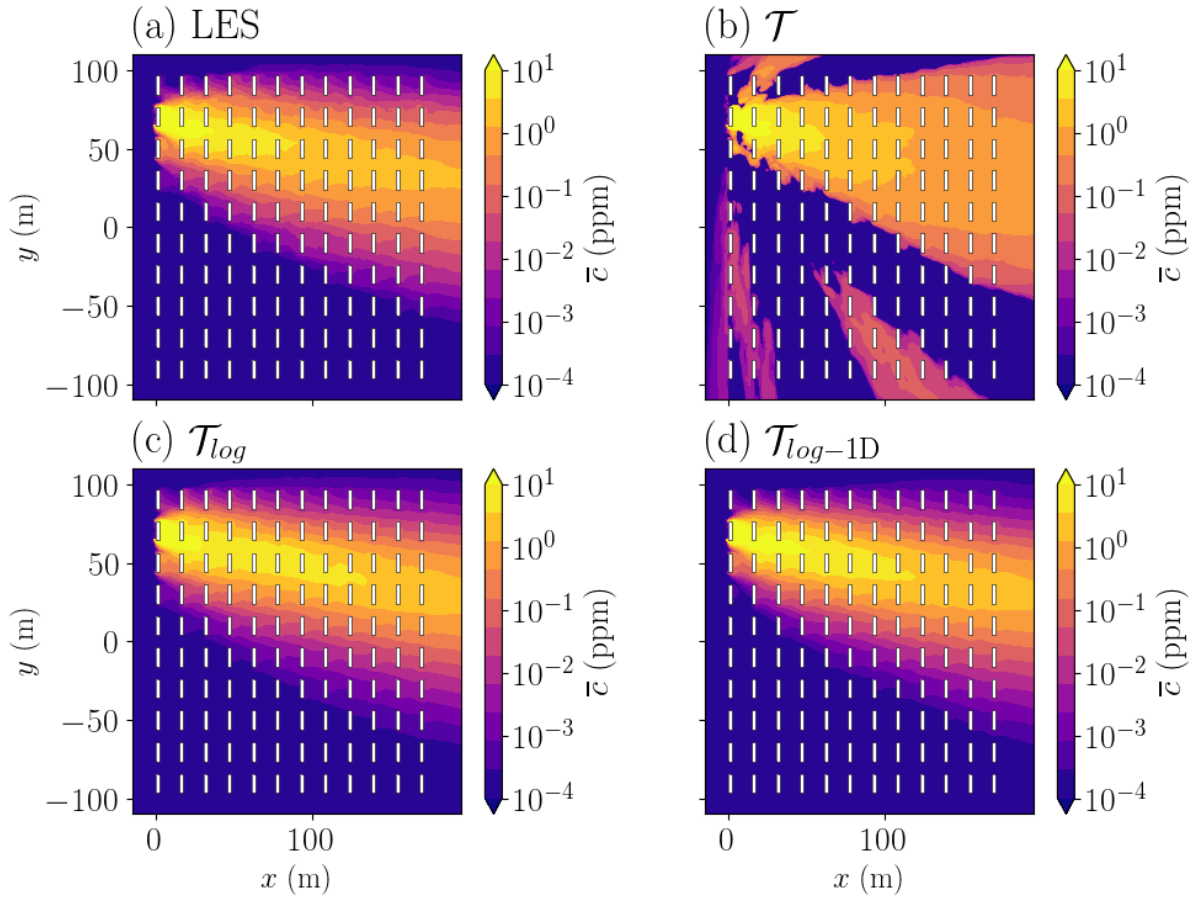


Figure IV.11: *Horizontal cuts at $z = 1.6\,m$ of the mean concentration field projected to the latent space spanned by the $L = 10$ first POD modes and then reconstructed using the POD inverse transform (Eq. IV.8). Results are given for the LES test sample #91 with $(\alpha_{inlet}^{(91)}, u_*^{(91)}) = (-8.9\,°, 0.38\,m\,s^{-1})$ (a). Three different field preprocessing are compared: the standard centering (Eq. IV.21) (b), the log-transformation (Eq. IV.22) (c), and the log-transformation with friction velocity normalization (Eq. IV.24) (d).*

echoes the plumes obtained for the different wind direction α_{inlet} of the train set. This suggests an important contribution of α_{inlet} to the variance of the train set. Near the source (IV.12b, d, and f), no clear pattern emerges which is probably linked to the fact less weight is given to this area when building the POD on the log-concentration fields.

It is important to note that the overall accuracy of the POD–GPRs model is limited by the POD projection error, as GPRs are built in the latent space defined by the POD reduced base (Fig. IV.1a), and predictions are reconstructed by POD inverse projection (Fig. IV.1b). Therefore, we decide not to use the standard preprocessing $\mathcal{T}$ for the rest of the study as the standard preprocessing has poor overall accuracy and yields distorted unphysical concentration fields (Fig IV.11b). An exception is made in Appendix. B.3, in which we propose an approach combining POD–GPRs with both linear and log-transformation of the fields to get the best of both approaches.

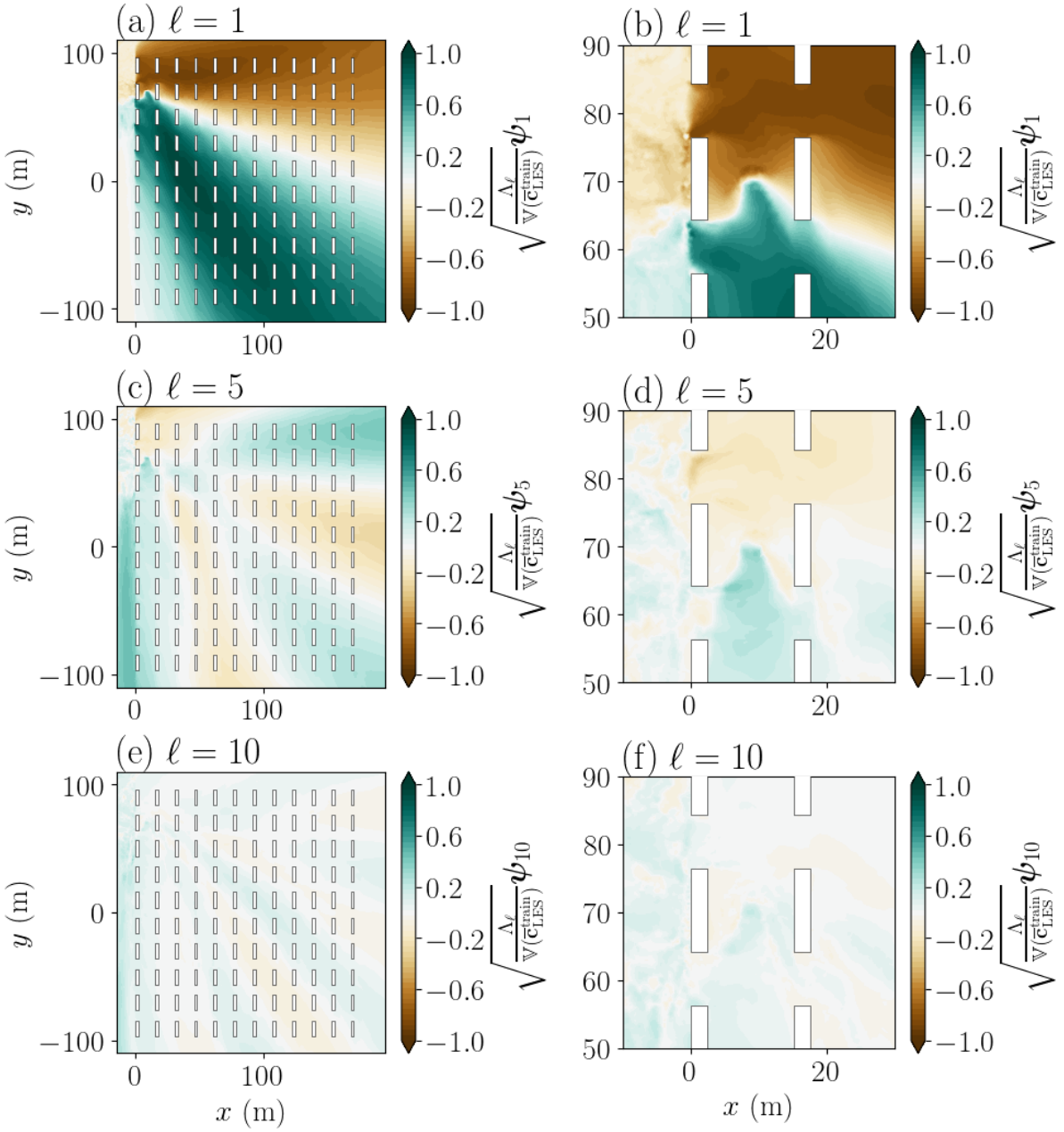


Figure IV.12: *Horizontal cuts at $z = 1.6\,m$ of the normalized POD modes obtained with $\mathcal{T}_{log}$ preprocessing. The first row corresponds to a global view of normalized modes within the canopy, while the second row shows a zoom at the source level. Each line corresponds to a specific mode, $\ell = 1$, 5, and 10 respectively.*

IV.5.3 Ability of Gaussian process regressors to represent internal variability

In this section, we optimize the GPRs and validate their ability to predict the POD-reduced coefficients and to account for the effect of internal variability. Results are given using a Matérn kernel function with $\nu = 5/2$ (Eq. IV.18) and the $\mathcal{T}_{log}$ fields preprocessing (Eq. IV.22). Similar results were found with $\mathcal{T}_{log\text{-}1D}$.

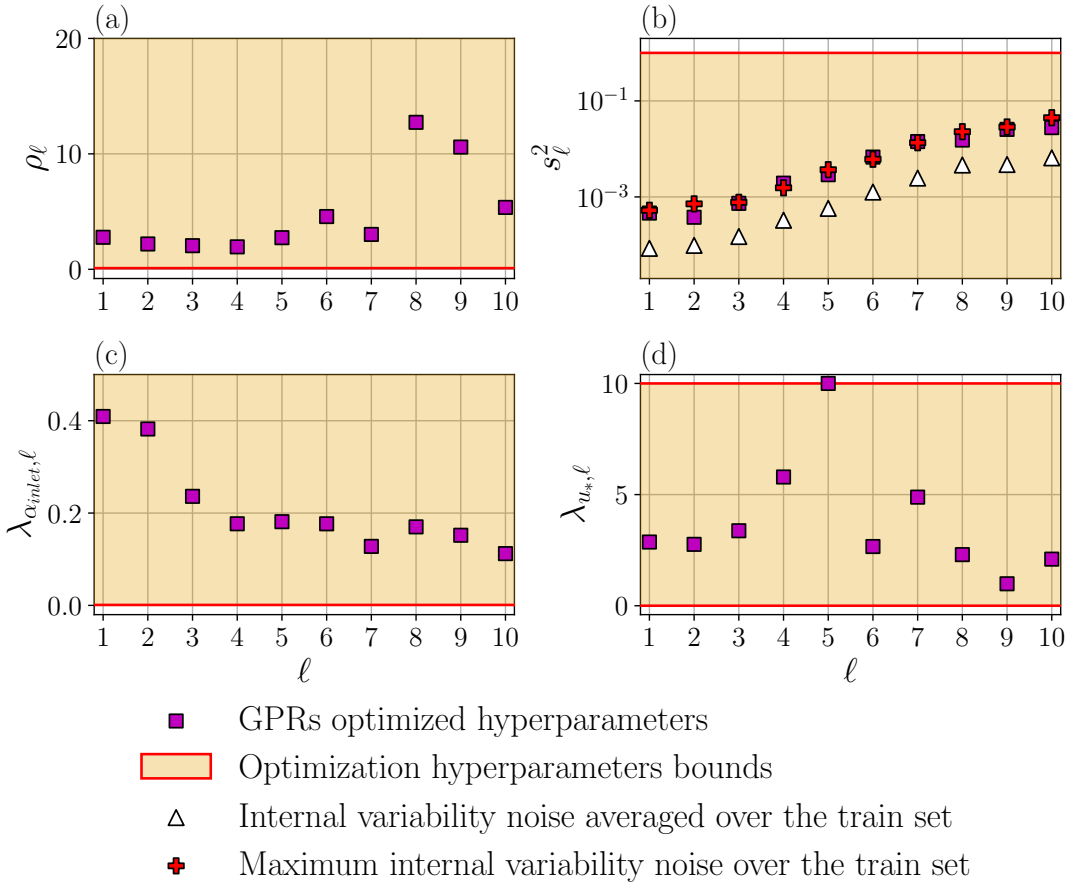


Figure IV.13: *Gaussian Process Regressors hyperparameters obtained by maximum log-likelihood estimation (see Sect. IV.2.3) for each mode. These hyperparameters are the maximum allowable covariance ρ_ℓ (a), the noise variance s^2_ℓ (b), and the length scales $\lambda_{\alpha_{inlet},\ell}$ and $\lambda_{u_*,\ell}$ (c, d). Orange shaded areas, with red edges, correspond to the hyperparameters bounds prescribed for the optimization. White triangles and red plus symbols correspond respectively to the average and maximum noise on the reduced coefficients induced by internal variability over the train set and obtained using the bootstrap procedure depicted in Fig. IV.8.*

GPRs optimization results Figure IV.13 shows the hyperparameters found by Maximum Log-Likelihood (MLL) estimation. As we arbitrarily choose hyperparameter bounds for the optimization, we check that the optimization solutions are not constrained by

our choice. In this case, results are satisfying as only one optimized hyperparameter, $\lambda^*_{u_*,5}$, hits the bound (represented by the red edges in Fig. IV.13). The optimized maximum allowable covariance ρ^*_ℓ is around 3 for the first modes and 10 for the last modes (Fig. IV.13a). This order of magnitude is coherent since the reduced coefficients k_ℓ have unit variance (Eq. IV.10b). The increase of ρ^*_ℓ with ℓ means that GPRs should be less constrained for higher modes. This indicates that differences between neighboring $k_\ell(\boldsymbol{\theta})$ in the parameter space tend to increase with ℓ, which is probably related to the increase in noise (Fig. IV.9). We also note that the optimized length scales are shorter for α_{inlet} than for u_*. The GPRs predictions are therefore poorly correlated with train samples that have distant wind directions, which is reassuring since the shape of the plume and consequently the POD projection changes strongly with α_{inlet}.

Validation of the estimated noise To validate the noise variance s^2_ℓ inferred by MLL estimation, we compare it to the noise induced by the internal variability of the mean concentration and propagated to the latent space. To do so, we compute $s^2(\mu(k_\ell))$ the variance of the bootstrap replicates of the POD reduced coefficients k_ℓ generated using the procedure detailed in Sect. IV.5.1. We find that MLL estimation tends to overestimate the noise compared to internal variability averaged over the train set, but closely fits the maximum level of internal variability among the train set (Fig. IV.13b). The noise estimated by MLL is therefore aligned with the worst case, which here corresponds to a low friction velocity (more details on the distribution of the internal variability are presented in Sect. IV.6.4). Using the worst-case scenario for the complete inputs parameter space is fine as it is better not to underestimate the noise. Indeed, noise underestimation would result in the train set overfitting and unphysical POD–GPRs surface. Overall, we are very satisfied with these results that demonstrate that GPRs optimization can infer correctly the LES theoretical noise based on its sampled response surface (Fig. IV.7) and that that POD–GPRs could be used to account for the fields uncertainty related to internal variability.

Instead of inferring the noise variance hyperparameter s^2_ℓ, we have also considered directly prescribing it as equal to the estimated variance of the POD reduced coefficients averaged over the train set $\langle s^2(\mu(k_\ell))\rangle$. Results show slightly less accurate GPR predictions over the test set when prescribing the noise variance, probably because it removes a degree of freedom for optimization. In addition, estimating the noise prior requires a costly bootstrap procedure (Fig. IV.8), we therefore recommend inferring the noise variance hyperparameter s^2_ℓ when optimizing GPRs instead of prescribing it.

Validation of the POD reduced coefficients estimates Figure IV.14 shows the error of GPRs in predicting the POD reduced coefficients. The error is quantified by the Mean Squared Error MSE calculated over the train set (Fig. IV.14a), and test set (Fig. IV.14b). The error is less than 0.05 for every mode which typically represents about 5% of the POD reduced coefficient values. Moreover, the prediction error is much smaller for the first modes, which are the most important to predict because they contain the most information about the system (see Sect. IV.2.2). The fact that GPRs noise variance aligns with maximum, rather than average, internal variability (Fig. IV.13b)

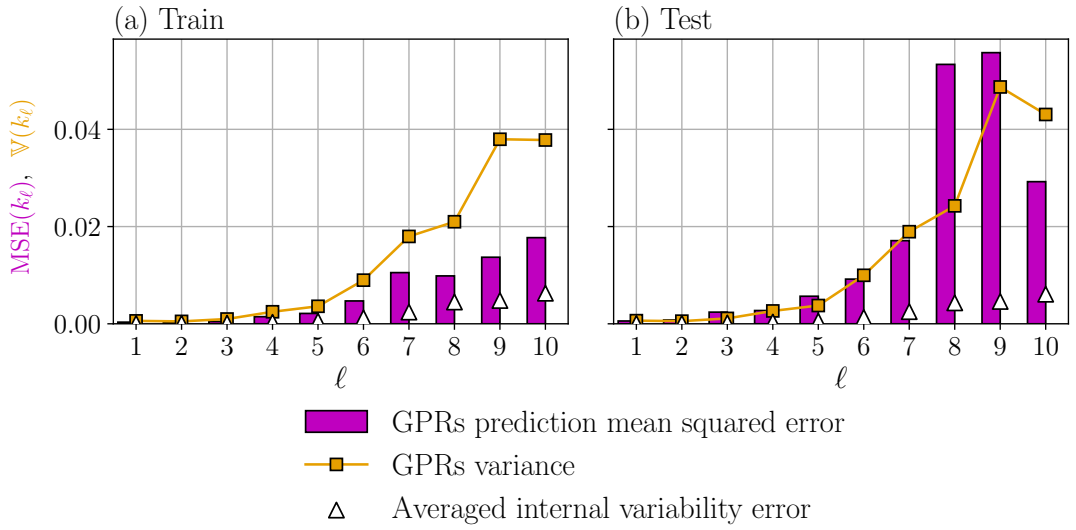


Figure IV.14: *Mean Squared Error* (MSE) *of the POD reduced coefficients estimated by the GPRs over the train set (a), or the test set (b). Results are given for each mode as vertical bars. The variance of the GPRs (Eq. IV.16b) is also represented by the orange line. White triangles correspond to the noise induced by internal variability averaged over the train and test sets, and obtained using the bootstrap procedure depicted in Fig. IV.8.*

explains why the MSE over the train set exceeds the averaged internal variability error (Fig. IV.14a). Finally, in addition to having very good overall accuracy, the variance of the GPRs (Eq. IV.16b) is in agreement with the error they commit, including on the test set (Fig. IV.14). We can therefore be confident in this estimate of the uncertainty of the GPRs predictions, which covers both the regression error and the fact that the data to be predicted are noisy.

IV.6 Validation of the POD–GPRs model

In this section, we provide a thorough validation of the POD–GPRs reduced-order model. Based on the findings of the previous section, the log-transformation preprocessing is used to build the POD–GPRs and we compare the performances obtained with or without normalizing by friction speed. We first assess in Sect. IV.6.1 the overall POD–GPRs accuracy over the test set, the sanity of its parametric response surface, as well as its efficiency. For each criterion, the 1–NN reduced-order model is used as a benchmark to evaluate the added value of the POD–GPRs. This comparison between the different versions of POD-GPRs and the 1-NN is deepened by investigating their robustness to the reduction in the size of the learning base (Sect. IV.6.2). We also perform cross-validation to assess the robustness of the POD–GPRs to the composition of the train set (Sect. IV.6.3), and to describe how POD–GPRs performance varies in parameter space (Sect. IV.6.4). Finally, we validate the relevance of our method for selecting the number of modes in Sect. IV.6.5.

IV.6.1 Analysis of the model reduction error

In this section, we validate the POD–GPRs predictions of mean concentration following the methodology introduced in Sect. IV.3. In particular, we use both the Nearest Neighbor (1–NN) and the mean internal variability error as references for the validation (see Sect. IV.3.1 and IV.3.4). We also compare the results obtained with the two fields preprocessing with log-scaling ($\mathcal{T}_{log}$ and $\mathcal{T}_{log\text{-}1\mathrm{D}}$) as they minimize the overall POD projection error (Sect. IV.5.2). In both cases, we use $L = 10$ modes in agreement with the informed choice made in Sect. IV.5.1.

Averaged prediction error The overall performance of the reduced-order models is quantified using the air quality metrics presented in Sect. III.3.2, page 90. Table IV.3 gives the scores averaged over the test set. The POD–GPR models yield very satisfactory overall results, with scores close to the error only due to internal variability (IV), which represents the best achievable accuracy. However, this is not true for the NMSE and FMS(1 ppm), indicating that POD–GPRs are less accurate at predicting high concentrations. We note that normalizing the fields by the friction velocity, i.e. using $\mathcal{T}_{log\text{-}1\mathrm{D}}$, improves the performances of the POD–GPRs over the high concentrations. In addition, Table IV.3 demonstrates that there is added value in using the POD–GPRs model, as it is more accurate than the trivial 1–NN model. We show in Sect. IV.6.2 that this gap between POD–GPRs and 1–NN increases when the size of the train set decreases. Finally, we note that the POD–GPRs prediction errors are almost identical to the POD projection errors (Table IV.2). This leads to the conclusion that the accuracy of the POD–GPR model is constrained by the accuracy of the POD.

Note that averaging the metric scores over the test set presents several limitations: in particular FB and MG are very low because of compensation errors, while the other metrics averaged scores are worsened by the presence of outliers. A detailed sample-by-sample analysis of the accuracy of the POD–GPRs model is provided in Sect. IV.6.4.

Table IV.3: *Comparison of the validation scores averaged over the test set for the POD–GPRs and 1–NN reduced-order models. Results for the POD–GPRs are given for the two fields preprocessing with log-scaling ($\mathcal{T}_{log}$ and $\mathcal{T}_{log\text{-}1D}$). Definitions of the validation metrics are given in Sect. III.3.2, page 90. Perfect scores of the different metrics are recalled in the second row. The third row corresponds to the mean level of error solely due to internal variability (see Sect. IV.3.4).*

	FB	NMSE	FAC2	MG	VG	FMS (1ppm)	FMS (0.01ppm)
Perfect score	0	0	1	1	1	1	1
Internal variability	0	1.80	0.95	1.00	1.39	0.83	0.93
1–NN	-0.03	7.75	0.87	0.99	2.22	0.74	0.87
POD–GPRs – $\mathcal{T}_{log}$	-0.04	20.6	0.91	1.00	1.39	0.75	0.92
POD–GPRs – $\mathcal{T}_{log\text{-}1D}$	-0.02	4.61	0.90	0.97	1.40	0.79	0.93

Snapshots comparison For a more detailed assessment of the accuracy of reduced-order models, we qualitatively examine their predictions. For instance, horizontal cuts of the mean concentration at $z = 1.6\,\mathrm{m}$ are given in Fig IV.15, for one LES test sample and the associated reduced-order model predictions. We choose the test sample #91 $(\alpha_{inlet}^{(91)}, u_*^{(91)}) = (-8.9\,^\circ, 0.38\,\mathrm{m\,s^{-1}})$, depicted as a yellow star in Fig. IV.5, as a reference sample for diagnostics because the validation scores obtained by the POD–GPRs model for this sample are representative of its overall accuracy, as shown in Fig. IV.19. A more detailed analysis of the model reduction error dispersion is given in Sect.IV.6.4. Each reduced-order model prediction is in overall good agreement with the test LES concentration field. 1–NN deviates from the plume towards negative angles as it uses a neighbor train field with $\alpha_{inlet} < \alpha_{inlet}^{(91)}$. Meanwhile, both versions of the POD–GPRs tend to overestimate the extension of the higher concentration isolines.

These observations are corroborated by the comparison between the LES and reduced-order model predictions of the 0.01 ppm and 5 ppm isolines (Fig. IV.16d, e, and f). In addition, every reduced-order model tends to overestimate the mean concentration between the containers, with particularly important errors at the source location (Fig. IV.16a, b, and c). This supports our first analysis, which associates high overall NMSE levels in Table IV.3 with errors in high concentrations. This is because, with the log-transformation applied to the concentration fields, the POD does not reproduce well the high concentrations, as already shown in Sect. IV.5.2. Moreover, the near-source area is naturally uncertain, as it is highly sensitive to internal variability due to its position near the edge of the canopy and strong concentration gradients. Finally, we note that when building the POD–GPRs on the fields normalized by the friction velocity with $\mathcal{T}_{log\text{-}1D}$, it overestimates less the extension of the 5 ppm isoline than his counterpart with $\mathcal{T}_{log}$ (Fig. IV.16b, and c), which is consistent with the overall trends presented in Table IV.3.

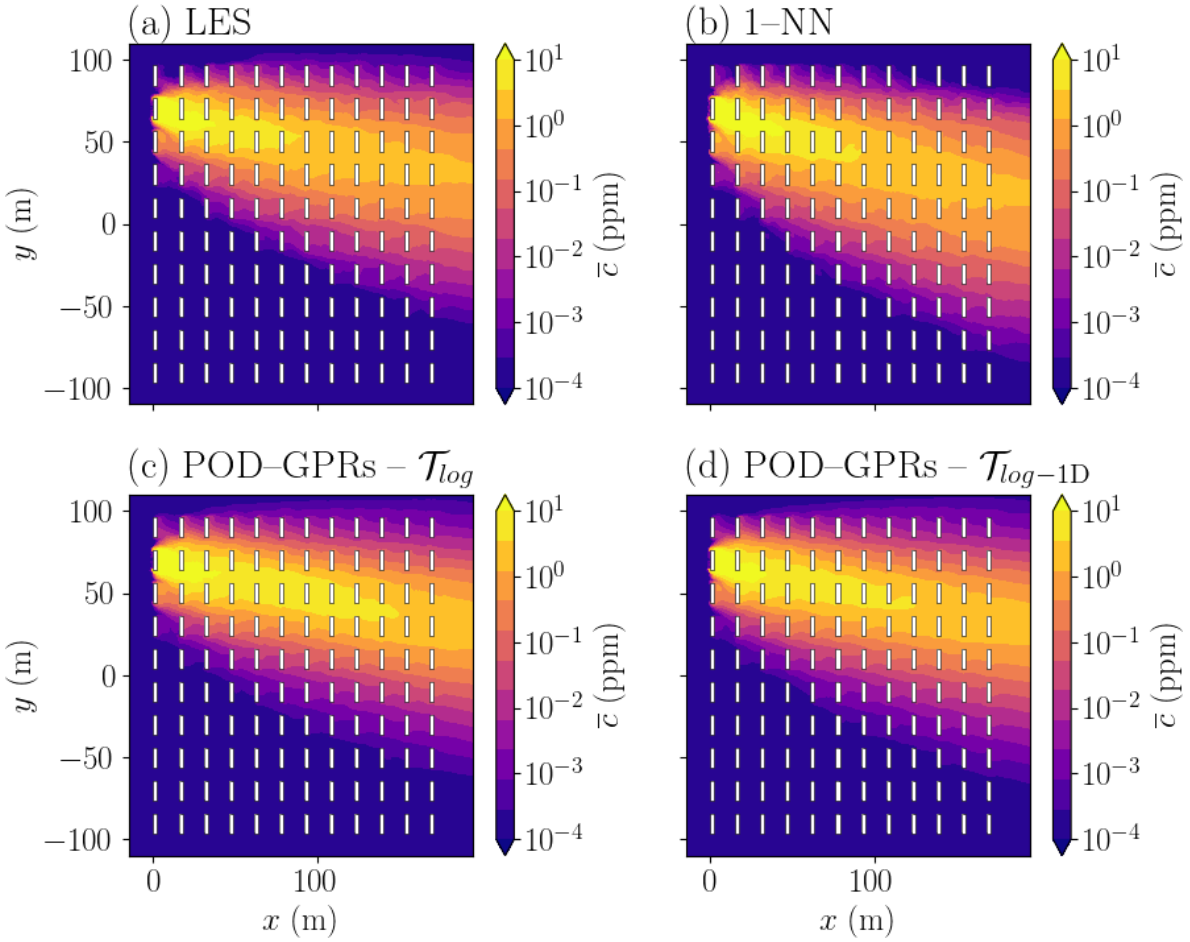


Figure IV.15: *Horizontal cuts at $z = 1.6\,m$ of the mean concentration estimated by the reference LES model (a), the 1–NN (b), and the POD–GPRs with the $\mathcal{T}_{log}$ (c) and $\mathcal{T}_{log1D}$ (d) pre-processing. Results are given for the test sample #91 with $(\alpha_{inlet}^{(91)}, u_*^{(91)}) = (-8.9\,°, 0.38\,m\,s^{-1})$.*

Analysis of the parametric response surfaces Another key aspect of reduced-order model validation is to verify the sanity of the response surface $\mathbf{y} = \mathcal{M}_{\text{ROM}}(\boldsymbol{\theta})$. It is particularly important in this thesis since we will use a reduced-order model in Chapter V to infer $\boldsymbol{\theta}$ given observations of $\mathbf{y}$. Figure IV.17a shows the mean concentration predicted by the POD–GPRs and 1–NN at the tower B at $z = 2\,m$ at constant friction velocity $u_*^{plot} = 0.45\,\text{m}\,\text{s}^{-1}$ and varying inlet wind direction. On the other hand, Figure IV.17b gives the response surface for $\alpha_{inlet}^{plot} = -43\,°$and varying u_*. In both cases, 1–NN and POD–GPRs show fine agreement with the LES samples closest to the two segments of parameter spaces thus scanned (see Fig. IV.5). As expected, the response surfaces of 1–NN are staircase-shaped and much less smooth than those of POD–GPRs which better reproduce the physics of the LES model as evidenced by the comparison with the LES response surface (Fig. IV.7).

The POD–GPRs built on the fields normalized by the friction velocity with $\mathcal{T}_{log\text{-}1D}$ gives more regular profiles than his counterpart with $\mathcal{T}_{log}$ (Fig IV.17). For the u_* profile,

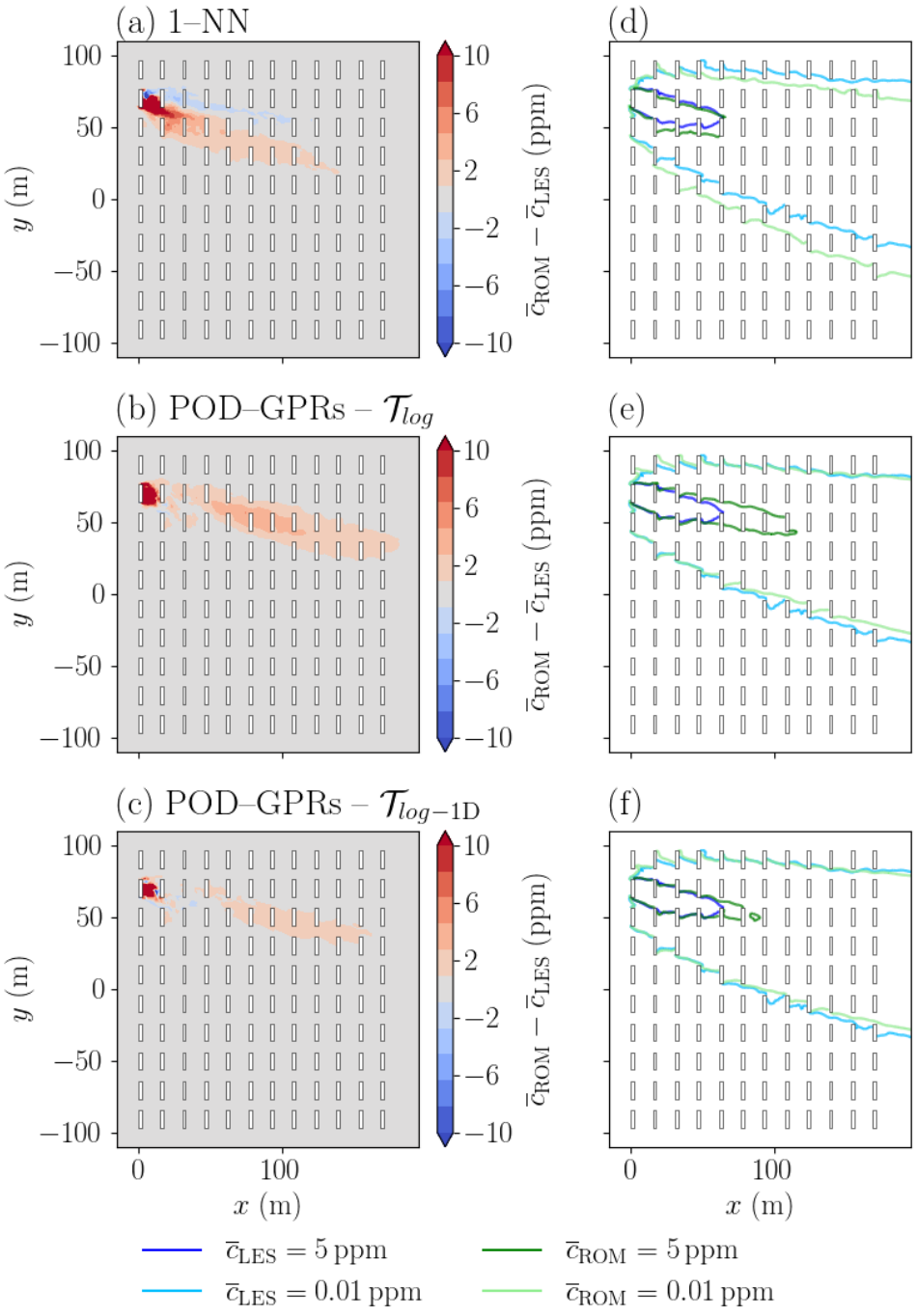


Figure IV.16: *Horizontal cuts at $z = 1.6\,m$ of the absolute difference between LES and reduced-order models mean concentration predictions, for the 1–NN (a), and the POD–GPRs with the $\mathcal{T}_{log}$ (b) and $\mathcal{T}_{log1D}$ (c) preprocessing. The predictions are also compared for two specific iso concentration levels: 0.01 ppm and 5 ppm (d, e, f), with the isolines predicted by LES and by the reduced-order model in blue and green respectively. Results are given for the test sample #91 with $(\alpha_{inlet}^{(91)}, u_*^{(91)}) = (-8.9\,°, 0.38\,m\,s^{-1})$.*

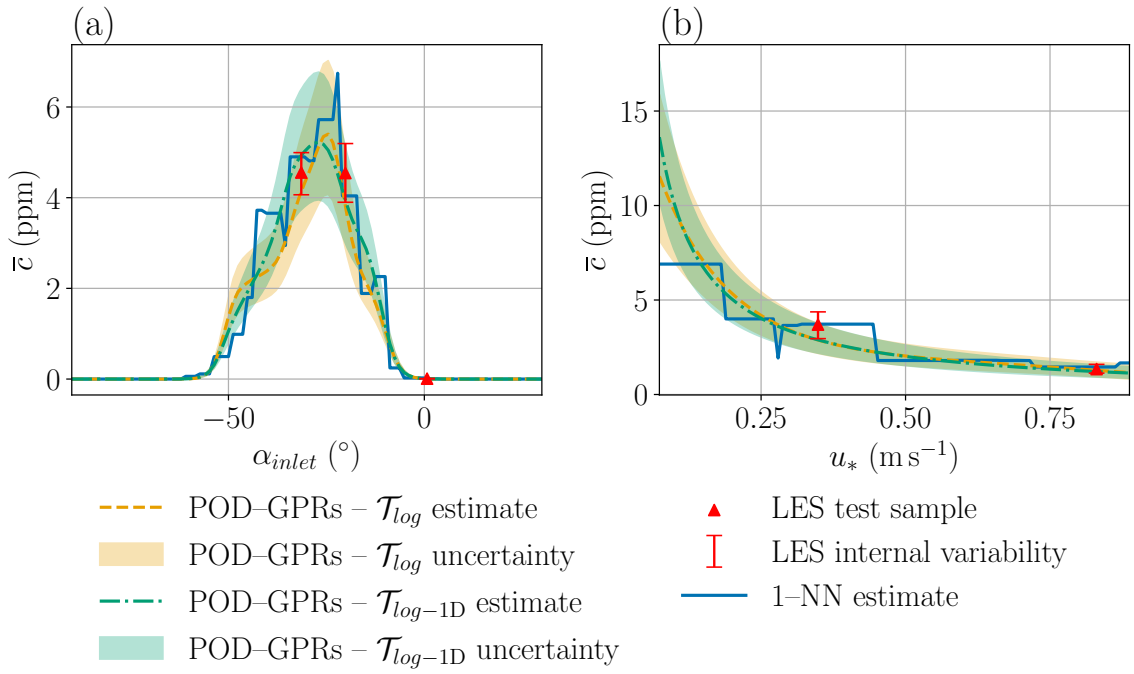


Figure IV.17: *Reduced-order model estimates of the mean concentration at tower B at $z = 2\,m$ as a function of the inlet wind direction α_{inlet} (a), and of the friction velocity u_* (b). When varying one parameter, the other is set constant to either $u_*^{plot} = 0.45\,m\,s^{-1}$ (a), or $\alpha_{inlet}^{plot} = -43\,°$ (b). The test samples closest to the two segments of parameter space thus scanned are represented by red triangles and are also identified in Fig. IV.5. Results are given for three different reduced-order models: POD–GPRs with $\mathcal{T}_{log}$, POD–GPRs with $\mathcal{T}_{log\text{-}1D}$, and 1–NN, as orange dashed lines, green dashed-dotted line and blue solid line respectively. Orange and green shaded areas correspond to the 95% confidence intervals estimated by the POD–GPRs according to the procedure detailed in Sect. IV.2.5. The uncertainty of the LES test samples due to internal variability is also depicted as red error bars.*

this is simply because its dependency on the friction velocity is imposed by the fields rescaling (Eq. IV.25). For the α_{inlet} profile, this is because, by collapsing the train set into one dimension, the number of neighboring train samples for one wind direction α_{inlet} drastically increases (Fig. IV.5b), thus smoothing the interpolation. It is interesting to note that the POD–GPRs model with $\mathcal{T}_{log}$ is also able to retrieve the inversely proportional dependence of concentration on friction speed (Fig. IV.17b).

One main feature of the proposed POD–GPRs modeling is that we have access to estimate the prediction uncertainty by propagating the GPRs variance in the physical space as developed in Sect. IV.2.5. Figure IV.17 also depicts the estimated 95% confidence intervals along the mean estimate profile. These intervals are based on the 2.5th and 97.5th of the log-normal distribution which is the distribution of the POD–GPRs estimates when using $\mathcal{T}_{log\text{-}1D}$ and $\mathcal{T}_{log\text{-}1D}$ by construction, as demonstrated in Sect. IV.2.5. Results show that these intervals effectively cover the departure between the LES model and the POD–GPRs. Moreover, the total POD–GPRs uncertainty is coherent with the

field uncertainty induced by internal variability but always larger since it also accounts for the regression error.

Model efficiency We evaluate the efficiency of the POD–GPRs model as we aim to use it in contexts requiring real-time prediction or multi-query evaluation for ensemble data assimilation. Table IV.4 gives the computational costs of the training, which can be performed offline, and of the prediction. The efficiency objective appears to be convincingly met by the POD–GPRs model with a prediction time of less than a tenth of a second and a total training time of about 30 s. Note that the POD–GPRs training time, which includes the fields preprocessing, the POD basis computation, and the GPRs optimization (Fig. IV.1a), is insignificant compared to the construction of the learning base, that cost 5.7 million core hours.

Table IV.4: *Computational costs of the POD–GPRs reduced-order models using two different preprocessing $\mathcal{T}_{log}$ and $\mathcal{T}_{log\text{-}1D}$. Computational costs are expressed in seconds separately for the complete reduced-order model training and a single prediction. Results are averaged over 10 training realizations and 1000 predictions. All the computations were performed using a single core of an Intel Ice Lake processor.*

Computational cost	Training	Prediction
POD–GPRs – $\mathcal{T}_{log}$	30 s	3.0×10^{-2} s
POD–GPRs – $\mathcal{T}_{log\text{-}1D}$	29 s	3.0×10^{-2} s

IV.6.2 Behavior of the reduced-order models for restricted train set

In this section, we assess how prediction errors of the different reduced-order models evolve as the size of the learning database decreases. This makes it possible to define a compromise between accuracy and database size, which can be used as a guideline for applying our approach to new case studies. It is especially important since the cost of building the LES training database is very high (Table IV.1). In addition, by decreasing the size of the train set without changing the test set, we gain insights into the ability of a model to generalize from limited training data.

To evaluate the robustness of each reduced-order model to the size of the train set, we train the model for decreasing train set size $N_{train} \in \{160, 128, 96, 64, 48, 32\}$ by randomly removing train samples. For the comparison to be fair, we always evaluate the averaged prediction errors over the same test set of $N_{test} = 40$ samples. Note that we do not use FB nor MG for this diagnostic as their averaged value over the test is too prone to error compensation (see Fig. IV.19). Figure IV.18 shows that, for FAC2, FMS(1 ppm), and FMS(0.01 ppm) the decrease in accuracy is fairly constant for each reduced-order model considered. By contrast, VG scores explode with the 1–NN but not with POD–GPRs. Concerning the NMSE, results are contrasted, while the evolution of the scores is quite

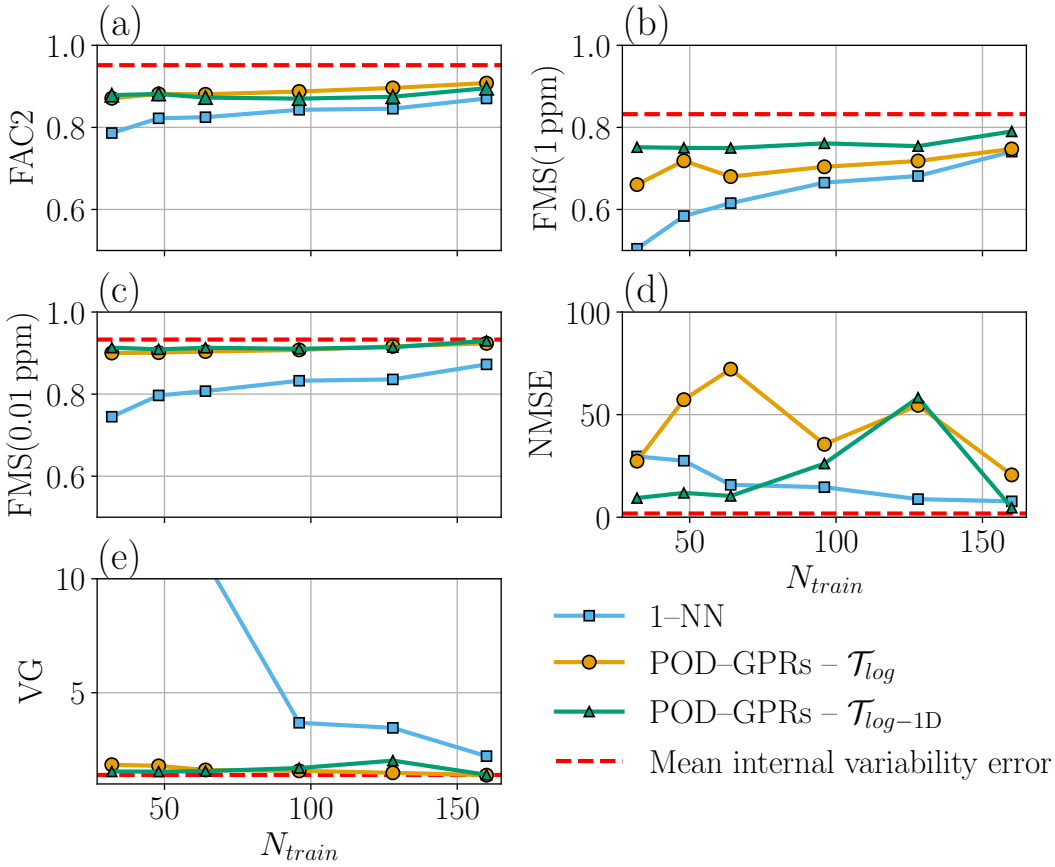


Figure IV.18: *Reduced-order model mean concentration prediction error for decreasing sizes of the train set. The prediction error is estimated using some of the standard air quality metrics presented in Sect. III.3.2, page 90: namely FAC2 (a), FMS(1 ppm) (b), FMS(0.01 ppm) (c), NMSE (d), and VG (e). Results are given for the 1–NN as blue squares and for the POD–GPRs with both $\mathcal{T}_{log}$ and $\mathcal{T}_{log-1D}$ preprocessing as orange circles and green triangles respectively.*

regular for the 1–NN it is not monotonous at all for the POD–GPRs. NMSE scores quickly deteriorate when decreasing the size of the train size, but eventually, improve for very small train sets. This behavior is probably linked to the high POD projection error near the source when using log-transformation (see Table IV.2 and Fig. IV.16). In general, the best performances are obtained by the POD–GPRs model using $\mathcal{T}_{log-1D}$. It is in particular the more accurate for the minimum train set size ($N_{train} = 32$). This is explained because normalizing the fields by the friction velocity increases the density of the train set (Fig. IV.5b).

Overall, the POD–GPRs models, and especially with $\mathcal{T}_{log-1D}$, cope better with the reduction in the size of the train set than the 1–NN model. This motivates the use of a more sophisticated model reduction method when the size of the learning database is limited.

IV.6.3 Sensitivity to the choice of the training data

We apply in this section the K-fold cross-validation methodology presented in Sect. IV.3.3 to test the sensitivity of the POD–GPRs model to the composition of the train set. Results are given for the POD–GPRs with the $\mathcal{T}_{log-1\mathrm{D}}$ preprocessing.

Table IV.5: *Comparison of the POD–GPRs mean concentration prediction error averaged over five test sets following the 5-fold cross-validation procedure presented in Sect. IV.3.3. Minimum and maximum errors over the different test sets are also given. The prediction error is evaluated by the standard air quality metrics presented in Sect. III.3.2, page 90. Results are given for the POD–GPRs built using the $\mathcal{T}_{log-1\mathrm{D}}$ preprocessing. Perfect scores of the different metrics are recalled in the second row. The third row corresponds to the mean level of error solely due to internal variability (see Sect. IV.3.4).*

	FB	NMSE	FAC2	MG	VG	FMS (1ppm)	FMS (0.01ppm)
Perfect score	0	0	1	1	1	1	1
Internal variability	0	1.80	0.95	1.00	1.39	0.83	0.93
Average score	-0.02	4.06	0.90	0.97	1.36	0.79	0.93
Minimum score	-0.03	3.65	0.90	0.96	1.30	0.78	0.93
Maximum score	-0.02	4.61	0.91	0.98	1.34	0.80	0.93

We use $K = 5$ folds for the cross-validation in order to keep the same test-train ratio (25%) as the one previously used in Sect. IV.6. Table IV.5 gives the POD–GPRs mean concentration prediction errors averaged over the five different test sets. Recall that we start over the training for each fold so that each POD–GPRs validation exercise is independent. As shown by minimum and maximum errors over each fold, apart from the quadratic metrics NMSE and VG, the scores are not affected by changes in the train set. The NMSE and VG metrics vary more since they measure data dispersion and are therefore more sensitive to the presence of outliers in the test blocks.

In conclusion, it is not necessary to proceed to a more in-depth analysis of the variance of the POD–GPRs validation scores since the metrics are not significantly sensitive to the choice of the training data.

IV.6.4 Dispersion of errors over test samples

We take advantage of this cross-validation methodology to examine the distribution of the POD–GPRs errors over the test samples. Indeed, each sample in the Halton sequence (Fig. IV.5) once belongs to one of the five test blocks. Interestingly, we find that the POD–GPRs errors are not uniformly distributed in the input parameter space. Figure IV.19 gives the air quality metrics distribution as a function of the friction velocity parameter. Overall, we note a quite large dispersion of the metrics scores, especially for the quadratic metrics NMSE and VG, with the presence of outliers indicating test samples for which

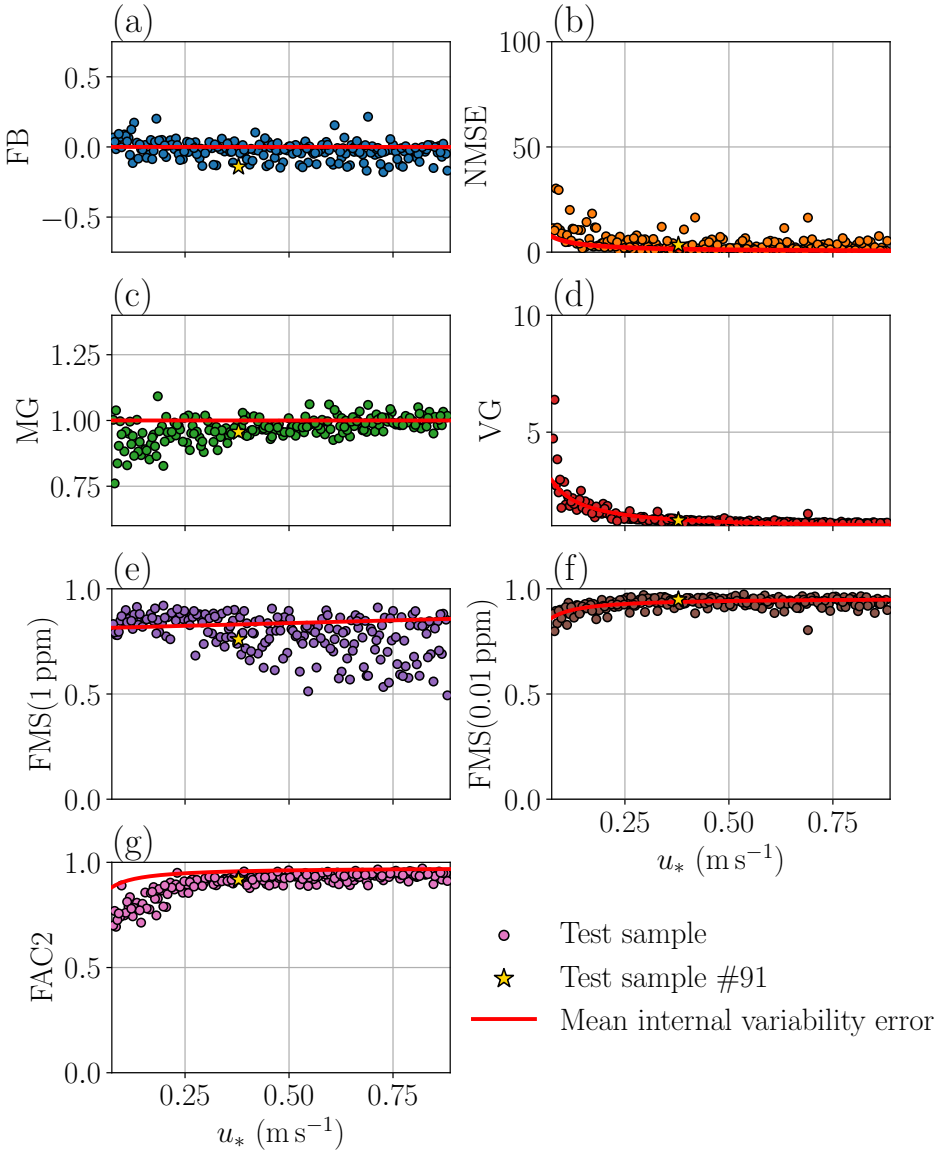


Figure IV.19: *POD–GPRs mean concentration prediction error for each sample evaluated using a 5-fold cross-validation procedure (Sect. IV.3.3) and sorted by increasing friction velocity. The prediction error is evaluated by the standard air quality metrics presented in Sect. III.3.2, page 90: namely FB (a), NMSE (b), MG (c), VG (d), FMS(1 ppm) (e), FMS(0.01 ppm) (f), and FAC2 (g). Results are given for the POD–GPRs built using the $\mathcal{T}_{log-1D}$ preprocessing. The test sample #91 used as an example for diagnostics in Sect. IV.6 is represented as a yellow star. Red lines correspond to an internal variability error model fitted for each metric.*

the POD–GPRs are less accurate. It also demonstrates that the very good global FB and MG scores presented in Table IV.3 are due to error compensation. Moreover, we highlight that POD–GPRs tend to perform less well for low friction velocities (Fig. IV.19b, c, d, f, g), except for the FMS(1 ppm) which deteriorates for high friction velocities. This singular behavior for the FMS(1 ppm) is due to a zoning effect as the area corresponding

to the 1-ppm isoline decreases when u_* decreases. The MG distribution indicates that the POD–GPRs model tends to overestimate the low concentration at low friction velocities. For the other metrics, the decrease in POD–GPRs accuracy is due to an increase in the internal variability when u_* decreases, which makes the mean concentration inherently noisier and therefore impossible to perfectly predict.

To demonstrate this point, we estimate the error solely due to internal variability for each metric and each sample using the methodology presented in Sect. IV.3.4. Then we fit a parametric model for each metric to account for the dependency of the internal variability on the friction velocity. The resulting internal variability error models are represented as red lines in Fig. IV.19. The mean internal variability error is zero for FB and MG as these metrics only measure systematic bias and not noise. For all the other metrics, the error solely due to internal variability is inversely dependent on the friction velocity. This is because when the advection slows down, the time correlation of the tracer concentration increases, and the sampling error due to the lack of independent realizations over the 200-s analysis period therefore also increases.

We have also investigated the dependence of the POD–GPRs accuracy on the inlet wind direction parameter α_{inlet}, and it appears that POD–GPRs errors are more uniformly distributed in this dimension. We only find zoning effects for FMS(1 ppm) and FMS(0.01 ppm) as the plume delimited by these concentration levels occupies a smaller volume within the analysis domain when the wind direction carries the plume outside the containers array, i.e. for $\alpha_{inlet} \approx 30\,^{\circ}$or $\alpha_{inlet} \approx -90\,^{\circ}$.

The same exercise was applied to the POD–GPRs built using the $\mathcal{T}_{log}$ fields preprocessing. We find the same structure of dependence of the error on the friction velocity but with higher dispersion of every score which explains the better overall performances obtained with $\mathcal{T}_{log-1\mathrm{D}}$ (Table IV.3). We also note that without field normalization by the friction velocity, the POD–GPRs model does not reproduce the MG bias and has better FAC2 scores for the low friction velocities. This may indicate that the friction velocity similarity used for rescaling does not hold perfectly for low friction velocities. Finally, in both cases, the ensemble-averaged POD–GPRs prediction error is quite close to the mean internal variability error (Fig. IV.19). Therefore, our model reduction approach is, on average, close to the best achievable accuracy but can still be improved concerning the dispersion of its errors.

IV.6.5 Why is it primordial to restrict the number of modes?

In Sect.IV.5.1, we present a method for selecting, a priori, the number of POD modes as a compromise between the total variance embedded in the POD reduced basis and the amount of noise related to internal variability carried by the modes. For the mean concentration field, this approach gives an optimal value of approximately $L = 10$ modes. In this section, we check the relevance of this choice a posteriori. Results are given with the fields preprocessing $\mathcal{T}_{log}$, but similar results are found using $\mathcal{T}_{log\text{-}1\mathrm{D}}$.

Figure IV.20 shows that increasing the number of modes L reduces the model reduction error averaged over the test. This result is supported by three different metrics, i.e. NMSE, FMS(1 ppm), and VG, that assess the ability of the reduced-order model to

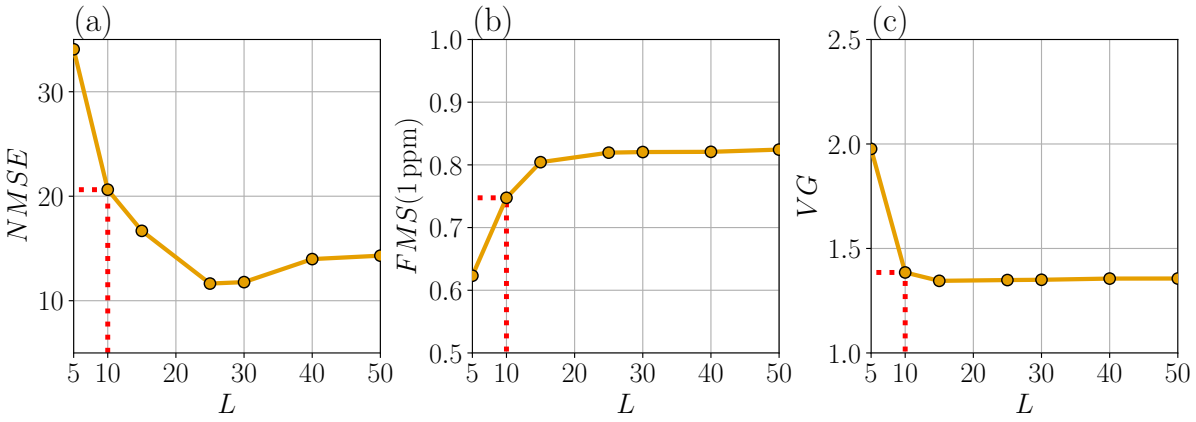


Figure IV.20: *POD–GPRs mean concentration estimation error as a function of the number of the modes L and evaluated with* NMSE *(a),* FMS(1 ppm) *(b), and* VG *(c) over the test set. Error levels corresponding to the selected number of modes ($L = 10$) are shown in red dotted lines.*

predict both high and low concentrations. As previously seen for the POD projection error (Fig. IV.9d), validation metrics reach a plateau but for a larger number of modes ($L \approx 15$–25) than the a priori retained number of modes ($L = 10$).

However, increasing the number of modes leads to a deterioration of the POD–GPRs response surfaces, as shown for the mean concentration at tower B at $z = 2\,\mathrm{m}$ in Fig. IV.21. It makes the response surfaces very noisy and even physically implausible. In particular, with $L = 50$ modes, the POD–GPRs model is no longer able to retrieve the inversely proportional dependence of concentration on friction speed expected from theory (Fig. IV.21b). This deterioration is because modes of high orders are likely to only account for noisy structures due to internal variability (Fig. IV.9a). They are therefore not informative on systematic structures due to the wind conditions, implying that high-order POD reduced coefficients are chaotically distributed in the input space (Fig IV.10c) and thus nearly impossible to predict correctly for the GPRs. These errors pile up to deteriorate the POD–GPRs predictions since they are defined as linear combinations of the reduced coefficients predicted by the GPRs (Eq. IV.4). In addition, increasing the number of modes also significantly increases the variance of the POD–GPRs because it is expressed as the exponential of the sum of the variances of the L GPRs that compose the POD–GPRs (Eqs. IV.29b and IV.32).

Interestingly the clear deterioration of the model response surface when using a large number of modes is not reflected in the overall performances (Fig. IV.20), except for the NMSE which ends up increasing when the number of modes becomes too large. This is explained because global metrics are somehow short-sighted as they compare the reduced-order model predictions with the test fields spatially before averaging each score obtained over the test set (Eq. IV.35). It demonstrates the need to avoid drawing conclusions based solely on scalar metrics.

In light of these tests, the a priori selection method for the number of modes presented in Sect. IV.5.1 seems very effective. Indeed, the chosen compromise appears to obtain

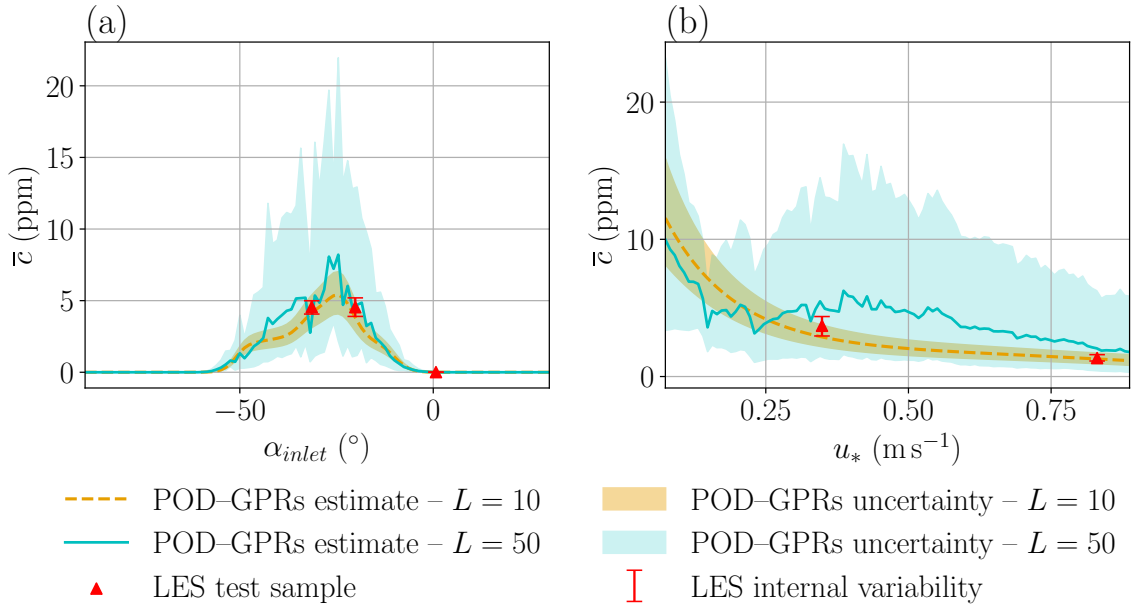


Figure IV.21: *POD–GPRs estimates of the mean concentration at tower B at $z = 2\,m$ as a function of the inlet wind direction α_{inlet} (a), and of the friction velocity u_* (b). When varying one parameter, the other is set constant to either $u_*^{plot} = 0.45\,m\,s^{-1}$ (a), or $\alpha_{inlet}^{plot} = -43\,°$ (b). The test samples closest to the two segments of parameter space thus scanned are represented by red triangles and are also identified in Fig. IV.5. Results are obtained using the $\mathcal{T}_{log}$ fields preprocessing and are given for two numbers of POD modes: $L = 10$ as the orange dashed line, and $L = 50$ as the cyan solid line. Associated shaded areas correspond to the 95% confidence intervals estimated by the POD–GPRs according to the procedure detailed in Sect. IV.2.5. The uncertainty of the LES test samples due to internal variability is also depicted as red error bars.*

very good validation scores (Table IV.3) while avoiding problems of response surface noise. We, therefore, continue to use the preselected value of $L = 10$ in the following.

Additional results obtained with the POD–GPRs reduced-order model, which are not essential to the main objective of this thesis, are presented in Appendix B. It includes an illustration of the ability of the POD–GPRs model to emulate other LES fields, an application to global sensitivity analysis, and a technique for getting the best out of both POD–GPRs predictions, with and without logarithmic transformation.

IV.7 Conclusion

Summary In this chapter, a data-driven reduced-order model is built as a surrogate for the LES dispersion model presented in Chapter II. To do so, we use the POD–GPRs method constructed in the thesis of Nony (2023), which involves a reduction step using Proper Orthogonal Decomposition (POD) and a regression step using independent Gaussian Process Regressors (GPRs). Following the conclusions of Chapter III on the main uncertainties of the LES model, we generate a database of LES realizations by varying the two parameters that have the most impact on the model predictions: the inlet wind direction and the friction velocity for which the theoretical dependency is known. The POD–GPRs reduced-order model is trained on one part of this database and we evaluate its ability to emulate the LES response surface for different wind conditions on the rest of the database. The construction and validation of the POD–GPRs model is focused on the mean concentration prediction as it is the quantity we seek to assimilate in the data assimilation framework proposed in Chapter V. Nevertheless, the POD–GPRs can emulate any other field predicted by the LES model, as illustrated in Appendix B.

Contributions This application to a more realistic context, i.e. the MUST dispersion field campaign, has enabled us to improve performance and our understanding of the POD–GPRs method even further. In particular, we show that the POD projection error heavily depends on the field preprocessing, which is, therefore, a key model design choice. In Sect. IV.5.2, we show that, for the mean concentration, using a log-transformation preprocessing instead of the linear centering used by Nony (2023) greatly improves the accuracy of the POD. Besides, we emphasize that validation should not be based solely on the Q2 metric, since it can fail to detect non-physical predictions. We also demonstrate in Sect. IV.6.1 that normalizing the fields by the friction velocity during the preprocessing improves the overall performances of the POD–GPRs, and we therefore recommend doing so.

Link with internal variability One of the major new contributions of this study is the use of the internal variability quantification approach proposed in Chapter III to build and validate the POD–GPRs model in an informed way. In particular, we develop in Sect. IV.5.1 a method to propagate the internal variability of the LES fields to the coefficients of the POD projection. From this method, we: i) show that POD modes of high order mostly encode the noise in the LES train samples, ii) propose a criterion to choose a priori the number of modes, iii) validate the GPRs noise hyperparameters found during their optimization in Sect. IV.5.3. In our case, we find an optimal value of 10 modes which is far less than the recommendation from Nony (2023). In addition, we validate a posteriori the relevance of this choice as we show that using more modes can, despite the better apparent performance, lead to non-physical POD–GPRs response surfaces (Sect. IV.6.5). Finally, internal variability is also used throughout POD–GPRs validation as it represents the best achievable accuracy.

Estimation of the POD–GPRs uncertainty To answer the general objective of this thesis of quantifying and controlling the dispersion model uncertainties, we propose a method to quantify the uncertainties on the POD–GPRs predictions. To this end, we derive in Sect. IV.2.5 the equations for propagating the variance predicted by the GPRs to the quantity of interest. We show that the resulting uncertainty accounts for both the GPR regression errors and the noise induced by internal variability. Finally, we verify in Sect. IV.6 that the estimated uncertainty properly explains the differences between LES test samples and POD–GPRs predictions. This, therefore, constitutes a new key asset of the POD–GPRs methodology. In particular, we use it in Chapter V for data assimilation to account for the reduced-order model error and prevent putting too much confidence in the model during the analysis. Nevertheless, note that POD projection error is not embedded yet in our estimation of the POD–GPrs uncertainty, and quantifying it would be an interesting prospect.

Choice of the best reduced-order model and perspectives Finally, we show that the POD–GPRs reduced-order model is very accurate at emulating the LES mean concentration predictions. It predicts particularly well the general shape of the plume and provides good-looking response surfaces for varying wind conditions (Sect. IV.6.1). In addition, the POD–GPRs model is very efficient, with a prediction of a complete 3D field in less than a tenth of a second. We also demonstrate that there is an added value in using POD–GPRs compared to a trivial model such as the Nearest Neighbor, both in terms of accuracy and robustness to decrease in the train set size. To conclude on the POD–GPRs limits, error dispersion among test samples is quite important and could be improved. More importantly, the POD–GPRs noticeably lacks accuracy in areas of high concentration. We demonstrate in Sect. IV.5.2 that it comes from POD projection errors. In Appendix B.3, we investigate the use of a pragmatic approach called Mixture-Of-Experts which combines POD–GPRs with both linear and log-transformation of the fields to get the best of both approaches. To improve it even more, better dimension reduction would be key in our opinion. This could be done by looking to better field prescaling, such as the one used by Mons et al. (2017), or using a mixture of PODs (Tipping and Bishop 1999), or nonlinear reduction techniques such as convolutional autoencoders (Murata et al. 2020; Nony 2023), but is outside the scope the present study. Another direct perspective of this work is its extension to handle the dependency on other parameters such as the source location or the atmospheric stability condition.

Chapter V

Data assimilation for wind condition estimation

In this chapter, we assemble and test the data assimilation (DA) system to address the thesis problematic of efficiently estimating and reducing uncertainties in LES models for microscale atmospheric dispersion.

To begin with, we present a comprehensive overview of related studies that we have identified in the literature. From their cross-analysis emerge important trends that have motivated the design of our DA system, in particular the choice of correcting meteorological boundary conditions and the use of a reduced-order model. We have also identified avenues for improvement based on this cross-analysis, particularly in error modeling, which motivates the main methodological contributions of this chapter.

To continue, we provide a concise introduction to fundamental DA concepts and subsequently present the ensemble Kalman filter (EnKF) algorithm, our chosen solution for solving the DA problem in a statistical way.

We then explain in detail how this algorithm is applied to the MUST trial 2681829 to estimate meteorological boundary conditions using local concentration measurements averaged over a finite time horizon. A major effort is made in our DA system to define realistic error models for observations, background parameters and model predictions, which include the uncertainty linked to the internal variability of the ABL.

A first step in validating the DA system is carried out with twin experiments, i.e. assimilating synthetic observations generated from an LES prediction. This provides a controlled environment for checking and calibrating the EnKF. In particular, we investigate the effect of the number of members and the effect of sampling noise from ensemble generation. In a second step, the DA system is tested by assimilating the real field measurements from the MUST field campaign. We evaluate the system's ability to correct the large-scale meteorological boundary conditions, and how these corrections improve mean concentration predictions.

Chapter outline

V.1 Introduction 171
V.1.1 Data assimilation review on the improvement of microscale atmospheric CFD models 171
V.1.2 Strategy adopted in this thesis 175
V.2 Data assimilation theoretical framework 178
V.2.1 Data assimilation problem and notations 178
V.2.2 The ensemble Kalman filter 181
V.3 Application to the MUST trial 2681829 184
V.3.1 Concentration observations and anamorphosis 184
V.3.2 Observation error covariance matrix 185
V.3.3 Uncertainty in background boundary conditions 187
V.3.4 Choice of the background parameters 189
V.3.5 Model definition and error 191
V.4 Validation and calibration of the data assimilation system . 193
V.4.1 Twin experiment principle 193
V.4.2 Results of the baseline twin experiment 195
V.4.3 Effect of the ensemble size and sampling error estimation . . . 198
V.5 Assimilation of the real field measurements 200
V.5.1 Results for the MUST experimental data assimilation 200
V.5.2 Effect of the concentration anamorphosis threshold 205
V.6 Conclusion 207

V.1 Introduction

This introduction aims to explain how data assimilation (DA) can improve CFD model predictions in microscale atmospheric contexts and to further elaborate the design of our reduced-cost DA system, whose general flowchart was already sketched in Chapter I. We provide a review of related studies in Sect. V.1.1, before positioning our solution in Sect. V.1.2.

Throughout this chapter, we use the main concepts, vocabulary and techniques of DA introduced in Sect. I.3.2, page 35.

V.1.1 Data assimilation review on the improvement of microscale atmospheric CFD models

Recent studies have highlighted the significant capabilities of DA in improving microscale or local-scale atmospheric CFD models. In this section, we present a cross-analysis of the main studies we have identified in the literature, summarized in Table V.1. Some of these studies are applied to microscale dispersion in urban environment (Mons et al. 2017; Aristodemou et al. 2019), including to the MUST experiment (Defforge et al. 2021); others to pedestrian wind comfort/natural ventilation (Sousa et al. 2018; Sousa and Gorlé 2019) or to wind energy resource assessment (Defforge et al. 2019; Bauweraerts and Meyers 2021).

Choice of the control vector A key aspect of DA system design is the choice of the control vector which defines the variables that are actually inferred by DA by translating discrepancies between the observations and the model counterparts into a correction in the control space. Concerning the control vector definition, the studies presented herein can be grouped into two main approaches: those opting for state estimation to correct initial conditions for the model as in traditional DA applications, and those opting for parameter estimation (respectively in blue and green in Table V.1). Defforge (2019) explains that this choice should be based on the ratio between the spatial scale and the temporal scale of the problem under consideration. In large-scale atmospheric simulation, the spatial extent of the domain L is much larger than the typical scale of motion, $L \gg T \times U$ with T the horizon time of the simulation and U the characteristic wind speed. This implies that initial conditions have a much larger impact on model predictions than boundary conditions, as illustrated in Fig. V.1a. The opposite occurs at the microscale ($L \ll T \times U$), which means that initial conditions quickly fade away and boundary conditions become predominant (Fig. V.1b). This theoretical argument is confirmed by the study of Aristodemou et al. (2019), which shows very limited persistence of the state correction with a microscale dispersion LES model.

Choice of the CFD approach LES is chosen for the problem of reconstructing a sequence of instantaneous atmospheric states (Aristodemou et al. 2019; Bauweraerts and Meyers 2021) since RANS only provides ensemble-averaged predictions. Meanwhile, both

Table V.1: *Summary of studies assessing the potential of DA at improving microscale atmospheric CFD predictions. For each selected study, we report the chosen algorithm, CFD modeling approach, and control vector, as well as the study case. We specify if the assimilated observations are synthetic, i.e. simulated, or real field measurements. We also mention when a reduced-order model or a reduced space is used within the DA framework. Studies opting for parameter estimation are highlighted in blue, and those opting for state estimation are in orange.* U_{inlet} *and* α_{inlet} *denote the incoming wind velocity and direction;* $\mathbf{x}_s$ *and* I_s *correspond to pollutant source location and intensity;* $u_{\mathrm{BC}}, v_{\mathrm{BC}}, k_{\mathrm{BC}}$ *are vertical profiles of wind velocity components and turbulent kinetic energy imposed as boundary conditions;* $\mathbf{c}$ *and* $\mathbf{u}$ *are the concentration and velocity fields, respectively.*

Study	DA algorithm	Study case	Model	Control Vector	Observations	Reduced
Mons et al. (2017)	POD–EnVar[1]	Shinjuku area of Tokyo	VLES[2]	$(U_{inlet}, \alpha_{inlet}, \mathbf{x}_s, I_s)$	Synthetic	$\emptyset$
Sousa et al. (2018) Sousa and Gorlé (2019)	EnKF	Stanford campus	RANS	$(U_{inlet}, \alpha_{inlet})$	Synthetic Real	Model
Aristodemou et al. (2019)	3D–Var	Small neighborhood in central London	LES	$\mathbf{c}$ *(state)*	Wind tunnel measurements	State
Defforge et al. (2019) Defforge et al. (2021)	IEnKS[3]	Wind turbine site MUST	RANS	$(u_{\mathrm{BC}}, v_{\mathrm{BC}})$ $(u_{\mathrm{BC}}, v_{\mathrm{BC}}, k_{\mathrm{BC}})$	Synth., Real Real	$\emptyset$
Bauweraerts and Meyers (2021)	4D–Var	Idealized ABL	LES	$\mathbf{u}$ *(state)*	Synthetic lidar measurements	State

[1] See Tian et al. (2011) and Yang et al. (2015).
[2] Very Large Eddy Simulation.
[3] See Bocquet and Sakov (2014).

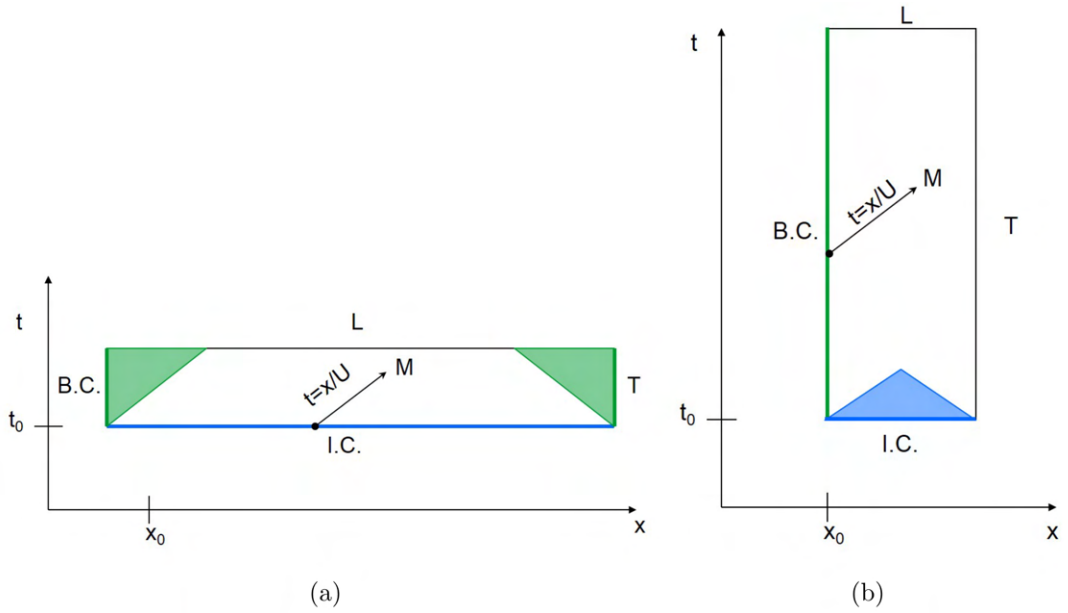


Figure V.1: *Schematic view from Defforge (2019) of the influence of initial and boundary conditions on the predictions of an atmospheric numerical model. Two scenarios are represented: large-scale meteorology (a), which typically corresponds to regional-to-synoptic scales in Table I.1, and microscale meteorology (b). L, T, and U are the associated characteristic scales of length, time, and velocity. a) The prediction is mainly influenced by the initial conditions (IC), except for the green-shaded areas. b) Conversely, the prediction is mostly defined by the boundary conditions (BC).*

approaches are suitable for parameter estimation using time-averaged observations (Table V.1). To the best of our knowledge, no study in this context has compared DA performance using RANS and LES. One can expect the DA problem to be better posed using LES as it has reduced turbulence modeling uncertainties (Gousseau et al. 2011; García-Sanchez et al. 2018). However, the increased computational cost of LES may limit the number of model predictions in the DA algorithm, and thus its accuracy. Finally, we note that, by adding the turbulent kinetic energy in the control vector, the DA system of Defforge et al. (2021) based on a RANS model can estimate wind profiles that are not in equilibrium. While the perturbation is readily imposed on the turbulent kinetic energy which is a transported quantity in RANS, it requires more care in LES as realistic perturbations have to be added to the resolved turbulent flow field.

These studies also show the relevant use of reduction techniques. For instance, Aristodemou et al. (2019) and Bauweraerts and Meyers (2021) reduce the state dimension using singular value decomposition and proper orthogonal decomposition (POD). Mons et al. (2017) also employ POD but in the ensemble space. In this way, they mitigate the sampling error caused by the limited number of model evaluations for DA given the high cost of CFD models (20 to 60 model evaluations in their study). Finally, Sousa et al. (2018) build a reduced-order model based on polynomial chaos expansion (PCE) (Wiener 1938) as a surrogate for the CFD model. This drastically reduces the DA computational

cost and enables the use of more ensemble members compared to other studies (Mons et al. 2017; Defforge et al. 2019, 2021), at the expense of introducing a new source of error in the DA scheme, i.e. the model reduction error. However, this disadvantage may be very limited, for example, Rochoux et al. (2014a) show that replacing a fire propagation model with a surrogate model based on PCE leads to similar DA performance. Note that it is possible to adopt a different approach to solving the CFD model cost problem, for instance, Defforge et al. (2019) use a state-of-the-art DA algorithm, the IEnKS (Bocquet and Sakov 2014), to reduce the number of model predictions required compared to a standard 3D–Var algorithm.

Choice of the observable The choice of observable (i.e. the quantity of interest in the observation space) is another important aspect that emerges from this literature cross-analysis. First, we mention that some studies present twin experiments, i.e. assimilate synthetic observations obtained from a model prediction. This is the first usual level of validation of a DA system, ensuring that the inverse problem is well-posed. Some studies then go a step further, assimilating field observations to demonstrate the robustness and maturity of their DA system. The rather exploratory state estimation studies of Aristodemou et al. (2019) and Bauweraerts and Meyers (2021) assimilate a large amount of data, covering entire planes, thanks to laboratory measurement resources and a modeled lidar. All other studies assimilate a small number of local observations (between 1 and 33) from measurement masts. For dispersion studies, Mons et al. (2017) and Defforge et al. (2021) both demonstrate that assimilating wind measurements in addition to tracer concentration measurements greatly improves the reconstruction of the control vector. Sousa et al. (2018) also show that sensor position has a critical effect on DA performance. In particular, the DA problem becomes severely ill-posed if sensors are placed in areas where the local flow is mostly determined by the local building arrangement. To a priori optimize sensor positions for DA, Sousa et al. (2018) propose a brute force strategy, while Mons et al. (2017) develop a method based on model sensitivity analysis. As these methods require multiple model predictions, both studies employ a reduced-order model as a surrogate for the CFD model.

Error modeling We note that error modeling – which is central to DA – can be improved compared to the studies presented in Table V.1. On the one hand, background parameter errors are rigorously estimated from micro-climatology constructed from nearby weather stations (Sousa et al. 2018; Sousa and Gorlé 2019; Defforge et al. 2021) or mesoscale atmospheric model simulations (Defforge et al. 2019). However, this can introduce representativeness errors as the meteorological conditions in the area of interest can significantly differ from observations at more distant stations or from large-scale model simulations due to local effects (Pimont et al. 2017). On the other hand, observation errors are chosen arbitrarily in all studies, at best by an expert's choice. Observation errors are also assumed to be uncorrelated, which is probably incorrect as stated by Defforge et al. (2021). However, Sousa et al. (2018) show through sensitivity tests that observation error estimation significantly affects DA accuracy. Ultimately, model error is never taken into account in the DA algorithm, except for the study by

Sousa and Gorlé (2019), which quantifies the reduced-order model error compared to the CFD model used as a reference. Still, this may underestimate the total uncertainty in CFD model predictions.

To further improve DA applications in microscale atmospheric applications, there is a need for an appropriate representation of both numerical predictions and observation errors (Mons et al. 2017). In addition, we note that most studies only take into account errors originating from the measuring instruments, but as Chapter III shows, the uncertainty linked to the internal variability of the ABL can be of the same order of magnitude or even significantly larger than measurement errors. This emphasizes the need to represent internal variability in error models.

V.1.2 Strategy adopted in this thesis

In this thesis, we propose a supplementary proof of concept of the potential of DA to enhance microscale atmospheric CFD models. Drawing on the state of the art presented in Sect. V.1.1, we present our own DA system in detail in this section.

Choice of the control vector First, we choose to use DA for estimating the large-scale meteorological boundary conditions of the LES model of the MUST field experiment presented in Chapter II. This is motivated by the huge impact of boundary condition uncertainties on microscale CFD models (García-Sánchez et al. 2014; Lucas et al. 2016; Wise et al. 2018); reducing them via DA would thereby significantly improve prediction accuracy (Mons et al. 2017; Sousa et al. 2018; Sousa and Gorlé 2019; Defforge et al. 2019, 2021). In this thesis, the control vector $\boldsymbol{\theta}$ is therefore defined as

$$\boldsymbol{\theta} = (\alpha_{inlet}, u_*)^{\mathrm{T}}, \tag{V.1}$$

with α_{inlet} and u_* the direction and friction velocity of the incident flow. We selected these two boundary condition parameters as they are the ones that have the greatest impact on LES predictions, according to the sensitivity tests presented in Chapter III. More complex parameterization of the boundary conditions could be adopted, thereby increasing the control space dimension and thus the number of degrees of freedom to solve the DA problem (Defforge et al. 2019). However, this aspect is outside the scope of the present study. Moreover, we do not include the pollutant source position in the control vector, as we assume that this is perfectly identified in the MUST field experiment (Biltoft 2001).

Use of a reduced-order model Following the approach proposed by Rochoux et al. (2014a) and Sousa et al. (2018), we use a reduced-order model as a surrogate for the LES model to drastically reduce the cost of the DA system, and even open the way to real-time applications. To do so, we rely on the POD–GPRs surrogate model built in Chapter IV, which emulates the LES surface response for varying incoming wind directions and velocities, and which can provide accurate mean concentration field predictions in less than a tenth of a second. This allows us to carry out a larger number of experiments to adjust

and optimize the DA system we propose, which is particularly relevant as this system is built from scratch. In addition, benefiting from the estimation of the POD–GPRs uncertainty (see Sect. IV.2.5, page 131), we take into account the model reduction error in the DA system. Finally, we emphasize that this approach does not make the CFD model cost problem disappear, but shifts this cost problem to the offline construction of the reduced-order model, which requires learning from a database of LES predictions.

Choice of the main quantity of interest In this thesis, we aim to improve the prediction of the mean concentration field using in-situ concentration measurements. We do not assimilate other measurements such as wind speed and direction provided by local anemometers, as this would greatly simplify the DA problem. The relationship between large-scale meteorological conditions and local flow is indeed relatively straightforward given the simplified urban canopy of the MUST case. In this way, our DA system would require less measurement instrumentation efforts to be used in an operational context. Here, we consider averages over the 200-s analysis period defined by Yee and Biltoft (2004). This is a first step in developing a sequential DA system that assimilates a series of observations over small periods to correct in real time the concentration field evolution.

Accounting for microscale internal variability An important innovation in our approach is to take into account the irreducible uncertainties related to the microscale internal variability of the ABL in the modeling of observation and model errors. For this purpose, we benefit directly from the developments made in Chapters III and IV to i) quantify the associated uncertainty using the stationary bootstrap algorithm from Politis and Romano (1994), and ii) take into account this uncertainty in the reduced-order model error thanks to the ability of GPR models to accurately identify the noise related to internal variability. This is an important step forward in guaranteeing the robustness of DA for real-world microscale dispersion applications.

Choice of the data assimilation algorithm Among the various DA algorithms reported in the literature (Carrassi et al. 2018), we choose to use an EnKF (Evensen 1994, 2003; Houtekamer and Mitchell 1998). It is an off-the-shelf and easy-to-implement algorithm, which is regarded as a benchmark solution for DA problems (Asch et al. 2016; Carrassi et al. 2018). This algorithm relies on an ensemble of model queries to handle non-linearity in the model response surface and to mitigate the effect of model errors, which is relevant in our case given the strong impact of internal variability in model predictions (see Chapter III). The independence of the ensemble members, which makes the algorithm highly parallelizable, and the use of a reduced-order model make it possible to use very large ensembles to reduce sampling errors while still providing fast estimates. Finally, as an ensemble-based method, the EnKF provides a probabilistic estimate of the state of the system, which is fully in line with the general concern of this thesis and is highly valuable for risk assessment applications.

Figure V.2 shows a schematic of the complete reduced-cost DA system designed in this thesis to reduce uncertainty on meteorological boundary conditions and thus improve

microscale dispersion predictions based on LES. Notations as well as the EnKF algorithm are properly introduced in the following section.

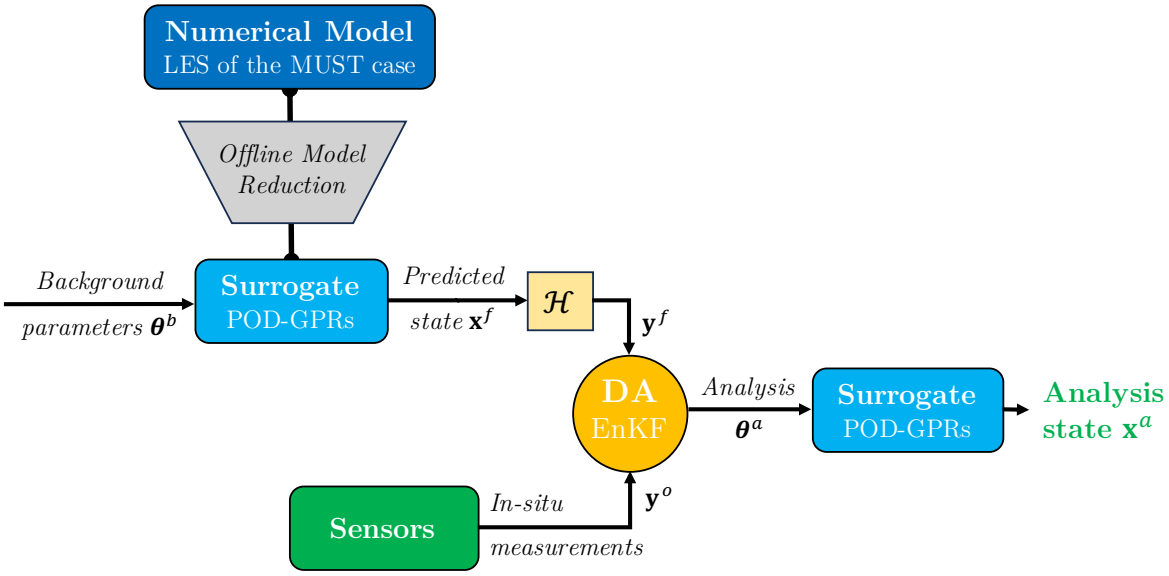


Figure V.2: *Specification of the reduced-cost DA system designed to address the thesis problematic. The pre-trained POD–GPRs reduced-order model constructed in Chapter IV is used as a surrogate of the microscale LES dispersion model presented in Chapter II. The POD–GPRs prediction of the physical system state $\mathbf{x}^f$, corresponding to prior and uncertain meteorological boundary condition parameters $\boldsymbol{\theta}^b$, is mapped onto the observation space using the observation operator $\mathcal{H}$. An EnKF algorithm is then used to estimate corrected input parameters $\boldsymbol{\theta}^a$ based on in-situ measurements $\mathbf{y}^o$. The corresponding corrected estimate of the system state $\mathbf{x}^a$ is finally obtained by a new reduced-order model prediction.*

V.2 Data assimilation theoretical framework

In this section, we introduce the essential theoretical basis of DA based on the book by Asch et al. (2016), before presenting the EnKF algorithm. Wherever possible, we use standard notations as defined by Ide et al. (1997).

V.2.1 Data assimilation problem and notations

The objective of DA is to improve the estimation of the control vector by optimally combining available observations, the prediction of the physical model, and all prior knowledge on the control vector, taking into account their respective uncertainties. In this thesis, the control vector is restricted to meteorological boundary condition parameters, as motivated in Sect. V.1.2. The following presentation is therefore given from the perspective of parameter estimation. Moreover, we do not consider the time dimension as the DA problem we are interested in is about estimating time-averaged variables.

The background error corresponds to the difference between the prior estimation of the control vector $\boldsymbol{\theta}^b$, called *background*, and its true value $\boldsymbol{\theta}^t$:

$$\mathbf{e}^b = \boldsymbol{\theta}^b - \boldsymbol{\theta}^t, \tag{V.2}$$

that is unknown. However, it is possible to estimate a priori the background error statistics, and in particular the background error covariance matrix:

$$\mathbf{B}_{ij} = \mathbb{E}\left[[\mathbf{e}^b]_i [\mathbf{e}^b]_j\right]. \tag{V.3}$$

This is a symmetric matrix of dimension $d \times d$, with d the dimension of the control space. It is assumed to be positive definite, hence invertible. Moreover, the background error is assumed to be unbiased, i.e. $\mathbb{E}[\mathbf{e}^b] = 0$, or the bias has been subtracted. These are standard DA assumptions (Asch et al. 2016).

The matrix $\mathbf{B}$ is a key input to the DA problem since it quantifies the uncertainty on the background control vector, and therefore the confidence that can be attributed to it.

The model error corresponds to the misfit between the true physical system state $\mathbf{x}^t$ and the state $\mathbf{x}^f$ predicted by the numerical model $\mathcal{M}$ using the background control vector as input:

$$\begin{aligned} \mathbf{e}^m &= \mathbf{x}^f - \mathbf{x}^t, \\ &= \mathcal{M}(\boldsymbol{\theta}^b) - \mathbf{x}^t. \end{aligned} \tag{V.4}$$

The state vector $\mathbf{x} \in \mathbb{R}^N$ denotes a field or a collection of fields discretized on a grid of dimension N. The model error represents all errors that occur during a model prediction; they can be linked to the choice of the computational mesh, the choice of the physical parameterizations, the calibration of the numerical or physical parameters, etc. Note that considering the very fine mesh used in this thesis (resolution of 60 cm in the area of interest, see Sect. IV.4.3.4, page 141), the representativeness error related to mesh

interpolation is assumed to be negligible here. The approach we propose to account for the model error $\mathbf{e}^m$ is described in Sect. V.3.5.

The observation error quantifies how available observations[1] $\mathbf{y}^o$ differ from the truth:

$$\mathbf{e}^o = \mathbf{y}^o - \mathcal{H}(\mathbf{x}^t), \tag{V.5}$$

where $\mathcal{H}$ denotes the observation operator that maps the model state vector onto the observation space. It can be a complex and nonlinear function, typically when using indirect satellite or lidar observations. However, in this thesis, since the available observations are direct in-situ measurements, the $\mathcal{H}$ operator is a simple selection matrix that interpolates the state vector at sensor positions. We also introduce the observation error covariance matrix[2]:

$$\mathbf{R}_{ij} = \mathbb{E}\left[[\mathbf{e}^o]_i[\mathbf{e}^o]_j\right], \tag{V.6}$$

of dimension $p \times p$ with p the number of available observations. As for the background error, a standard DA assumption is that observation error is unbiased, i.e. $\mathbb{E}(\mathbf{e}^o) = 0$ (Asch et al. 2016). The matrix $\mathbf{R}$ is an input of primary importance for DA, as it quantifies the uncertainty related to observations.

Estimation problem

DA aims to optimally estimate the control vector $\boldsymbol{\theta}$ using the background estimation $\boldsymbol{\theta}^b$ and associated uncertainty $\mathbf{B}$, the predictive model $\mathcal{M}$, the available measurements $\mathbf{y}^o$ and related uncertainty $\mathbf{R}$.

The DA solution to this estimation problem $\boldsymbol{\theta}^a$ is called the *analysis*. There are of course many ways to define what is an optimal estimation. Yet, in DA the optimal solution is defined as the one that minimizes the magnitude of the analysis error:

$$\mathbf{e}^a = \boldsymbol{\theta}^a - \boldsymbol{\theta}^t, \tag{V.7}$$

which can be done by minimizing the trace of the analysis error covariance matrix $\mathbf{A}$ defined as follows:

$$\mathbf{A}_{ij} = \mathbb{E}\left[[\mathbf{e}^a]_i[\mathbf{e}^a]_j\right]. \tag{V.8}$$

In addition to the analysis, DA techniques can estimate $\mathbf{A}$ and thus provide information on the uncertainty that comes with the analysis $\boldsymbol{\theta}^a$ considering all sources of uncertainty.

[1]The vector of observations, usually denoted $\mathbf{y}^o$ in DA (Ide et al. 1997), should not be confused with the predictions of the reduced models noted $\mathbf{y}$ in Chapter IV.

[2]Note that the Reynolds tensor is denoted $\mathbf{R}$ in Chapters II, III and IV, but the observation error covariance matrix is also denoted by $\mathbf{R}$ by convention in DA (Ide et al. 1997). As the Reynolds tensor is not involved in this chapter, we retain the usual notation for the observation error covariance matrix for the rest of the chapter.

Optimal estimation under simplified assumptions Assuming that the observation operator and the model operator are linear and perfect and that the background estimation is unbiased, the analysis can be written using statistical interpolation as a linear combination of all available information:

$$\boldsymbol{\theta}^a = \boldsymbol{\theta}^b + \mathbf{K}\left(\mathbf{y}^o - \mathcal{G}(\boldsymbol{\theta}^b)\right), \tag{V.9}$$

where $\mathcal{G} = \mathcal{H} \circ \mathcal{M}$ is the generalized observation operator combining model integration and interpolation at sensor positions for parameter estimation, and where $\mathbf{K}$ is a gain matrix. By minimizing the trace of the analysis error covariance matrix $\mathbf{A}$ associated with Eq. V.9, it is possible to derive the optimal gain $\mathbf{K}^*$, also known as the Kalman gain:

$$\mathbf{K}^* = \mathbf{B}\mathbf{G}^{\mathrm{T}}\left(\mathbf{R} + \mathbf{G}\mathbf{B}\mathbf{G}^{\mathrm{T}}\right)^{-1}, \tag{V.10}$$

where $\mathbf{G}$ is the linear generalized observation operator. By replacing $\mathbf{K}$ by $\mathbf{K}^*$ in Eq. V.9, we obtain the so-called best linear unbiased estimator (BLUE), whose error covariance matrix reads

$$\begin{aligned}\mathbf{A} &= (\mathbf{I} - \mathbf{K}^*\mathbf{G})\mathbf{B}(\mathbf{I} - \mathbf{K}^*\mathbf{G})^{\mathrm{T}} + \mathbf{K}^*\mathbf{R}\mathbf{K}^{*\mathrm{T}}, \\ &= (\mathbf{I} - \mathbf{K}^*\mathbf{G})\,\mathbf{B},\end{aligned} \tag{V.11}$$

with $\mathbf{I}$ the identity matrix. It is possible to show that this linear solution to the DA problem is also an analytical solution to Bayes' theorem when considering linear model and observation operators, and that background and observation errors follow Gaussian distributions of zero mean (Asch et al. 2016).

Note that if we also consider that control and observation spaces are one-dimensional, the Kalman gain can be simply written as

$$K^* = \frac{{\sigma_b}^2}{{\sigma_b}^2 + {\sigma_o}^2}, \tag{V.12}$$

with σ_b and σ_o the background and observation error variance, i.e. the one-dimension counterparts of $\mathbf{B}$ and $\mathbf{R}$. This highlights that the analysis (Eq. V.9) is the result of a balance between the background and observation uncertainties. Thus, if we underestimate background uncertainty, the analysis will tend to stick to the prior estimate as $K \to 0$. Conversely, if we underestimate the observation error, the analysis will overfit the observations and will ignore the background information. In this one-dimensional case, the analysis error variance σ_a can be expressed as

$$\frac{1}{{\sigma_a}^2} = \frac{1}{{\sigma_b}^2} + \frac{1}{{\sigma_o}^2}, \tag{V.13}$$

meaning that the confidence in the analysis is the sum of the confidence in the observation and the confidence in the background. All the proofs of the results stated in this section can be found in the book by Asch et al. (2016) or in the DA introduction lectures notes from Bocquet (2014).

This very simplified example in 1-D (Eqs. V.12–V.13) highlights that uncertainty estimation is at the core of DA, and the importance of providing accurate error estimates in a DA algorithm. Note that in real applications, the dimensions of the control, state and observation spaces can be very large, and the model and observation operators can be highly nonlinear, thereby requiring the use of more advanced DA techniques.

V.2.2 The ensemble Kalman filter

As motivated in Sect. V.1.2, we choose to solve the DA problem using the EnKF algorithm (Evensen 1994, 2003; Houtekamer and Mitchell 1998; Ehrendorfer 2007). It is a variant of the Kalman filter (Eqs V.9–V.10), which requires neither model linearization, nor the propagation of large matrices as in Eqs V.10–V.11, which can be difficult when dealing with high dimensional spaces. The EnKF relies on a collection of control vectors, called the members of the ensemble, which are meant to be representative of the uncertainty in the control vector and which are used to compute the Kalman Gain using a Monte Carlo approach. Yet, it differs from the particle filter (Gordon et al. 1993) as it relies on the assumptions that observation and background distributions are normally distributed. Since the model dynamics are evaluated for each ensemble member, the EnKF is able to (at least partly) account for model non-linearities. However, it only provides an exact solution to Bayes' theorem when the model is linear and when error distributions are Gaussian. Nevertheless, the EnKF has proven to be very efficient on a large number of academic and operational DA problems including nonlinear problems (Asch et al. 2016).

Among the reported implementations of the EnKF (Houtekamer and Zhang 2016), we use in this thesis the stochastic EnKF (Burgers et al. 1998; Houtekamer and Mitchell 1998). Note that EnKF is usually applied sequentially as observations become available, but since we are considering time-averaged quantities in this study, it can be simplified to a single assimilation cycle. The complete procedure is given in Algorithm 1. It relies on two main steps:

a) the prediction step, in which a prediction is made for each of the N_e members $(\boldsymbol{\theta}_i^b)_{i=1}^{N_e}$ that sample the background control vector distribution $\mathcal{N}(0, \mathbf{B})$,

b) the analysis step, where the error cross-covariance matrices $\mathbf{B}\mathbf{G}^{\mathrm{T}}$ and $\mathbf{G}\mathbf{B}\mathbf{G}^{\mathrm{T}}$ are estimated using the ensemble members as follows:

$$\mathbf{B}\mathbf{G}^{\mathrm{T}} \approx \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\boldsymbol{\theta}_i^b - \overline{\boldsymbol{\theta}}^b) \left(\mathcal{H}(\mathbf{x}_i^f) - \overline{\mathcal{H}(\mathbf{x}_i^f)} \right)^{\mathrm{T}} \tag{V.14}$$

$$\mathbf{G}\mathbf{B}\mathbf{G}^{\mathrm{T}} \approx \frac{1}{N_e - 1} \sum_{i=1}^{N_e} \left(\mathcal{H}(\mathbf{x}_i^f) - \overline{\mathcal{H}(\mathbf{x}_i^f)} \right) \left(\mathcal{H}(\mathbf{x}_i^f) - \overline{\mathcal{H}(\mathbf{x}_i^f)} \right)^{\mathrm{T}}. \tag{V.15}$$

where $(\mathbf{x}_i^f)_{i=1}^{N_e} = (\mathcal{M}(\boldsymbol{\theta}_i^b))_{i=1}^{N_e}$ is the ensemble of predictions associated to the background control vector ensemble obtained from the first step. These estimates are then used to compute the Kalman gain using Eq. V.10, and the analysis is estimated independently for each member using Eq. V.9.

Algorithm 1 Stochastic Ensemble Kalman Filter (EnKF) for parameter estimation

Inputs

▷ Ensemble size N_e

▷ Background parameters $\boldsymbol{\theta}^b \in \mathbb{R}^d$

▷ Background error covariance matrix $\mathbf{B} \in \mathbb{R}^{d \times d}$

▷ Measurements $\mathbf{y} \in \mathrm{R}^p$

▷ Observation error covariance matrix $\mathbf{R} \in \mathbb{R}^{p \times p}$

Outputs

▷ Parameter analysis ensemble $\mathbf{E}^a = (\boldsymbol{\theta}_1^a, \ldots, \boldsymbol{\theta}_i^a, \ldots, \boldsymbol{\theta}_{N_e}^a) \in \mathbb{R}^{d \times N_e}$

Initialization

Sample the background ensemble

$$\mathbf{E}^b = (\boldsymbol{\theta}_1^b, ..., \boldsymbol{\theta}_i^b, ..., \boldsymbol{\theta}_{N_e}^b)^{\mathrm{T}} \text{ with } \boldsymbol{\theta}_i^b \sim \mathcal{N}(\boldsymbol{\theta}^b, \mathbf{B})$$

a) Prediction step

for $i = 1, ..., N_e$ **do**

Propagate the parameters in state space

$$\mathbf{x}_i^f = \mathcal{M}(\boldsymbol{\theta}_i^b)$$

b) Analysis step

Estimate the error covariance matrix

$$\mathbf{BG}^{\mathrm{T}} = \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\boldsymbol{\theta}_i^b - \overline{\boldsymbol{\theta}}^f) \left(\mathcal{H}(\mathbf{x}_i^f) - \overline{\mathcal{H}(\mathbf{x}_i^f)} \right)^{\mathrm{T}}$$

$$\mathbf{GBG}^{\mathrm{T}} = \frac{1}{N_e - 1} \sum_{i=1}^{N_e} \left(\mathcal{H}(\mathbf{x}_i^f) - \overline{\mathcal{H}(\mathbf{x}_i^f)} \right) \left(\mathcal{H}(\mathbf{x}_i^f) - \overline{\mathcal{H}(\mathbf{x}_i^f)} \right)^{\mathrm{T}}$$

Compute the Kalman gain $\mathbf{K}^* = \mathbf{BG}^{\mathrm{T}}(\mathbf{GBG}^{\mathrm{T}} + \mathbf{R})^{-1}$

Sample an ensemble of perturbed observations

$$\mathbf{Y} = (\mathbf{y}_1^o, ..., \mathbf{y}_{N_e}^o)^{\mathrm{T}} \text{ with } \mathbf{y}_i^o \sim \mathcal{N}(\mathbf{y}, \mathbf{R})$$

for $i = 1, ..., N_e$ **do**

Update each ensemble member $\boldsymbol{\theta}_i^a = \boldsymbol{\theta}_i^b + \mathbf{K}^*(\mathbf{y}_i^o - \mathcal{H}(\mathbf{x}_i^b))$

From the ensemble of analysis members obtained from the EnKF (Algorithm 1), we derive the optimal estimate as the ensemble-average vector:

$$\overline{\boldsymbol{\theta}^a} = \frac{1}{N_e} \sum_{i=1}^{N_e} \boldsymbol{\theta}_i^a, \tag{V.16}$$

but also an estimation of the analysis error covariance matrix:

$$\mathbf{A} = \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\boldsymbol{\theta}_i^a - \overline{\boldsymbol{\theta}^a})(\boldsymbol{\theta}_i^a - \overline{\boldsymbol{\theta}^a})^{\mathrm{T}}. \tag{V.17}$$

It is worth noting that each member is independent, implying that each prediction and analysis are also independent, making the EnKF easily parallelizable and thus computationally efficient. Only $\mathbf{BG}^{\mathrm{T}}$ and $\mathbf{GBG}^{\mathrm{T}}$ computations require all the members together.

The main specificity of the stochastic EnKF lies in the fact that, during the analysis, each member is compared with a different observation vector $\mathbf{y}_i^o$, obtained by perturbing the actual observations $\mathbf{y}_i^o \sim \mathcal{N}(\mathbf{y}^o, \mathbf{R})$. If the observations were not perturbed, the analysis error would be underestimated (Burgers et al. 1998; Houtekamer and Mitchell 1998). In particular, under the assumptions of linear operators ($\mathcal{H}$ and $\mathcal{M}$) and normally distributed errors, and in the limit of a large ensemble size, the resulting analysis error covariance matrix (Eq. V.17) would tend to

$$\mathbf{A} = (\mathbf{I} - \mathbf{K}^*\mathbf{G})\mathbf{B}(\mathbf{I} - \mathbf{K}^*\mathbf{G})^{\mathrm{T}},$$

instead of the theoretical optimal value (Eq. V.11).

Beyond the intrinsic limitations linked to its optimality assumptions (linear model and Gaussian-distributed errors), the EnKF is affected by sampling noise resulting from the restriction on the ensemble size (Ehrendorfer 2007; Asch et al. 2016). While sampling errors due to observation perturbations can be avoided by using more sophisticated implementations such as the ensemble-transform Kalman filter (Bishop et al. 2001; Houtekamer and Zhang 2016), sampling errors due to background ensemble sampling are inherent to the EnKF. In Sect. V.4.3, we quantify the background sampling noise for our case study.

Two techniques are commonly used in DA to enhance EnKF estimations: localization and inflation (Anderson and Anderson 1999; Hamill et al. 2001; Ehrendorfer 2007). Localization aims at filtering non-physical correlations between variables of the control vector that can appear in the analysis covariance matrix $\mathbf{A}$ as assimilation cycles progress. Inflation consists of increasing the spread of the ensemble after each analysis to avoid ensemble collapse. However, we do not use these standard techniques in this thesis since i) we only consider one cycle, which prevents the accumulation of errors; and ii) the dimension of the control vector is very low ($d = 2$) which limits sampling errors compared to state estimation problems where $d = N \gg N_e$.

V.3 Application to the MUST trial 2681829

In this section, we design a numerical experiment to evaluate the ability of the EnKF algorithm to improve the accuracy of the LES model of the MUST trial 2681829 presented in Chapter II. We define a scenario where meteorological data available are incomplete or highly uncertain and in which the wind boundary conditions of the model are inferred using in-situ measurements provided by the observation network defined in Sect. V.3.1. We provide realistic models for the so-called observation and background error covariance matrices $\mathbf{R}$ (Sect. V.3.2) and $\mathbf{B}$ (Sect. V.3.3). The prior wind conditions are specified in Sect. V.3.4, and the predictive model alongside its uncertainty in Sect. V.3.5.

V.3.1 Concentration observations and anamorphosis

As a reference observation network, we use one of the setups employed by Defforge et al. (2021) for the same MUST trial. It includes 13 tracer concentration sensors, distributed over towers C (at $z = 1, 2, 3\,\mathrm{m}$), D ($z = 1, 2, 3\,\mathrm{m}$) and T ($z = 1, 2, 4, 6, 8, 10\,\mathrm{m}$), plus the DPID sensor #26 (at $z = 1.6\,\mathrm{m}$). Figure V.3 shows the location of these different sensors within the container array. Sensor measurements that are not assimilated are used for validation. We use time-averaged measurements over the 200-s analysis period defined by Yee and Biltoft (2004). As mentioned in Sect. V.1.2, we limit our study to the assimilation of concentration measurements.

We exclude the observations that are under the sensor detection threshold for both assimilation and validation steps. Indeed, while these near-zero concentration measurements may be informative about the orientation and size of the plume, they may also correspond to a problem with the sensors (as in the case of two sensors at tower C and one at tower D), and assimilating them could lead to inconsistent results.

As tracer concentration values are always positive or zero, observation errors are not normally distributed. To ensure better compliance with the assumptions of the EnKF (Algorithm 1), we use the concentration anamorphosis used by Liu et al. (2017) and Defforge et al. (2021). This consists in assimilating the log-transformed mean concentration observations $\tilde{\mathbf{y}}$ instead of the actual concentration measurements $\mathbf{y}$:

$$\tilde{\mathbf{y}} = \ln(\max(\mathbf{y}, 0) + y_t), \tag{V.18}$$

with y_t a given threshold to avoid putting too much weight on low-concentration values. In this study, we use the maximum sensor detection threshold $y_t = 0.04\,\mathrm{ppm}$. A sensitivity test of the DA performance to the choice of y_t is carried out in Sect. V.5.2. This transformation relies on the assumption that concentration observations follow a log-normal distribution, which is a fairly standard assumption (Cassiani et al. 2020). Note that Eq. V.18 is the same transformation as the preprocessing (Eq. IV.22) used in Chapter IV, and with which we obtain the best POD reduction accuracy (see Sect. IV.5.2, page 148).

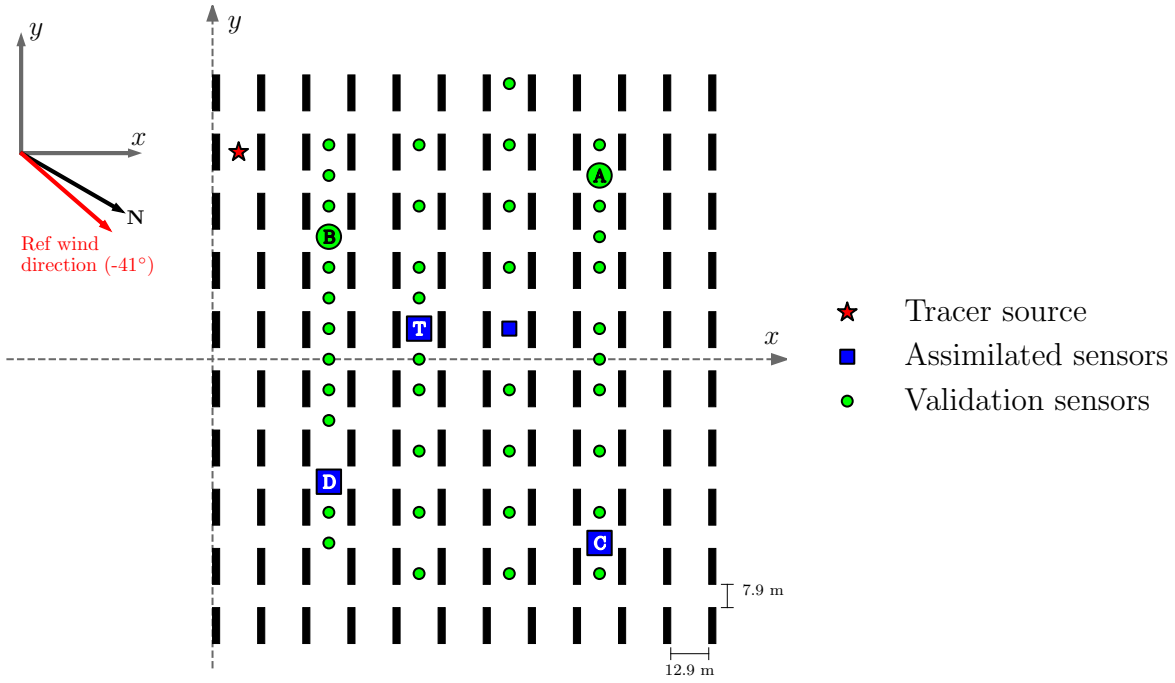


Figure V.3: *Schematic view of the reference sensor network used for DA. Blue squares correspond to sensors from which measurements are assimilated, and green circles correspond to validation sensors. Towers A, B, C, D and the central tower T are identified by the corresponding letters. Black rectangles represent the shipping containers used to mimic the urban canopy. The propylene-source location of trial 2681829 is indicated by the red star. The red arrow corresponds to the reference upstream wind direction estimated using all measurements available (see Sect. II.4.2.1, page 62). For a more detailed overview of the different types of sensors, we refer the reader to Fig. II.3, page 59.*

V.3.2 Observation error covariance matrix

We consider three contributions to the observation error: i) the sensor uncertainty $\mathbf{e}^{\mu}$, ii) the uncertainty induced by the internal variability of the ABL $\mathbf{e}^{\nu}$ (see Chapter III), and iii) the representativeness error $\mathbf{e}^{\mathbf{r}}$ that corresponds to the information lost in the discretization of the system state. The observation error (Eq. V.5) thus reads:

$$\mathbf{e}^{o} = \mathbf{e}^{\mu} + \mathbf{e}^{\nu} + \mathbf{e}^{r}. \tag{V.19}$$

Among these three sources of errors, we already assumed in Sect. V.2.1 that representativeness errors $\mathbf{e}^{r}$ are negligible considering the fine grid resolution used to represent the system state. For the instrument measurement error $\mathbf{e}^{\mu}$, we did not find information either in the technical report (Biltoft 2001) or in the main MUST modeling studies (Yee and Biltoft 2004; Milliez and Carissimo 2007; Dejoan et al. 2010; Nagel et al. 2022). Moreover, we demonstrated in Chapter III that concentration measurements are subject to significant aleatory uncertainty due to the internal variability of the ABL. Therefore, we make the assumption that internal variability $\mathbf{e}^{\nu}$ accounts for most of the observation

error, hence:

$$\mathbf{e}^o \approx \mathbf{e}^\nu. \tag{V.20}$$

Note that this is the case for the wind velocity as we demonstrate in Sect. III.4.1.2, page 94, that internal variability outweighs anemometer measurement errors.

To estimate statistics of the observation error due to internal variability, we use the stationary bootstrap procedure based on sub-averages resampling and presented in Sect. III.2, page 80. In particular, the observation error covariance matrix (Eq. V.6) is estimated as

$$\mathbf{R} = \frac{1}{B-1}\sum_{k=1}^{B}\left(\mu_k(\tilde{\mathbf{y}}) - \widehat{\mu(\tilde{\mathbf{y}})}\right)\left(\mu_k(\tilde{\mathbf{y}}) - \widehat{\mu(\tilde{\mathbf{y}})}\right)^T, \tag{V.21}$$

where $\{\mu_k(\tilde{\mathbf{y}})\}_{k=1}^{B}$ is the set of B bootstrap replicates of the vector of observations $\tilde{\mathbf{y}}$ after anamorphosis (Eq. V.18), and where $\widehat{\mu(\tilde{\mathbf{y}})} = \frac{1}{B}\sum_{k=1}^{B}\mu_k(\tilde{\mathbf{y}})$ is the ensemble average of the replicates. To limit sampling errors, we use a very large number of bootstrap replicates ($B = 50\,000$). One asset of this approach is that it provides estimates of the cross-covariance terms, which characterize error correlation between sensors, which is often assumed to be zero due to a lack of information (Sousa et al. 2018; Aristodemou et al. 2019; Defforge et al. 2021).

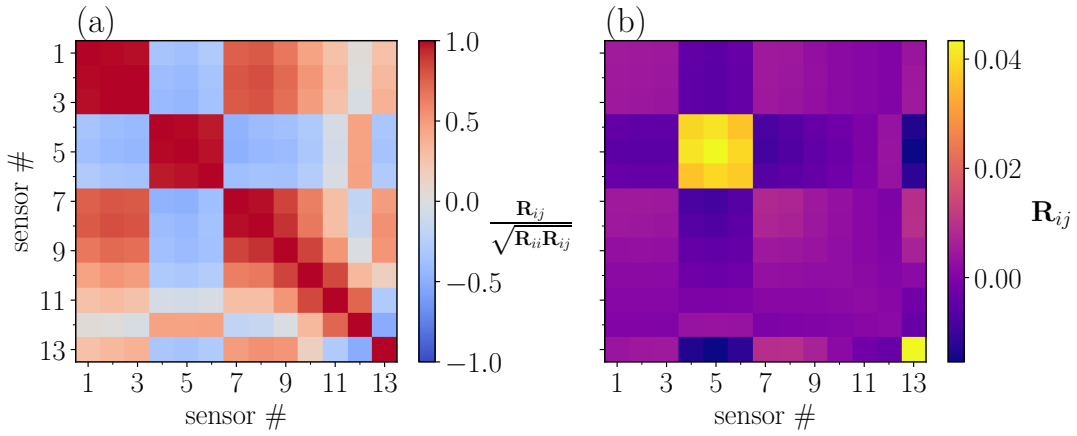


Figure V.4: *Observation error covariance matrix* $\mathbf{R}$ *(Eq. V.6) estimated using bootstrap replicates of the log-transformed concentration measurements (Eq. V.21) with (a) and without normalization (b). Rows and column indices correspond to the sensors from which measurements are assimilated, arranged in the following order: tower C (1, 2, 3 m), tower D (1, 2, 3 m), tower T (1, 2, 4, 6, 8, 10 m), and DPID sensor #26 at* $z = 1.6$ *m.*

Figure V.4 shows the estimated observation error covariance matrix. The distribution of the covariances is realistic with normalized covariances that are unitary on the diagonal and that decrease as the distance between sensors increases (Fig. V.4a). Moreover, Figure V.4 b shows that absolute covariances are particularly high for tower D sensors and the DPID sensor #26. This is explained by their location near the plume edge (see

Fig. V.3), which is especially sensitive to internal variability as demonstrated in Chapter III (see for example Fig. III.8, page 97). In addition, the logarithmic anamorphosis (Eq. V.18) relatively increases the variance of low concentration values also located on the plume edges. Tower D sensors seem to have a low correlation with other sensors, whereas measurements from tower C and T sensors, and to a lesser extent from DPID sensor #26, are significantly correlated. This is because the tower D sensors are located on the edge of the plume, while the others are aligned with the direction of the flow.

V.3.3 Uncertainty in background boundary conditions

This section aims to quantify the uncertainty in the estimation of the background boundary conditions $\boldsymbol{\theta}^b = (\alpha_{inlet}, u_*)^{\mathrm{T}}$ when using measurements of neighboring meteorological stations. We assume that data from tower S, located 30 meters from the area of interest, are not available. The reason for this is twofold: on the one hand, we aim to demonstrate that the proposed DA system can be generalized to a large number of urban configurations that are not necessarily as well instrumented as the MUST field campaign; on the other hand, this allows us to use the measurements from tower S for validation. To define the background parameters and quantify the associated uncertainty, we instead use data measured by the SAMS meteorological stations located a few kilometers around the container array. We have access to 12 full days of acquisition with a temporal resolution varying from 5 to 15 minutes and for the four stations #8, #9, #15, and #16 shown in Fig. II.2a, page 58. For uncertainty quantification, we exclude stations #15 and #16 as we find significant measurement biases compared to the other two stations. This is probably explained by the fact that these stations are closer to the mountains.

Figures V.5a, b, and c show the distribution of the differences in wind direction and friction velocity measurements between SAMS stations #8 and #9. These differences are indicative of the error caused by using measurements from the SAMS stations to define the boundary condition background parameters. Results show significant dispersion of the errors, especially for the wind direction for which the standard deviation reaches 65 °. To validate these estimates, we also take a look at the differences between measurements from the SAMS station #8 and tower S anemometers (d, e, f). This involves fewer samples as only 40 hours of common acquisitions are available. Tower S measurements are averaged over 200-s periods. The resulting distributions are much narrower, especially for the wind direction, whose standard deviation is nearly divided by three. This is because i) the distance between SAMS station #8 and tower S is much smaller than that between stations #8 and #9, ii) the measurements from SAMS station #9 can be influenced by its proximity to more mountainous terrain (Fig. II.2a, page 58), and iii) over the 12 days of acquisition used for the first estimate, there are many time instants for which the wind speed is almost zero and where the wind direction measurement becomes erratic.

Estimated error distributions are approximately centered around zero, which supports the hypothesis that background estimates are unbiased (see Sect. V.2.1). We also recall that, within the EnKF framework, background (and observation) errors are assumed to be normally distributed (see Sect. V.2.2). Using a Kolmogorov-Smirnov test (Massey 1951), we find that the plausibility of such distribution is rejected in every case. Looking at the

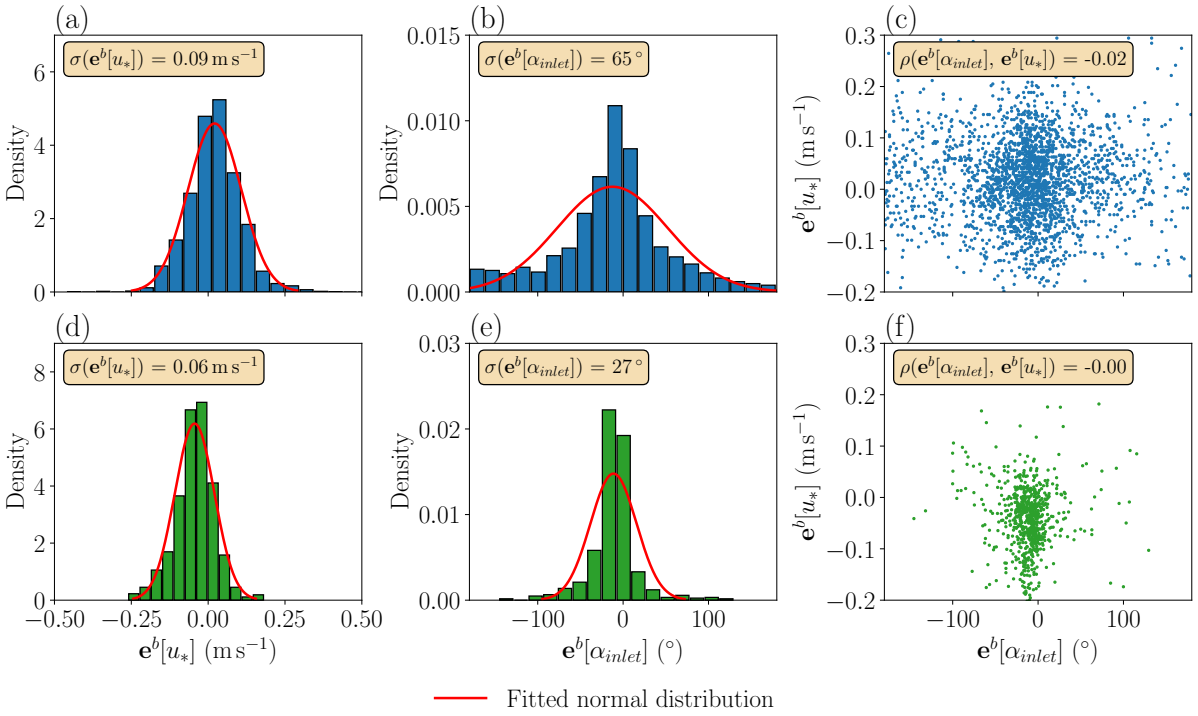


Figure V.5: *Estimation error distributions for the incoming wind friction velocity (a, d) and direction (b, e), and associated scatter plots (c, f). The first row corresponds to the differences between measurements at SAMS stations #8 and #9 over 12 days during the MUST campaign (see Fig. II.2, page 58); and the second row shows the differences between SAMS station #8 and tower S anemometers during the trials at our disposal, giving a total of 40 h of acquisition. The red line corresponds to the fitted normal distribution (obtained by maximum likelihood estimation).*

histograms presented in Fig. V.5, the hypothesis of normally distributed errors is deemed reasonable for friction velocity but not for wind direction. This is a standard problem in circular statistics (Jammalamadaka and SenGupta 2001), which is simply explained by the fact that wind direction is periodic and thus defined over the $[-\pi, \pi]$ interval, while normal distributions are defined on the set of real numbers. Other distributions, such as the von Mises distribution (Mardia and Zemroch 1975), are more appropriate to deal with circular quantity. However, adapting the EnKF (Algorithm 1) to these distributions is outside the scope of the present work. Another possible solution would be to use a particle filter algorithm (Gordon et al. 1993) to account for the real error distributions.

Finally, we define the background error covariance matrix as follows:

$$\mathbf{B} = \begin{pmatrix} \sigma^2_{\alpha_{inlet}} & 0 \\ 0 & \sigma^2_{u_*} \end{pmatrix} \tag{V.22}$$

with $\sigma_{\alpha_{inlet}} = 25\,°$ and $\sigma_{u_*} = 0.09\,\mathrm{m\,s^{-1}}$ the standard deviation of the inlet wind direction and friction velocity background errors, estimated based on the error distributions presented in Fig. V.5. For friction velocity, we choose the worst-case scenario (Fig. V.5a).

This is not feasible for the wind direction because of the definition range of the predictive model, i.e. $[-90\,°, 30\,°]$ (see Sect. V.3.5). We therefore limit the standard deviation of the wind direction background error to $25\,°$, which is consistent with the error estimated using tower S (Fig. V.5e). Nevertheless, we point out that in a real scenario, local measurements such as the ones obtained at tower S would not be available to define the background parameters and therefore could not be used to define the background error covariance matrix. Increasing the operability range of the reduced-order model is a straightforward way to improve this work, but the pollutant plume would not cross the area of interest for such angles. Finally, defining $\mathbf{B}$ as a diagonal matrix is justified because friction velocity and wind direction errors are uncorrelated (Fig. V.5c, f).

V.3.4 Choice of the background parameters

The mean wind direction and friction velocity measured by the SAMS stations during the pollutant release of the MUST trial 2681829 are reported in Table V.2. Except for the station #16, SAMS measurements are really close to our reference best estimates using tower S and ASU anemometer measurements (see Sect. II.4.2.1, page 62). This proximity is explained by the fact that the MUST test site is a flat desert where the ABL can develop homogeneously (Defforge 2019).

However, we aim to test the ability of our DA system to correct larger biases in the wind boundary conditions, more representative of the departure between meteorological stations and local conditions that can occur in a real urban canopy (García-Sánchez et al. 2014; Sousa and Gorlé 2019). We therefore choose to significantly perturb the measurements to define the background parameters used in the DA algorithm. This is also done by Defforge et al. (2021): in this study, a $10\,°$ deviation is applied to the background wind direction. The chosen background parameters $\boldsymbol{\theta}^b = (\alpha^b_{inlet}, u^b_*)$ are reported in Table V.2. The difference between the background friction velocity and the reference value measured by tower S and ASU anemometers is equal to $0.16\,\mathrm{m\,s^{-1}}$, i.e. corresponding to twice the standard deviation of the background error we estimate (Eq. V.22). For wind direction, we impose a deviation $\Delta\alpha_{inlet} = +16\,°$. This is less than one standard deviation but this is comparable to the difference measured by the SAMS station #16, which is heavily influenced by its local topography (see Fig. II.2a, page 58), and this is larger than the deviation used by Defforge et al. (2021).

Table V.2: *Values of the wind direction α_{inlet}, and friction velocity u_* measured near the area of interest by tower S and ASU anemometers and measured further away by the SAMS meteorological stations. Station locations are shown in Fig. II.2, page 58. The last column corresponds to the background parameters $\boldsymbol{\theta}^b$ actually used in the present DA experiments.*

	Tower S & ASU	SAMS #8	SAMS #9	SAMS #15	SAMS #16	$\boldsymbol{\theta}^b$
α_{inlet}	$-41\,°$	$-45\,°$	$-53\,°$	$-37\,°$	$-58\,°$	$-25\,°$
u_*	$0.73\,\mathrm{m\,s^{-1}}$	$0.79\,\mathrm{m\,s^{-1}}$	$0.78\,\mathrm{m\,s^{-1}}$	$0.79\,\mathrm{m\,s^{-1}}$	$0.91\,\mathrm{m\,s^{-1}}$	$0.57\,\mathrm{m\,s^{-1}}$

Since friction velocity is always larger or equal to zero, its distribution cannot be

properly represented by a normal distribution as shown in Fig. V.6. Since this problem is analogous to that of assimilating concentration measurements presented in Sect. V.3.1, we choose to apply an anamorphosis to friction velocity. A log-normal distribution is fitted to all available observations at tower S and SAMS stations #8 and #9 by maximum likelihood estimation (Fig. V.6). Note that the distribution of the tower S measurements is slightly shifted towards higher friction velocity values because very few experiments were carried out in near-zero wind conditions (Yee and Biltoft 2004). We check the likelihood of this distribution using the Kolmogorov-Smirnov test (Massey 1951). We find a p-value > 0.05, indicating that this distribution assumption is not rejected based on the 3 198 samples considered. In the end, we propose to apply the same logarithmic anamorphosis as for concentration (Eq. V.18). We found that the threshold has a relatively limited impact on the DA estimates; we retain the value of $u_t = 0.04\,\mathrm{m\,s^{-1}}$.

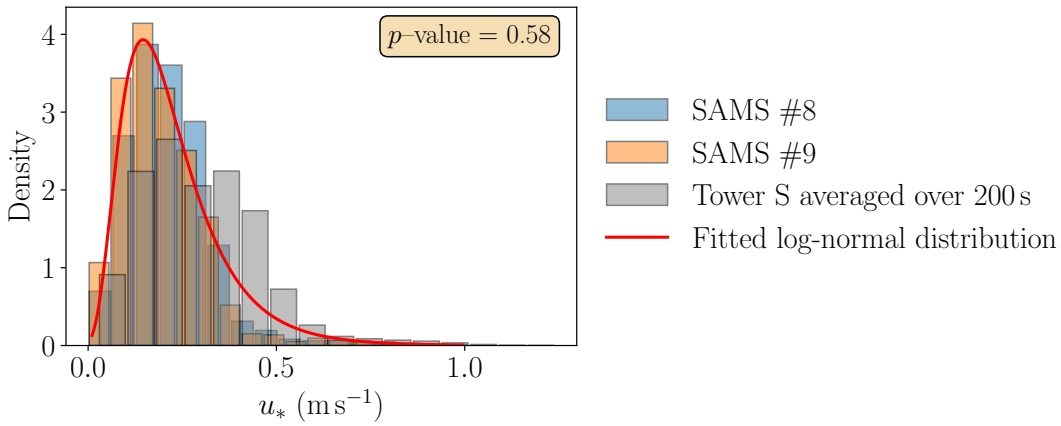


Figure V.6: *Friction velocity distribution estimated at SAMS stations #8 and #9 over 12 days during the MUST field campaign and at tower S, over 200-s periods, during the 40 h of available measurements. The red line corresponds to the fitted log-normal distribution on all available measurements by maximum likelihood estimation.*

This means that the DA algorithm is going to estimate the logarithmic friction velocity $\widetilde{u}_* = \ln(u_* + u_t)$ instead of the friction velocity u_*. The background error covariance matrix $\mathbf{B}$ must therefore be modified accordingly. For this purpose, we use the following formula:

$$\sigma^2 = \ln\left(1 + \frac{\mathbb{V}(X)}{\mathbb{E}(X)^2}\right), \tag{V.23}$$

which links the variance σ^2 of $\ln(X)$ with the variance and expected value of X when X follows a log-normal distribution. In our case, X is the non-transformed friction velocity u_*, and we take for its expected value the chosen background value u_*^b reported in Table V.2. Note that this transformation does not drastically change the background error distribution compared to the one presented in Fig. V.5b, e.

V.3.5 Model definition and error

As motivated in Sect.V.1.2, we use the POD–GPRs reduced-order model built in Chapter IV as a surrogate of the LES dispersion model within the DA system. In our case, this reduces the prediction cost from 20 000 core hours to 30 milliseconds (see Table IV.4, page 160). This opens up the possibility of using a much larger number of members (N_e) in the EnKF in order to reduce sampling errors and to better account for the model response surface.

Concerning the parameters of the POD–GPRs model, we truncate the POD basis at $L = 10$ modes following diagnostics from Chapter IV, and we use the $\mathcal{T}_{log\text{–}1\mathrm{D}}$ pre-processing, which includes logarithmic transformation (Eq. IV.22) and friction velocity normalization (Eq. IV.24). Using these parameters, the POD–GPRs reduced-order model very well reproduces the LES model predictions, except in the direct vicinity of the source (see Sect. IV.6, page 155). The error added in the DA process due to the reduced-order model is therefore limited.

In addition, when using friction velocity normalization, POD–GPRs predictions depend only linearly on friction velocity. By assuming that the similarity theory holds for any friction velocity $u_* \in \mathbb{R}_+^*$, the definition range of the model can be extended to:

$$\Omega_{\boldsymbol{\theta}} = [-90\,^\circ,\ 30\,^\circ] \times \mathbb{R}_+^*. \tag{V.24}$$

This is convenient for ensemble-based DA methods as it allows representing larger uncertainty in the friction velocity. However, one should keep in mind that the POD–GPRs model was not validated for friction velocity outside the $[0.0740,\ 0.8875]\,\mathrm{m\,s^{-1}}$ interval. Extra caution should be taken with predictions for very low friction velocity as the LES model was not designed for near-zero wind conditions, and model reduction errors related to internal variability tend to increase inversely proportional to friction speed (as shown in Sect. IV.6.3).

Accounting for the model error is important as it prevents the DA algorithm from overtrusting model predictions, while they are inherently uncertain as demonstrated in Chapter III. For this purpose, Sousa and Gorlé (2019) perturb model predictions in the EnKF algorithm (Algorithm 1) by adding a Gaussian noise term. In this thesis, we directly take advantage of the fact that POD–GPRs not only predict time-averaged concentration but also their error distribution as shown in Fig. V.7. As demonstrated in Chapter IV, these distributions explain both internal variability and model reduction errors. We therefore define the model prediction function $\mathcal{M}$ as a random realization of the POD–GPRs (several examples are represented by the dashed lines in Fig. V.7) obtained from the posterior distribution of the GPRs used in the reduced-order model (Eqs. IV.15–IV.16, page 127) instead of the mean prediction corresponding to the solid line in Fig. V.7.

This approach allows for a more realistic representation of the model error distributions compared to the approach of Sousa and Gorlé (2019), especially for low concentration values, as POD–GPRs predictions follow log-normal distributions. More importantly, our method directly accounts for the spatial correlation in the model errors. Finally, it is worth noting that the model randomness adds variability in the EnKF analysis step

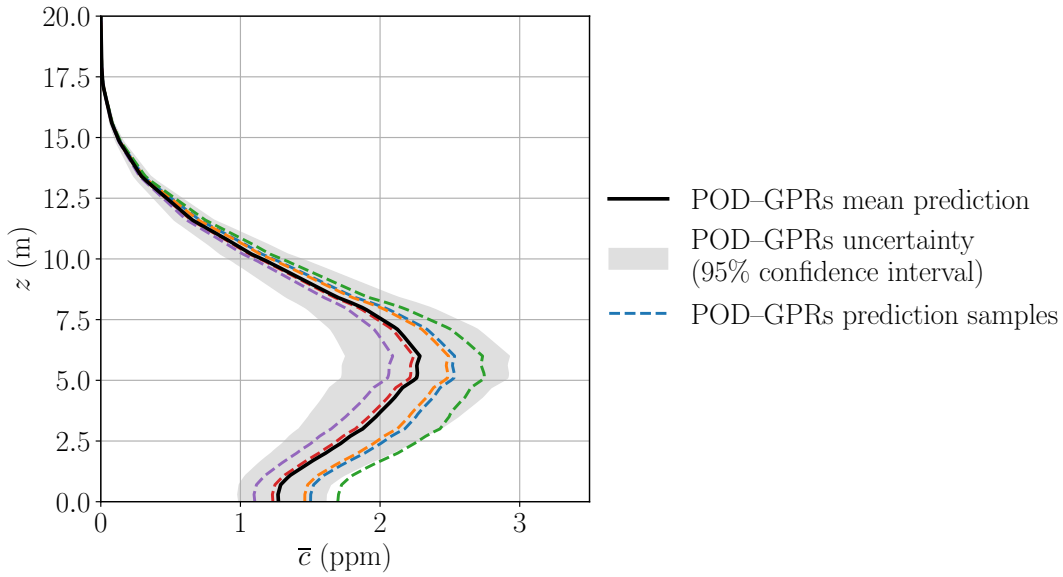


Figure V.7: *Vertical profile of the mean concentration at tower B estimated by the POD–GPRs reduced-order model with $\mathcal{T}_{log-1D}$ preprocessing. The mean estimate is represented by the black solid line, while the gray shaded area corresponds to the associated 95% confidence interval estimated using the procedure detailed in Sect. IV.2.5, page 131. Dashed lines correspond to random samples of POD–GPRs predictions, obtained from the posterior distributions of the Gaussian processes.*

(Algorithm 1), which can help to avoid ensemble collapse. This last point is particularly of interest from an uncertainty quantification perspective, as it prevents the error associated with the final DA estimation from being underestimated.

V.4 Validation and calibration of the data assimilation system

This section presents a first step in the validation and calibration of the EnKF algorithm used to correct and reduce the uncertainty on the LES wind boundary condition parameters. This is done using the twin experiment framework properly defined in Sect. V.4.1, in which synthetic observations are assimilated instead of the real measurements. The operability of the DA system is thus assessed in Sect. V.4.2. Next, we further investigate the effect of the sampling error inherent to the EnKF algorithm in Sect. V.4.3, which enables us to calibrate the optimal ensemble size for the EnKF.

V.4.1 Twin experiment principle

In twin experiments, also known as observing system simulation experiments (OSSEs) (Arnold and Dey 1986; Hoffman and Atlas 2016), the true control vector $\boldsymbol{\theta}^t$ and system state $\mathbf{x}^t$ are assumed to be known. Synthetic observations are then generated from $\mathbf{x}^t$ using the observation operator and adding a random noise representative of the observation error:

$$\mathbf{y}^o = \mathcal{H}(\mathbf{x}^t) + \epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{R}). \tag{V.25}$$

In this thesis, we use the baseline mean concentration field LES prediction presented in Chapters II and III as the true state. The associated true control vector is thus defined as our best guess of the boundary conditions using all measurements available: i.e. $\boldsymbol{\theta}^t = (-41\,^\circ, 0.73\,\mathrm{m\,s^{-1}})^\mathrm{T}$. A schematic of the twin experiment framework is shown in Fig. V.8.

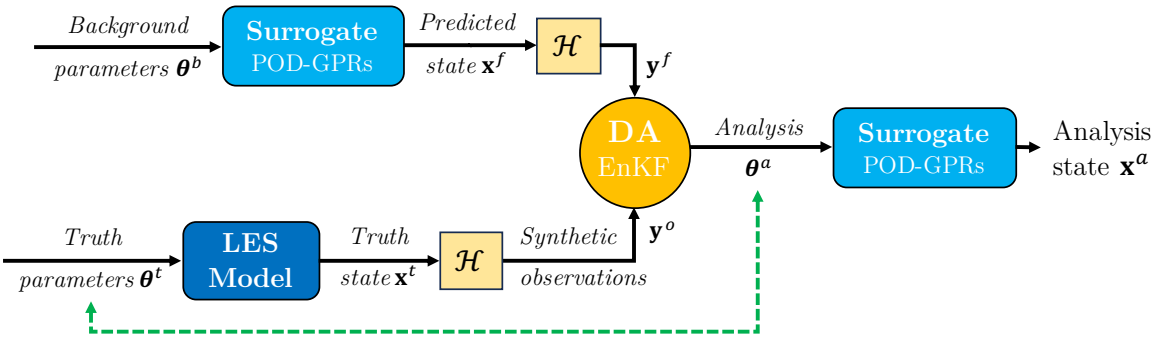


Figure V.8: *Principle of the twin experiment setup used to set up and validate the reduced-cost DA system designed in this thesis. As represented by the green dashed line, the objective of the experiment is to assess the ability of the DA system to correctly infer the true input parameters $\boldsymbol{\theta}_t$ by assimilating synthetic observations $\mathbf{y}_o$ (Eq. V.25) obtained from the reference LES model prediction. The POD-GPRs reduced-order built in Chapter IV is used as a surrogate for the LES model within the EnKF (Algorithm 1).*

Twin experiments provide an ideal testing framework by giving access to the true system state and control vector, which are never known in real applications, making

it possible to quantify the analysis accuracy by comparing $\boldsymbol{\theta}^a$ and $\boldsymbol{\theta}^t$, as well as the resulting state prediction accuracy by comparing $\mathbf{x}^a$ and $\mathbf{x}^t$ (Fig. V.8). In addition, twin experiments enable the control of the model error, which is difficult to estimate accurately in real applications, thus enabling optimal analysis. Indeed, generating the true state using the same model as in the DA scheme ensures the existence of an exact solution to the inverse problem. With the framework adopted in this section (Fig. V.8), we make the inverse problem slightly more difficult by generating the true state with the LES model $\mathcal{M}_{\text{LES}}$ instead of the POD–GPR model $\mathcal{M}$ used in the DA scheme, which gives rise to a model error η:

$$\begin{aligned} \mathcal{M}(\boldsymbol{\theta}^t) &= \mathbf{x}^t + \eta, \\ &= \mathcal{M}_{\text{LES}}(\boldsymbol{\theta}^t) + \eta, \end{aligned} \tag{V.26}$$

corresponding to the model reduction error. This error is not zero since $\boldsymbol{\theta}^t$ does not belong to the POD–GPRs training database, but is very limited as demonstrated when validating the POD–GPRs in Chapter IV. Following this principle, more complex twin experiment frameworks can be developed by using different models to generate the true state to progressively complexify the DA estimation task (Halliwell et al. 2014).

Table V.3 summarizes the setup of the baseline twin experiment used to evaluate our reduced-cost DA system. The definition of background and observational data are further detailed in Sect. V.3. The ensemble size used in the EnKF is set to 500, and its influence on the DA estimation is further investigated in Sect. V.4.3.

Table V.3: *Set-up of the twin experiment for application to the MUST 2681829 trial. The first row corresponds to the true parameters used in the twin experiments. The background parameters and observations as well are their associated uncertainties are defined in detail in Sect. V.3.*

	Notation	Setup
Truth parameters	$\boldsymbol{\theta}^t = (\alpha^t_{inlet}, u^t_*)$	$(-41\,^\circ, 0.73\,\mathrm{m\,s^{-1}})$
Background parameters	$\boldsymbol{\theta}^b = \left(\alpha^b_{inlet}, u^b_*\right)$	$(-25\,^\circ, 0.57\,\mathrm{m\,s^{-1}})$
Background errors	$\mathbf{B} = \begin{pmatrix} \sigma^2_{\alpha_{inlet}} & 0 \\ 0 & \sigma^2_{u_*} \end{pmatrix}$	with $\sigma_{\alpha_{inlet}} = 25\,^\circ$, $\sigma_{u_*} = 0.09\,\mathrm{m\,s^{-1}}$
Observation network	$\mathbf{y}$	13 observations of concentration at towers C (1, 2, 3 m), D (1, 2, 3 m) T (1, 2, 4, 6, 8, 10 m) and DPID #26
Observation error	$\mathbf{R}$	See Sect. V.3.2
EnKF ensemble size	N_e	500
Anamorphosis threshold	(y_t, u_t)	$(0.04\,\mathrm{ppm}, 0.04\,\mathrm{m\,s^{-1}})$

Note that the observation error covariance matrix $\mathbf{R}$ is re-estimated specifically for the twin experiments according to the procedure presented in Sect. V.3.2, in order to

take into account the actual aleatory uncertainty of the synthetic observations predicted by the LES, which is not equal to that of the field measurements. In particular, since the LES model underestimates the effect of the internal variability of the ABL as shown in Fig. III.5, page 86, the resulting uncertainty on observations is lower in twin experiments compared to when assimilating the real measurements.

V.4.2 Results of the baseline twin experiment

In this section, we give the estimation results obtained with the baseline twin experiments defined in Table V.3. We first evaluate the ability of the EnKF (Algorithm 1) to infer the control vector $\boldsymbol{\theta} = (\alpha_{inlet}, u_*)^{\mathrm{T}}$, since this is what is estimated in the first place (Fig. V.8).

Table V.4: *Control vector estimations in the baseline twin experiment, for the background (before DA), and the analysis (after DA). The two latter columns correspond to the parameter standard deviations. Background error standard deviations are estimated in advance in Sect. V.3.3. Analysis error standard deviations are estimated from the analysis ensemble (Eq. V.17). The first row corresponds to the true parameters used in the LES model prediction from which the synthetic observations are obtained for this twin experiment.*

	α_{inlet}	u_*	$\sigma(\alpha_{inlet})$	$\sigma(u_*)$
Truth	-41°	$0.73\,\mathrm{m\,s^{-1}}$	/	/
Background	-25 °	$0.57\,\mathrm{m\,s^{-1}}$	25 °	$0.09\,\mathrm{m\,s^{-1}}$
Analysis	-33 °	$0.62\,\mathrm{m\,s^{-1}}$	7.3 °	$0.11\,\mathrm{m\,s^{-1}}$

Table V.4 validates that the EnKF is working reliably, since i) it reduces the bias in the control vector estimation compared to the background, and ii) it provides realistic uncertainty predictions that explain well the estimation errors, as $|\alpha^a_{inlet} - \alpha^t_{inlet}| < 2\sigma(\alpha^a_{inlet})$ and $|u^a_* - u^t_*| < 2\sigma(u^a_*)$. We also note that DA reduces the uncertainty on the wind direction estimation when compared to the background uncertainty, but not for the friction velocity. Indeed, the uncertainty on u_* even slightly increases after the analysis, meaning that the DA algorithm is not able to improve the estimation of u_* with the available noisy observations and given the model prediction uncertainty defined in Sect. V.3.5. This is confirmed by the distribution of the EnKF ensemble in the control space (Fig. V.9), whose spread according to α_{inlet} is greatly reduced after the analysis, while the spread according to u_* is unchanged. This distribution is explained by the fact that the concentration observations are more sensitive to variations in the wind direction, which induce large plume position errors than to friction velocity, which has a linear effect on the concentration.

In the EnKF framework, the best estimate of the control vector is defined as the mean of the analysis (Eq. V.16). Using this estimate as input for the reduced-order model provides a new concentration field estimation corresponding to the corrected wind boundary conditions (Fig. V.8). As the true state $\mathbf{x}^t$ is entirely known in the twin experiment framework, we can assess in detail the accuracy of the state estimated by DA.

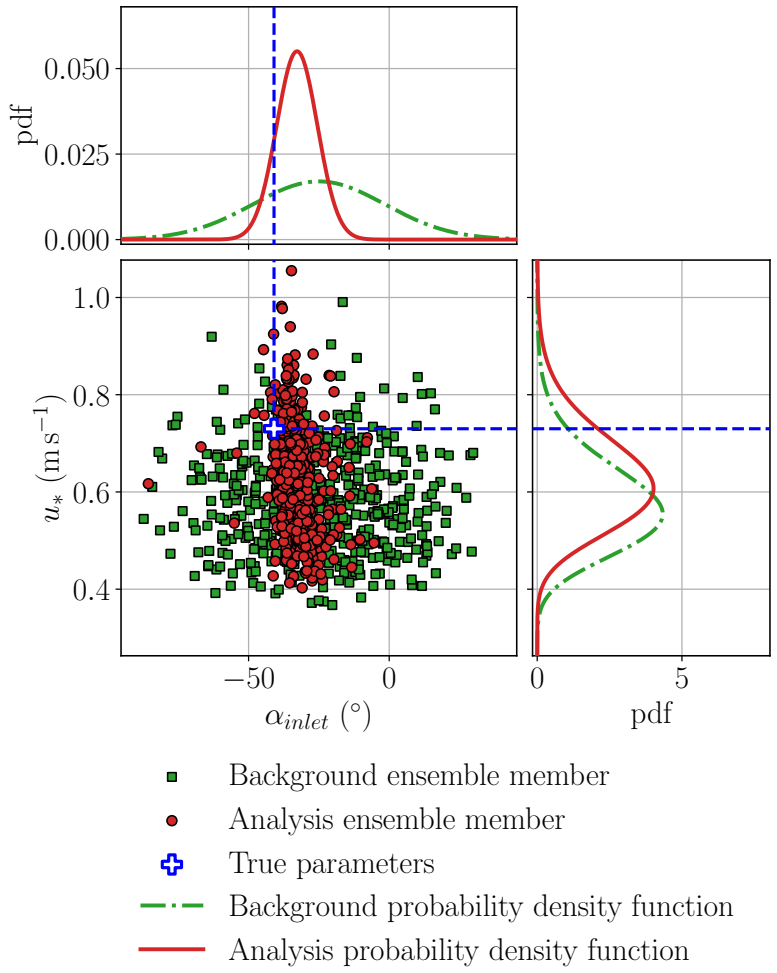


Figure V.9: *Background and analysis EnKF ensembles obtained in the baseline twin experiment. Background and analysis ensemble members are represented as green squares and red circles respectively. The associated normal and log-normal probability density functions (pdf) of the inlet wind direction and of the friction velocity are represented in the upper and right panels respectively. The blue plus symbol indicates the true control vector* $(\alpha^t_{inlet}, u^t_*) = (-41\,°, 0.73\,m\,s^{-1})$.

Figure V.10 compares two isolines (1 ppm and 0.01 ppm) for the background, analysis, and true mean concentration fields, at $z = 1.6$ m. As expected, the gain in accuracy in the control space propagates well to the state space, and the analysis estimation is more in line with the true concentration field. In particular, we see that the analysis plume centerline is better aligned with the true plume centerline. Moreover, the plume spread, which was overestimated by the background, is well reduced especially at higher concentrations. The accuracy of the analysis for the 1-ppm iso-concentration illustrates in part why the DA system does not further correct for friction velocity (Fig. V.10a).

To better assess the accuracy of the mean concentration DA estimations, we compute the air quality metrics defined in Sect. III.3.2, page 90. Metric scores calculated for the

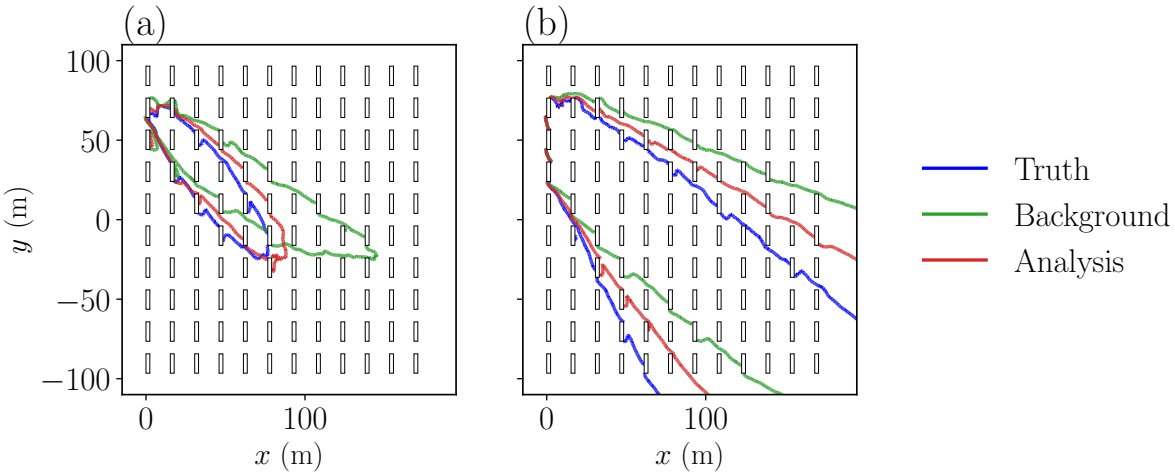


Figure V.10: *Mean concentration field estimations at $z = 1.6\,m$ for two iso-concentration levels: 1 ppm (a) and 0.01 ppm (b). The isolines corresponding to the background and analysis predictions are represented in green and red respectively, while the blue isolines represent the true concentration field predicted by LES.*

complete fields are given in Table V.5. All the metrics considered show significant improvement in the accuracy of the mean concentration field estimation, thereby indicating that the estimation is enhanced both for the far field and the near field. Note that the analysis estimation of the mean concentration remains globally overestimated as FB < 0 and MG < 1. This is simply explained by the underestimation of the friction velocity (Table V.4). Overall, the accuracy of the mean concentration field prediction can still be improved as Table V.5 shows that there is still a significant gap between the validation scores obtained at the analysis step and those considering only errors related to internal variability, which represents the best scores that could be achieved in principle. This is consistent with the fact that the DA system does not perfectly correct the control vector.

Table V.5: *Mean concentration prediction errors for the background and analysis when compared to the true concentration field. Definitions of the validation metrics are given in Sect. III.3.2, page 90. Perfect scores of the different metrics are recalled in the second row. The third row corresponds to the mean level of error solely due to internal variability estimated using the bootstrap propagation method explained in Sect. IV.3.4, page 135.*

	FB	NMSE	FAC2	MG	VG	FMS (1ppm)	FMS (0.01ppm)
Perfect score	0	0	1	1	1	1	1
Internal variability	0	0.42	0.96	1.00	1.11	0.87	0.95
Background	-0.41	26.8	0.53	0.71	6.0×10^3	0.28	0.51
Analysis	-0.29	7.55	0.65	0.77	12.6	0.66	0.74

Computational cost of the DA system is shown to be very limited with a total runtime of 50 s on average using a single core of an Intel Ice Lake processor. Almost all of this time corresponds to the $N_e + 1$ reduced-order model queries. The prediction time of the reduced model is a little slower than the one announced in Table IV.4, page 160, because of implementation reasons as i) we do not use the mean POD–GPRs prediction but a random realization (see Fig. V.7), and ii) we have not optimized the call to the reduced-order model within the EnKF algorithm. The full cost of the DA system therefore scales linearly with N_e, which allows the use of even larger ensemble sizes. Moreover, the DA system could be significantly accelerated as the reduced-order model calls are independent and thus parallelizable (Algorithm 1). Yet, we did not push optimization any further in this thesis, as the return time is already very satisfactory. To conclude, we have achieved one of the major objectives of this thesis, namely the construction of a fast modeling system that could be suitable for real-time applications.

V.4.3 Effect of the ensemble size and sampling error estimation

In this section, we jointly address the issue of the choice of EnKF ensemble size and the effect of sampling errors on the performance of the proposed DA system. This is done by repeating the twin experiment described in Table V.3, varying N_e and varying background ensemble sampling.

In particular, we perform multiple independent realizations of the EnKF by varying the random seed for the background ensemble sampling in Algorithm 1, in order to estimate the analysis error statistics using a Monte-Carlo approach. Tests show that between 100 and 200 realizations are enough to achieve convergence of the Monte-Carlo estimates. Figure V.11 shows the averaged DA estimation error, along with the associated 95% confidence interval estimated by Monte Carlo. Both estimations of the inlet wind direction and friction velocity quickly converge with the ensemble size, with almost no change in the mean prediction from $N_e = 50$ onwards. However, we find that background sampling error in the EnKF induces a significant variability in the control vector estimation. This variability is attenuated slowly as N_e increases, presumably in $\mathcal{O}(1/\sqrt{N_e})$. With an ensemble of $N_e = 500$ members, we estimate that the sampling error results in a relative standard error of approximately 2.5% for both the wind direction and friction velocity estimations. This error is deemed reasonable given the other forms of uncertainty involved and is more than covered by the analysis uncertainty estimation (Table. V.4). We therefore estimate that using $N_e = 500$ members provides a good compromise between sampling error and computational cost. For the results previously presented in Sect. V.4.2, we chose a background sampling seed that yields performances closed to the averaged DA performances as shown by the yellow stars in Fig. V.11.

Similar trends are found for the effect of ensemble size and sampling noise on the analysis ensemble variances, i.e. $\sigma(\alpha^a_{inlet})$ and $\sigma(u^a_*)$. However, as a second-order statistic, it achieves convergence for a higher ensemble size, between 250 and 500 members.

Note that we have also investigated the effect of the other independent sampling seeds involved in:

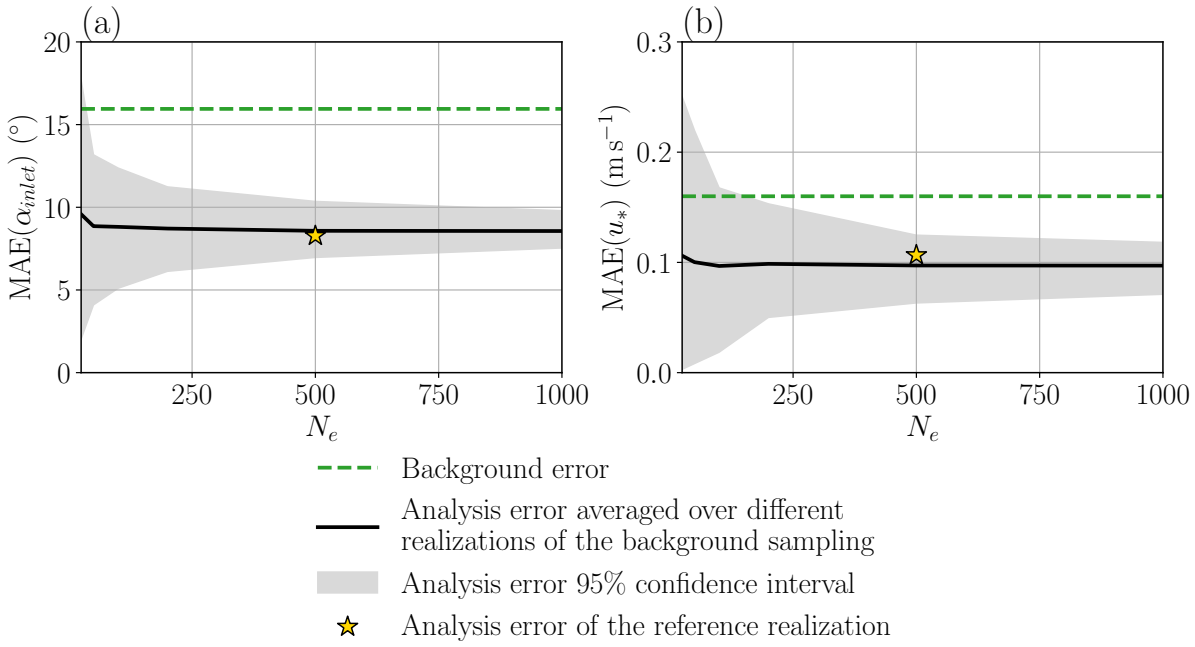


Figure V.11: *Mean absolute errors of the DA estimation of inlet wind direction (a) and friction velocity (b), for varying EnKF ensemble size N_e. The black solid line corresponds to the averaged analysis error, estimated by a Monte-Carlo procedure with between 100 and 200 realizations of the EnKF (Algorithm 1) corresponding to different random seeds for the background ensemble sampling. The gray shaded area represents the 95% confidence interval of the analysis error thus obtained. The baseline background ensemble sampling used in Sect. V.4.2 is depicted as a yellow star.*

- the generation of the synthetic observations (Eq. V.25) that are assimilated within the twin experiment framework,
- the perturbation of the observations in the stochastic EnKF (Algorithm 1),
- the random POD–GPRs realizations used to account for model error as explained in Sect. V.3.5,

and it turns out that they all have less impact on DA performance than the background sampling seed.

Further analysis is required to provide insights into the variability of DA estimates due to the choice of the background sampling seed but this is beyond the scope of the present work.

V.5 Assimilation of the real field measurements

In this section, we apply the reduced-cost DA system designed in this thesis (Fig. V.2) to estimate the wind boundary condition parameters using the actual measurements from the MUST trial 2681829. Section V.5.1 presents the results obtained, with the aim of assessing the ability of the DA system to improve the accuracy of the LES model in a case where the available meteorological data are incomplete or highly uncertain. An additional test is then carried out to investigate the sensitivity of the DA system to the choice of threshold in the concentration anamorphosis.

V.5.1 Results for the MUST experimental data assimilation

This section presents the nominal results obtained by the DA system when using the actual MUST field measurements. Outside of the measurements assimilated and their associated uncertainties, which are defined in Sect. V.3.1 and V.3.2, the rest of the set-up is the same as for the baseline twin experiment defined in Table V.3. From the findings of Sect. V.4.3, an ensemble size of $N_e = 500$ is used in the EnKF. The analysis of the results focuses on the assessment of:

1. the ability of the EnKF to estimate the wind boundary conditions when assimilating real concentration measurements,

2. the effect of the control vector correction on the mean concentration prediction.

Estimation of the wind condition parameters

As opposed to twin experiments, when applying the DA system to a real case with actual measurements we do not know the true control vector anymore. Nevertheless, we have a fairly representative estimate of wind boundary condition parameters thanks to measurements from the tower S and ASU anemometers, that we deliberately did not use in the DA system and background definition. The resulting reference control vector is thus the same as the true control vector used in the twin experiments, i.e. $\boldsymbol{\theta}^{(ref)} = (-41\,^\circ, 0.73\,\mathrm{m\,s^{-1}})^\mathrm{T}$.

Figure V.12 shows the distribution in control space of the background and analysis ensembles of the EnKF. The reference estimation of the control vector is also represented by the yellow cross. The EnKF is not able to infer the friction velocity from the real measurements, as it does not reduce either the background bias compared to $u_*^{(ref)}$ or the background uncertainty. This is explained by the lack of sensitivity of the observations to the friction velocity, which makes the inverse problem difficult to solve as already shown in the quasi-ideal twin experiment framework (Fig. V.9). Concerning the inlet wind direction, the EnKF tends to over-correct the inlet wind direction compared to the reference measurements. Nevertheless, it achieves reducing the uncertainty compared to background information, while still explaining the analysis error with respect to the reference value $\alpha_{inlet}^{(ref)}$. These results thus confirm the reliability of the proposed DA system to estimate the wind direction, but also its inability to correct the friction velocity.

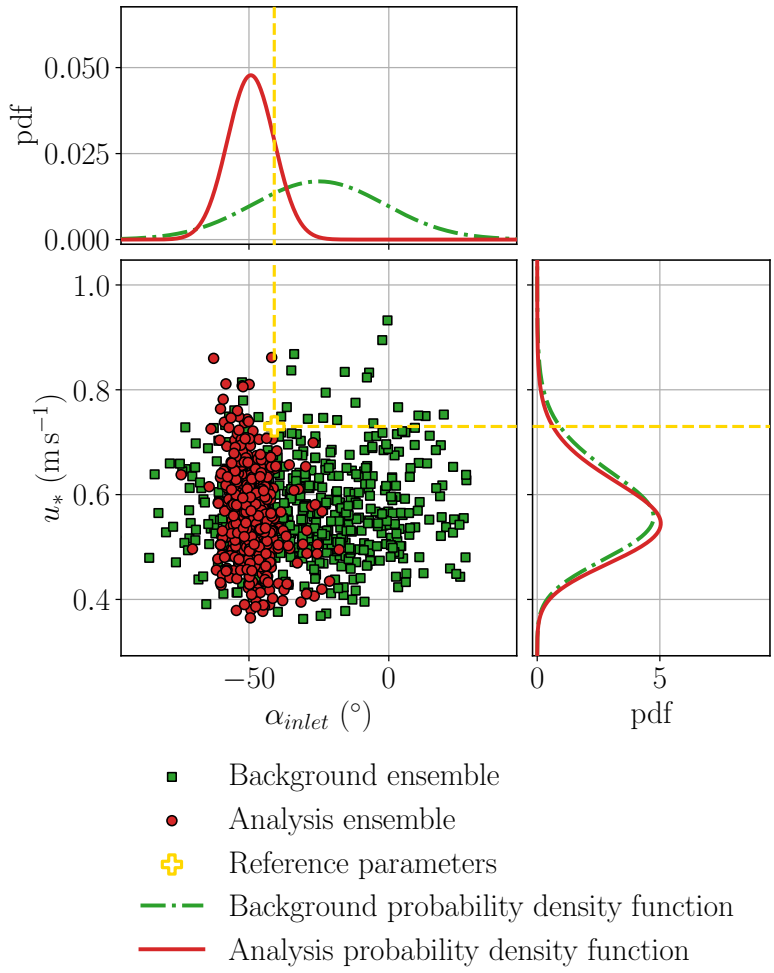


Figure V.12: *Background and analysis EnKF ensembles obtained assimilating the MUST field measurements. Background and analysis ensemble members are represented as green squares and red circles respectively. The associated normal and log-normal probability density functions (pdf) of the inlet wind direction and of the friction velocity are represented in the upper and right panels respectively. The yellow plus symbol indicates the reference control vector* $(\alpha_{inlet}^{(ref)}, u_*^{(ref)}) = (-41\,°, 0.73\,m\,s^{-1})$ *based on measurements from tower S and ASU anemometers.*

We also found that DA estimations are sensitive to background sampling error (not shown) and that the resulting analysis error is of the same order of magnitude as in the twin experiments (Sect. V.4.3). The results presented in this section are obtained for a background sampling seed that yields DA performances representative of the averaged accuracy estimated by Monte Carlo.

Effect on the mean concentration prediction

Using the corrected control vector as input to the surrogate model (Fig. V.2), we can update the background estimate of the mean concentration field to determine whether or not the analysis improves accuracy over field measurements. Table V.6 compares the air

Table V.6: *Air quality metrics (Sect. III.3.2, page 90) quantifying the global agreement between the background and analysis mean concentration prediction and experimental measurements. Metrics are computed for all available observations and for the validation observations subset, i.e. those that are not assimilated (Fig. V.3). Sensors for which the experimental mean concentration is under the detection threshold are excluded. Perfect scores of the different metrics are recalled in the second row.*

		FB	NMSE	FAC2	MG	VG
	Perfect score	0	0	1	1	1
All obs.	Background	-0.15	1.66	$^{14}/_{47} = 29.8\,\%$	1.80	18.9
	Analysis	0.42	0.60	$^{22}/_{47} = 46.8\,\%$	1.94	4.61
Validation obs.	Background	-0.11	1.80	$^{6}/_{34} = 17.7\,\%$	2.03	32.5
	Analysis	0.32	0.39	$^{22}/_{34} = 64.71\,\%$	1.88	3.92

quality metrics scores obtained with the background and analysis estimates. Metrics are computed over either the subset of validation observations, i.e. that are not assimilated by the DA system (see their location in Fig. V.3), or over every observation available. The results show that the EnKF succeeds in improving mean concentration estimation, with an improvement in all scores except FB. The value of fractional bias obtained by the analysis (FB>0) indicates that it overall underestimates the mean concentration. Since the analysis underestimates the friction velocity (Fig. V.12), which should lead to an overestimation of the mean concentration, it is more likely due to a plume position error. The fact that the analysis not only improves accuracy at the assimilated sensor location but also at the validation sensor level, is explained by the choice of boundary condition parameter correction, which ensures a physically consistent correction of the average concentration. This demonstrates the ability of our DA system to extrapolate information from local measurements. Note that DA performance is even slightly better for sensors that are not assimilated. However, the difference in performance is not significant, given the limited number of sensors considered and the fact that validation measurements are also subject to internal variability.

To better understand the effect of the boundary condition parameter correction on the mean concentration estimation, we show the resulting mean concentration horizontal and vertical profiles in Figs. V.13 and V.14. The profiles shown correspond to the predictions obtained with the ensemble-averaged control vectors for background (Table V.3) and analysis (Eq. V.16). The shaded areas represent the 95% confidence interval resulting from the POD–GPRs model error, which covers both the model reduction error and the internal variability error predicted by LES. Note that the variance resulting from the uncertainty on the control vector (Fig. V.12) is not shown in Figs.V.13 and V.14. The overcorrection of wind direction by the EnKF directly affects the plume, as illustrated by the increasing shift in the position of the plume centerline (Fig. V.13) and by the concentration overestimation at tower D (Fig. V.14c). While this deviation of the plume explains quite well the discrepancy with DPID measurements on the horizontal, the

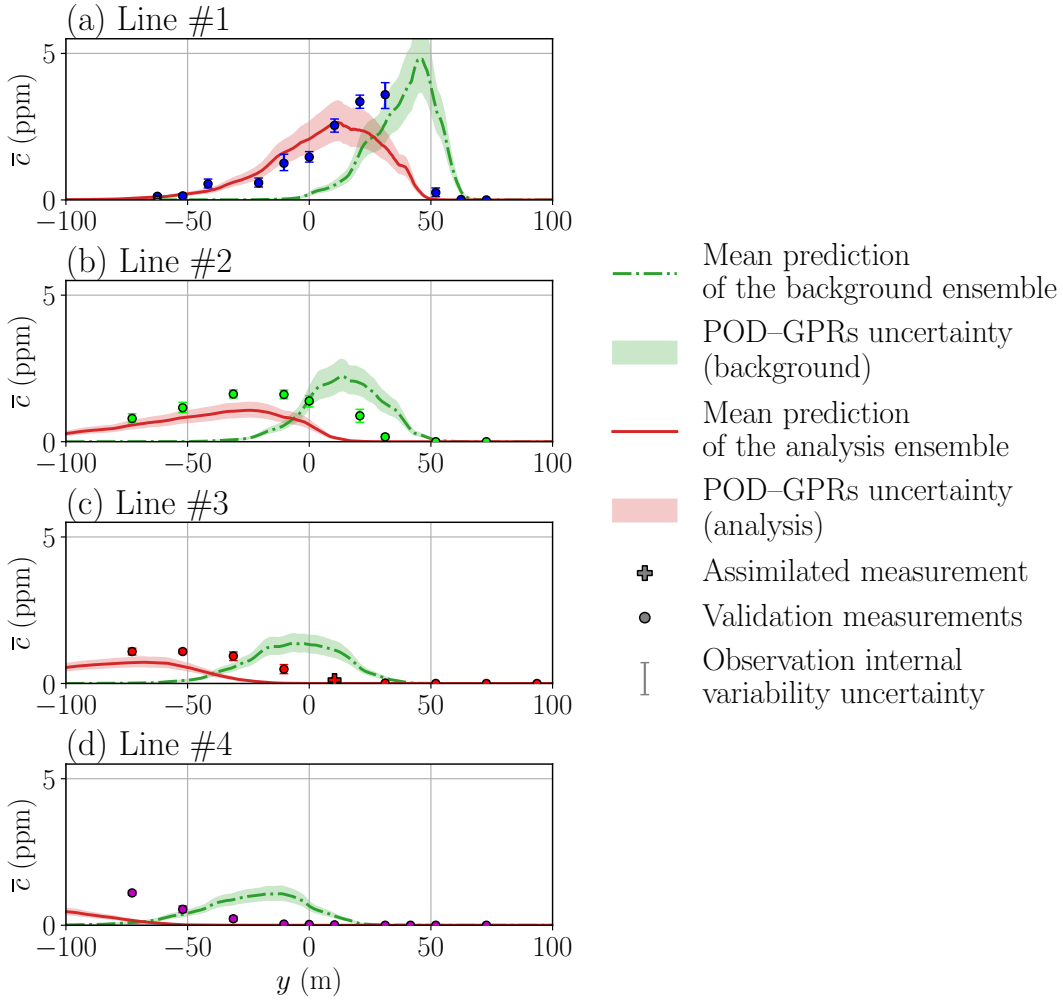


Figure V.13: *Horizontal profiles of mean concentration $\bar{c}$ predicted by the POD–GPRs model for the ensemble-averaged background and analysis control vectors in red solid and green dashed-dotted lines respectively. Each row corresponds to a line of DPID sensors represented with a distinct color in Fig. II.3, page 59. Cross and circle symbols respectively correspond to the assimilated and validation measurements. Shaded areas represent the 95% confidence interval associated with the POD–GPRs prediction (see Sect. IV.2.5, page 131), while error bars correspond to the measurement uncertainty induced by the internal variability of the ABL and estimated by stationary bootstrap (see Sect. III.2, page 80).*

results are more ambiguous for the vertical profiles as the analysis significantly improves the prediction at tower B, but degrades it at towers T and C (Fig. V.14). This result is counter-intuitive, since we assimilate measurements from towers T, C and D but not from tower B. We explain it because i) measurements from tower D carry too much weight in the analysis, as shown in Sect. V.5.2; and ii) the lower sensitivity of the model to friction speed compared to wind direction, which leads the DA system to overcorrect wind direction to compensate for errors that are in fact due to the underestimation of friction speed.

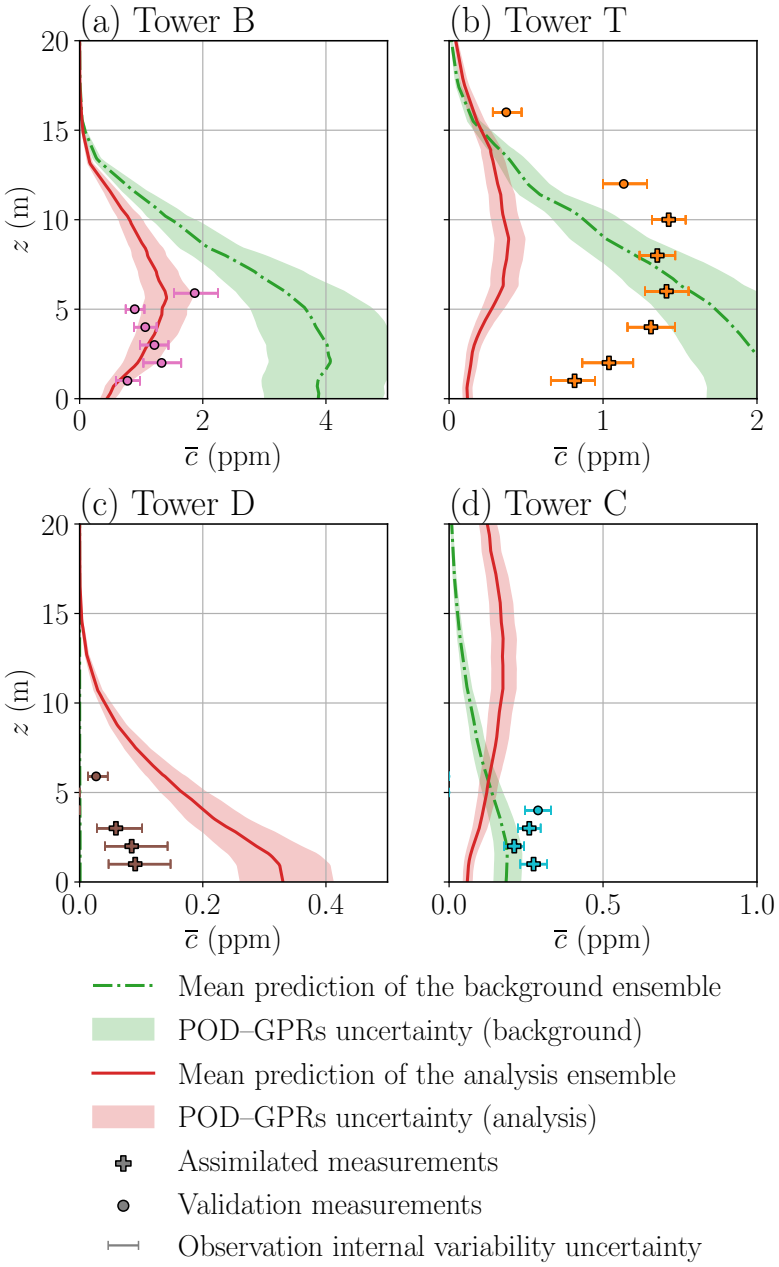


Figure V.14: *Vertical profiles of mean concentration $\overline{c}$ predicted by the POD–GPRs model for the ensemble-averaged background and analysis control vector in red solid and green dashed-dotted lines respectively. Results are given at towers B, T, C and D locations (see Fig. V.3). Cross and circle symbols respectively correspond to the assimilated and validation measurements. Shaded areas represent the 95% confidence interval associated with the POD–GPRs prediction (see Sect. IV.2.5, page 131), while error bars correspond to the measurement uncertainty induced by the internal variability of the ABL and estimated by stationary bootstrap (see Sect. III.2, page 80).*

It is worth mentioning that even using the best possible estimate of the boundary condition parameters $\boldsymbol{\theta}^{(ref)}$, the profiles predicted by the LES model are not consistent between towers B, T and C, as illustrated in Fig. III.8, page 97. This may be due to a misrepresentation of the plume spread and elevation, as well as potential inconsistency between DPID and UVIC sensor measurements, as discussed in Sect. III.4.2.1, page 94. The current experiment shows that our DA system fails to infer a value for the control vector that corrects LES model inconsistencies. This limitation of the system is inherent in the choice of correcting only the meteorological boundary parameters, which prevents the correction of model biases explained by other sources of uncertainty. In particular, it does not address the model's structural uncertainty and thereby cannot correct biases in the state space.

Concerning the computational cost, assimilating real measurements instead of synthetic observations does not change anything. The total execution time is therefore the same as for the baseline twin experiment (Sect. V.4.2).

V.5.2 Effect of the concentration anamorphosis threshold

Figure V.15 gives an overview of the estimation accuracy at sensor location before and after the analysis, for two different concentration threshold y_t used for the concentration anamorphosis (Eq. V.18). In the baseline experiment presented in Sect. V.5.1, i.e. with $y_t = 0.04\,\text{ppm}$ chosen as the maximal sensor threshold in the experiment, the EnKF significantly improves the estimation accuracy on DPIDs line #1 as well as on towers B and D. The scatter plots presented in Figs. V.15a and b further demonstratres that the DA system i) improves overall estimation accuracy with a Pearson correlation coefficient ρ between predictions and observations increasing from 0.29 to 0.78, and ii) is able to improve the accuracy on sensors that are not assimilated thanks to the choice of correcting large-scale boundary condition parameters that affect the whole field.

However, we find that the DA system also deteriorates the prediction at certain locations, such as towers T and C, as already shown in Fig. V.14. We point out that a logarithmic scale is used in Fig. V.15, which gives a good idea of the differences seen by the EnKF analysis, given that a log-transformation is applied during the concentration anamorphosis (Eq. V.18). Thus, in the view of the DA algorithm, the largest discrepancy in the background estimation concerns the low concentrations at tower D, which thereby gives these observations a greater weight in the analysis. As a result, the EnKF tends to overcorrect the wind direction (Fig. V.12), which deteriorates the estimation at other towers T and C. Note that the discrepancy obtained for these relatively high concentrations is considered to be small by the DA system given the log-transformation used.

This higher sensitivity to differences in low concentration has led us to try modifying the concentration anamorphosis by changing the threshold y_t used. Indeed, for an equal relative deviation between prediction and observation, increasing y_t reduces the weight of differences in low concentrations in the analysis. Figure V.15c shows that with $y_t = 0.1\,\text{ppm}$ the improvement in prediction obtained after analysis is much more homogeneous than with the reference value of $0.04\,\text{ppm}$, despite a correlation coefficient of the same

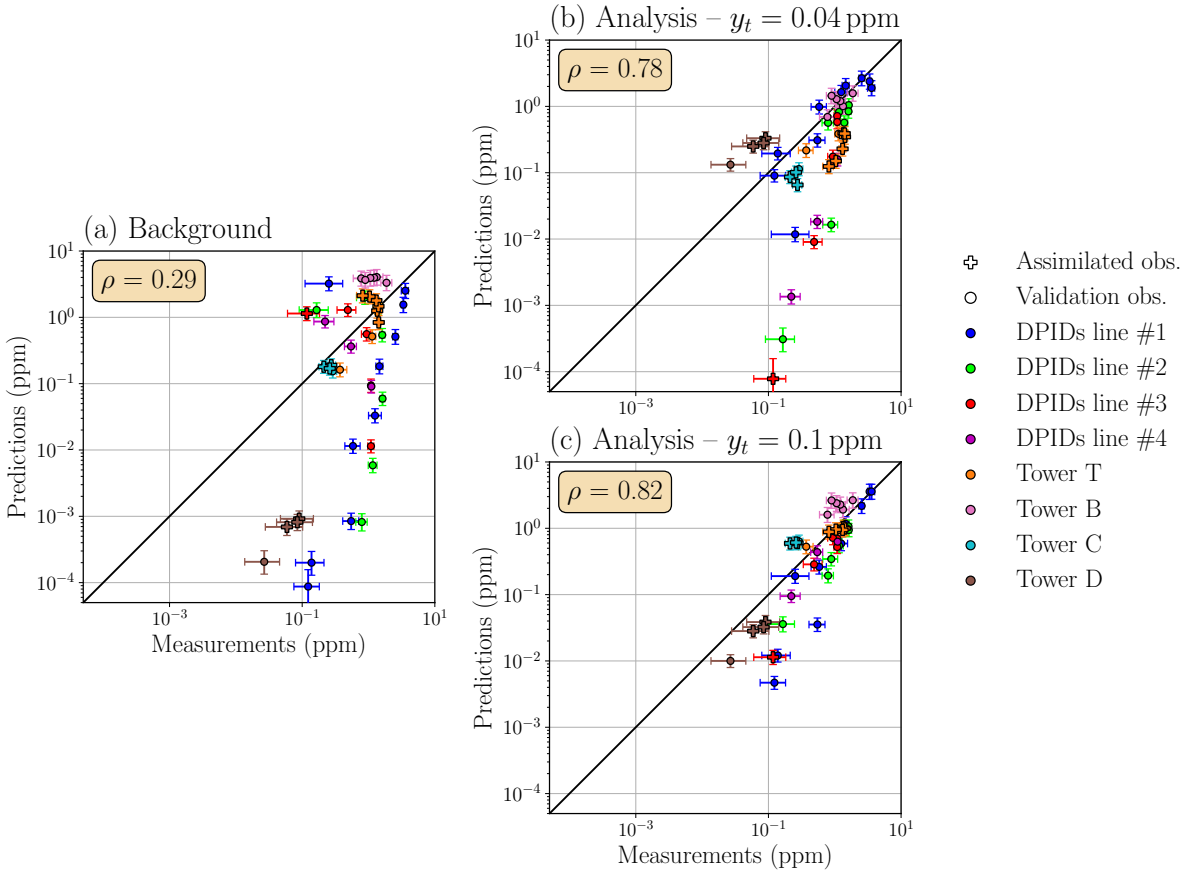


Figure V.15: *Scatter plots of the predicted versus measured mean concentration statistics for the background estimate (a) and two analysis estimates with different thresholds for the concentration anamorphosis: $y_t = 0.04\,ppm$ (b), and $y_t = 0.1\,ppm$ (c). Each sensor is represented with the same color as in Fig. II.3, page 59. Cross and circle symbols respectively represent assimilated and validation measurements. The Pearson correlation coefficient ρ is indicated in each case. The error bars represent the 95% confidence intervals estimated with stationary bootstrap for the measurements (see Sect. III.2, page 80) and using the POD–GPRs uncertainty estimation (see Sect. IV.2.5, page 131).*

order of magnitude ($\rho = 0.82$). In particular, the EnKF no longer deteriorates prediction at tower T, while continuing to improve the estimation at tower D. This is explained by the fact that, in this new experiment, the weights on assimilated observations are more homogeneous and the EnKF thus no longer overestimates wind direction. Indeed, we find that the analysis wind direction bias goes from $\overline{\alpha^a_{inlet}} - \alpha^{(ref)}_{inlet} = -8\,^\circ$ with $y_t = 0.04\,\mathrm{ppm}$ to just $-0.5\,^\circ$ with $y_t = 0.1\,\mathrm{ppm}$. However, this does not improve the estimation of friction velocity, since the bias of the analysis remains around $-0.15\,\mathrm{m\,s^{-1}}$.

V.6 Conclusion

Summary In this chapter, we have established a proof of concept of the reduced-cost DA system designed to address the thesis problematic of efficiently estimating and reducing uncertainties in LES models for microscale atmospheric dispersion. This system relies on an EnKF algorithm assimilating in-situ pollutant concentration measurements in order to infer two wind boundary condition parameters of the LES model presented in Chapter II: the inlet wind direction and friction velocity. The POD–GPRs reduced-order model built in Chapter IV replaces the LES model to successfully accelerate the estimation of the DA system. In addition, the EnKF choice enables us to move from deterministic prediction to probabilistic prediction that takes into account the various uncertainties involved in microscale atmospheric flows, thus fulfilling one of the main objectives of this thesis.

Improved error modeling Significant efforts have been made in this chapter to improve error modeling.

- For the background error, we carried out a detailed study comparing wind condition estimates from several weather stations to characterize representativeness and measurement errors that typically affect the LES model when its boundary conditions are calibrated from remote stations.

- For the observation error, we used the bootstrap approach developed in Chapter III to quantify the irreducible representativeness error caused by the internal variability of the ABL. This method provides efficient estimation of the complete observation error covariance matrix, including the cross-correlations which are often overlooked due to the lack of suitable estimates (Defforge et al. 2021).

- For the model error, we propose to take it into account in the EnKF by replacing the POD–GPRs mean prediction with random realizations sampled from GPR posterior distributions. As shown in Chapter IV, this allows us to account for both the model reduction error and the LES internal variability error. This approach improves on that of Sousa and Gorlé (2019) by preserving the spatial coherence of the errors.

All these developments constitute the main methodological contribution of this chapter and offer a cost-effective solution to the challenge raised by Mons et al. (2017) and Defforge et al. (2021), of providing realistic error models in the DA framework. This is an important step forward in ensuring the robustness of the DA system, by preventing it from giving too much credence to a source of information that is highly uncertain.

Validation of the DA system The proposed DA system was validated first in a twin experiment framework, i.e. using synthetic observations obtained from a LES model prediction, then using the real field measurements from the MUST trial 2681829. These experiments show that the EnKF succeeds in improving mean concentration by reducing the background bias in the wind direction. However, we have found that the DA system

has difficulty inferring friction velocity due to the low sensitivity of the log-transformed concentration to friction velocity. In addition, we show that a large enough ensemble size ($N_e \geq 500$ in our case) is required in the EnKF to limit the effect of background sampling errors. Finally, regarding computational cost, we demonstrate that the DA system achieves its objective with an ensemble prediction of the concentration field in less than a minute, without even taking advantage of the EnKF's parallelization opportunities.

Limitations and perspectives The assimilation of real observations has confronted the DA system we have designed with its own limitations, the analysis of which enables us to identify avenues for improvement.

- The inability of the DA system to infer the friction velocity limits the accuracy of the prediction and can lead to wind direction overcorrection. This could be solved by assimilating wind velocity measurements in addition to the concentration measurements. The relationship between local flow velocity and friction velocity is indeed much more direct, and it has been shown in the literature that assimilating different quantities significantly improves the DA estimation accuracy (Mons et al. 2017; Defforge et al. 2021).

- When assimilating real measurements, we found that the DA system fails to correct biases from the LES model that are not due to the meteorological boundary conditions. This inherent limitation of our system could be overcome by adopting a joint state/parameter estimation DA approach (Evensen 2009; Ruckstuhl and Janjić 2020; Zhang et al. 2019) to directly correct the mean concentration model biases while still inferring the large-scale forcing parameters.

- We show that DA estimations are significantly affected by the choice of the threshold used in the concentration anamorphosis, which ensures that the mean concentration observation and prediction are normally distributed. To our knowledge, there are no guidelines for selecting this parameter, so it would deserve further study in order to define an a priori selection method.

- Finally, since DA estimates are also known to be highly sensitive to the design of the assimilation sensor network (Mons et al. 2017; Sousa et al. 2018), testing and comparing optimal sensor placement strategies (Choi and How 2011; Peng et al. 2014; Dur et al. 2020; Deng et al. 2021b; Mons and Marquet 2021) seems a promising perspective to further improve the DA system presented in this chapter.

Conclusion and perspectives

Contributions and lessons learned

This thesis is a proof of concept of the ability of a reduced-cost data assimilation (DA) system to improve large-eddy simulation (LES) predictions of microscale dispersion in an idealized urban environment. This system quantifies and reduces uncertainty related to large-scale meteorological forcing by using information from in-situ measurements to solve an inverse problem. It relies on a reduced-order model to significantly speed up estimations. This is done while accounting for the internal variability of the atmospheric boundary layer (ABL), making the DA system more realistic and robust.

The main milestones in the construction of this system were

i) the further development of an LES model of one trial of the MUST field experiment (Biltoft 2001) corresponding to neutral stratification condition;

ii) the robust validation of this model by quantifying its main sources of uncertainty, comparing its predictions with experimental measurements, and investigating its sensitivities;

iii) the construction of a training database of 200 LES predictions, to train and optimize a reduced-order model that combines proper orthogonal decomposition (POD) and Gaussian process regressors (GPRs) to efficiently emulate the LES model response to changes in the meteorological boundary conditions;

iv) the assembling and testing of the reduced-cost data assimilation (DA) system, first via twin experiments, then using MUST field measurements.

The LES model, its reduced-order model, and the associated assimilation system have all proved to reach state-of-the-art performance levels. LES statistical predictions can be used to improve our understanding of complex flow patterns induced by the interaction between the ABL and the built environment, and their effect on microscale dispersion. Once trained for a specific urban configuration, the POD–GPRs reduced-order model provides faithful probabilistic dispersion predictions in less than a tenth of a second, that can be used to speed up the DA system and account for part of the prediction uncertainty. Finally, the DA system achieves to correct biases in large-scale meteorological conditions and provides robust and realistic uncertainty predictions.

During this thesis, we have paid particular attention to the internal variability of the ABL, which induces two forms of uncertainty in LES model predictions: i) an uncertainty in the model boundary conditions due to large-scale fluctuations in the atmosphere, and ii) a random and irreducible uncertainty corresponding to the effect of microscale fluctuations. The effect of uncertainty on the large-scale boundary conditions has already been quantified in microscale atmospheric flows (Wise et al. 2018; García-Sanchez and Gorlé 2018) and can be reduced through DA (Sousa et al. 2018; Sousa and Gorlé 2019; Defforge et al. 2021). However, to the best of our knowledge, the aleatory uncertainty inherent to microscale variability has been less thoroughly explored, and there are no ready-to-use methods for quantifying it. To overcome this shortcoming, we have developed and validated our approach based on the stationary bootstrap algorithm from Politis and Romano (1994). It is a promising, simple, and yet efficient method that does not require long acquisitions and relies on minimal statistical assumptions.

Using this bootstrap method, we demonstrate that microscale internal variability significantly affects not only LES predictions but also field observations. We also explain how it can be taken into account in model validation, thus meeting a need expressed by the scientific community (Schatzmann and Leitl 2011; Dauxois et al. 2021). The possibility to estimate the effect of internal variability has proven to be key throughout the whole thesis, as we used it to:

- demonstrate that structural uncertainties (e.g. related to turbulence modeling and numerical schemes) are not significant, unlike uncertainties linked to boundary conditions;
- make an informed choice on the number of reduced-basis modes used by the POD–GPRs model in order to avoid overfitting noisy structures linked to the aleatory uncertainty associated with internal variability, and thus optimize POD–GPRs predictions accuracy;
- take this uncertainty into account in the DA system to provide realistic error models for observations, background and model. These methodological improvements are made elegantly and without generating implementation or computational heaviness. They significantly improve the robustness of our DA system by preventing it from underestimating uncertainties, thereby filling a gap expressed by the community (Mons et al. 2017; Defforge et al. 2021).

From a broader point of view, this thesis underscores the potential for enhancing modeling chains by significantly speeding up advanced numerical model predictions without compromising their accuracy, all while assessing and reducing end-to-end uncertainties thanks to observational data. Such advancements hold significant importance for risk assessment applications.

Possible improvements

In the short-to-medium term, various improvements can be made to enhance the reduced-cost DA system built in this thesis and extend its scope. In this section, we summarize the main avenues for improving each component of our framework.

The LES microscale dispersion model can be further improved in terms of accuracy, as demonstrated by the limited level of agreement with field observations shown in Chapter III. Although we have explained some of these differences by the internal variability of the ABL or by the lack of vertical extent of the domain (Appendix A.2), there are still some differences that could not be corrected by our DA system. This means that model errors are not explained by meteorological boundary conditions uncertainty alone and motivates the exploration of new avenues. We could increase the level of details in the urban geometry representation as done by Santiago et al. (2010). We could also improve the meteorological boundary condition representativeness, starting by adopting more realistic wind direction and turbulent fluctuations vertical profiles, before going towards downscaling from mesoscale simulations (García-Sanchez and Gorlé 2018; Nagel et al. 2022). Comparison with other codes is also an exciting perspective, which could determine whether the error in our LES model is structural or not.

Note that the LES ensemble computed in this thesis was mainly used to build the reduced-order model for accelerating mean pollutant concentration predictions. This requires only a small fraction of the data collected during these high-resolution simulations. The substantial amount of data at hand merits further exploration to improve our understanding of the physical processes involved in microscale dispersion.

Internal variability quantification is mainly restricted to microscale fluctuations in this thesis since our LES model limits the size for the largest eddies. A rather straightforward and promising perspective is to quantify mesoscale variability, especially for longer analysis periods. For this purpose, a solution would be to use an LES model with inflow boundary conditions that are dynamically changing through a multi-scale meteorological model based on grid nesting (Wiersema et al. 2020; Nagel et al. 2022). Improving the representation of the large-scale atmospheric fluctuations is expected to fill the current gap between predicted and observed internal variability (see Chapter III).

The representation of the internal variability in the reduced-order model predictions could also be improved. Indeed, after optimization, the noise variance parameter of the Gaussian processes regressors (GPRs) fits the maximum variability observed in the training database (Chapter IV). However, the variance related to internal variability is not homogeneous in the space of input parameters, since it increases when the friction velocity decreases. To account for this heteroscedasticity of the response surface, it would be interesting to consider that the noise variance depends on the inputs, as proposed by Miyagusuku et al. (2015).

The reduced-order modeling approach we adopted remains limited in accuracy because of the error due to dimension reduction. We assume that this is mainly because the large disparity of scales in the mean concentration field is difficult to faithfully represent with a linear reduction method such as POD. We are therefore looking forward to replacing POD with non-linear reduction techniques such as convolutional autoencoders (Murata et al. 2020; Nony 2023). However, using autoencoders we could lose the hierarchical aspect of decomposition that was useful to filter out the effect of internal variability.

There is also room for improvement in the field preparation stage, as we found that it significantly affects the POD projection error. In particular, using a preprocessing based on log-transformation of the mean concentration improves the overall accuracy of field reconstruction, but deteriorates it in the vicinity of the source. Looking for optimal preprocessing for building the POD could therefore be another avenue to improve the reduction step. In Appendix B.3, we also show that it is possible to combine predictions of POD–GPRs built with different types of preprocessing to get the best of each preprocessing.

Data assimilation The experiments we have conducted in Chapter V to validate the DA system have several shortcomings:

- they are limited to the assimilation of concentration measurements and it would be interesting to investigate whether the gain in performance obtained by Mons et al. (2017) and Defforge et al. (2021) when assimilating wind speed and direction data can also be obtained in our case;
- we only correct an arbitrarily chosen bias on the meteorological boundary condition parameters, but in real applications, this bias is unknown, it would therefore be relevant to assess the robustness of our DA system for a wide range of biases;
- we show that DA predictions are strongly influenced by the threshold chosen for concentration anamorphosis. Further efforts could be made to study this sensitivity in order to understand how to make an informed choice of threshold value;
- according to observations we underestimate the background inlet wind direction error variance. However, when considering larger wind direction error variance the assumption of normally distributed errors no longer holds, highlighting the need to modify the ensemble Kalman filter (EnKF) for directional variables (Jammalamadaka and SenGupta 2001).

More importantly, when dealing with real observations, our DA system is not able to perfectly improve concentration predictions as it overcorrects parameters to compensate for model biases. This is an intrinsic limitation of our system, which can be overcome by modifying the definition of the control vector. In a first step, we could first discretize the boundary condition profiles to increase the parameter space dimension and thereby the number of degrees of freedom to solve the DA problem (Defforge et al. 2021). In a second step, we could investigate the use of joint state-parameters estimation DA algorithms (Evensen 2009; Smith et al. 2013; Ruckstuhl and Janjić 2020). This exciting perspective could allow for the correction of model biases (Zhang et al. 2019) but would require extra caution in the definition of the background error covariance matrix to guarantee the physical consistency of predictions. Given the huge size of the state vector ($N \approx 10^6$), this perspective may require combining reduction methods, such as the POD we are already using, with the DA algorithm (Tissot et al. 2013; Arcucci et al. 2019; Bauweraerts and Meyers 2021).

Finally, we note that recent efforts in associating machine learning techniques with DA (Cheng et al. 2023) could provide significant improvements to our framework, in particular, to better estimate model errors (Farchi et al. 2021).

Perspectives

In this section, we detail long-term perspectives arising from the work presented in this thesis. These prospects could significantly improve our reduced-cost DA system. In a first step, we examine how we might extend the scope of this system to a broader safety environment perspective. This is an opportunity to reflect on the scientific challenges associated with these new applications, such as integrating the temporal dimension to monitor unsteady phenomena. In the second step, we explore the specific but open question of assimilating data from mobile sensors to enrich the DA framework presented in this thesis.

Expanding the capacity of the data assimilation system

Towards broader safety environment applications To broaden the application scope of our reduced-cost DA system, we shall first extend its capabilities. This first includes complexifying the LES model to handle thermal stratification effects to gain in generality, and pollutant buoyancy to deal with dense or volatile pollutant species. Using the AVBP solver provides a way to model chemically reactive species and realistic sources. This typically includes the important dynamic and thermodynamic effects that occur when pressurized/liquified gases leak (Spicer and Tickle 2021). These developments are key to studying the dispersion of species widely used in industry, such as ammonia, and others for which there is growing interest, notably hydrogen as an energy carrier (Guilbert 2021) and carbon dioxide for capture in efforts to mitigate climate change (Delprat-Jannaud 2022). In the longer term, the choice to use AVBP also opens the door to new safety environment applications such as hydrogen combustion (Boivin et al. 2012), explosions (Vermorel et al. 2017), wildfires (Rochoux et al. 2014b), as well as the associated smoke dispersion.

These new applications will require well-instrumented validation experiments. We could first use other MUST trials to assess the ability of our model to simulate unstable and stable conditions (Biltoft 2001). The JOINT URBAN 2003 dispersion campaign (Allwine et al. 2004) in Oklahoma City (U.S.) or the dispersion wind-tunnel experiments reproducing Feyzin refinery (France) carried out by Vendel (2011) are also relevant case studies to account for more realistic urban canopy and to study puff releases. The more recent Jack Rabbit II chlorine dispersion field campaign (Fox et al. 2022), shown in Fig. VI.2a, is particularly suited to consider buoyancy and source effects. Finally, we note that major experimental efforts are underway to provide extensive measurements of real wildland fires, for example with the field-scale experiment shown in Fig. VI.2b or the more recent FASMEE campaign (Liu et al. 2019).

These new applications will also require developments in model reduction and data assimilation. As we opted for a non-intrusive model reduction approach, our DA system can be adapted to new case studies to integrate richer and more complex physics, provided that we enrich its learning base. Given the computational cost of most numerical models, limiting the number of simulations needed to train the reduced-order model is one of our main concerns. Based on the experience of this thesis, we will be able to use physical

similarity to reduce the number of simulations needed to account for different source intensities and thermal stratification conditions. With the framework now established, we will also be able to minimize the number of training simulations by incrementally constructing the reduced-order model and stopping as soon as we attain an accuracy plateau. In this sense, using adaptive sampling methods (Picheny et al. 2010; Braconnier et al. 2011) could be relevant to optimize the gain of accuracy from each new sample by prioritizing critical areas of the parameter space. Concerning the DA algorithm, the choice of the EnKF will remain convenient thanks to its simplicity, and its ability to deal with non-linear models and large control space dimensions.

In the following, we look at two specific issues arising from these potential applications.

Accounting for uncertainty in source location would allow our DA system to estimate pollutant release location in addition to meteorological conditions. This has direct application in safety monitoring for industrial sites such as refineries, where pollutants can leak from several different locations. However, specific methodological issues arise from the nature of source location uncertainty.

First, it would require a very large simulation ensemble to represent every possible combination of wind conditions and source location when training the reduced-order model. To address this challenge, Nony (2023) shows that adopting a multi-fidelity framework can save significant computational resources by decoupling sources of uncertainty. This relies on the fact that Navier-Stokes equations are independent of the pollutant transport equation when considering passive species. We could thus keep the reduced-order modeling approach adopted in this thesis to predict the wind flow field for various meteorological conditions, and then use these predictions as forcing for the decoupled transport equation to predict the tracer concentration field. The latter could then be solved for any source location possible using, for example, a highly parallelizable Lagrangian model.

Secondly, we note that the inverse problem of finding the source position is severely ill-posed (Hutchinson et al. 2017). Moreover, EnKF assumptions are not well-suited to describe the ensemble of possible source locations. We therefore should investigate the use of more direct algorithms such as Markov chain Monte Carlo (Gilks et al. 1995) to solve the Bayesian inverse problem. This algorithm is model-query intensive and therefore requires the multi-fidelity reduced-order model to be very efficient.

Tracking non-stationary phenomena, such as puff emissions, front fire propagation, or highly transient meteorological conditions, would require significant improvement of our reduced-cost DA system. The LES approach we use already provides an instantaneous representation of the atmospheric flow, and the EnKF is naturally suited for sequential estimation but it would require initial conditions to be included in the control vector. In our opinion, the greatest challenge lies in building a reduced-order model capable of predicting physical time series under uncertainty. It is an active research topic in fluid mechanics (Xie et al. 2018; Kim et al. 2019a; Deng et al. 2021a), but no perfect solution with limited database requirements has yet emerged, to the best of our knowledge.

In addition, we highlight that non-stationary phenomena, because of their characteristic time scales, are likely to be highly sensitive to the internal variability of the ABL. However, the internal variability quantification method we propose in this thesis relies on a stationarity assumption. For non-stationary applications, the standard approach is to perform several independent predictions corresponding to different turbulent states of the ABL (Harms et al. 2011; Costes et al. 2021), which is computationally demanding with LES. Methodological development towards accelerating internal variability quantification in non-stationary contexts is therefore of great interest.

Assimilation of observations from mobile sensors

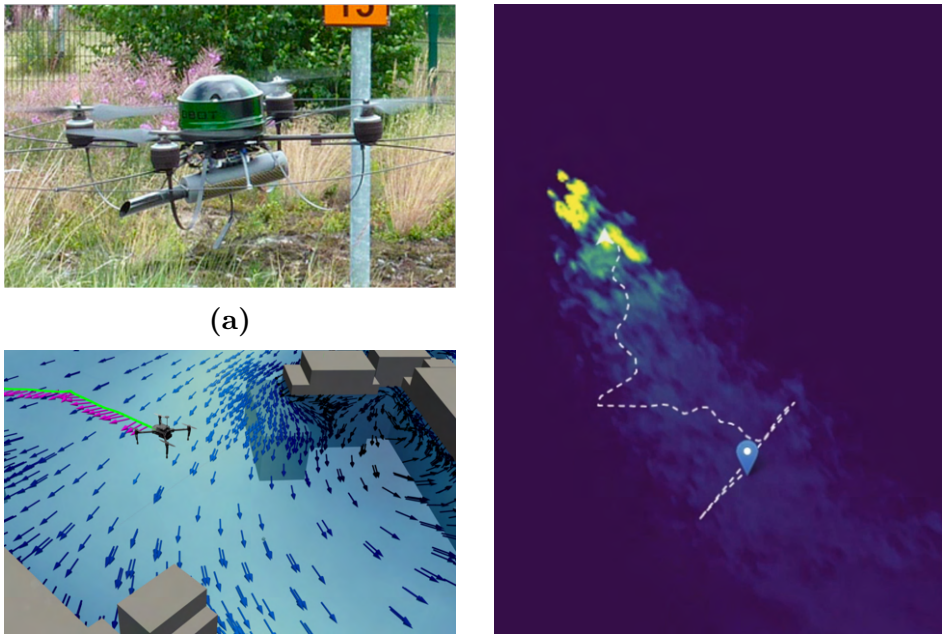
In recent years, unmanned aerial vehicles (UAVs), more commonly known as drones, have undergone remarkable technological advances, and are increasingly used for environmental monitoring (Manfreda et al. 2018), in applications such as agriculture (Kim et al. 2019b), river flooding (Perks et al. 2016), and wildland fire monitoring (Bailon-Ruiz and Lacroix 2020; Hu et al. 2022). By carrying various types of sensors, UAVs make it possible to i) sample a large area with a limited sensor budget, ii) take high-altitude pictures of a phenomenon at higher resolution than conventional means (aircraft, satellites) and lower cost (see Fig. VI.2a), and iii) target the area to be measured to maximize information gathering.

For these reasons, UAVs also have a great potential for microscale dispersion applications (Villa et al. 2016a). To date, many studies have focussed on plume tracking and pollutant source localization (Brink and Pebesma 2014; Montes et al. 2014; Letheren et al. 2016), which required the development of on-board gas sensors (Neumann et al. 2013; Villa et al. 2016b), see for example Fig. VI.1a. Instead of relying on an ill-posed inverse problem to identify pollutant source location, mobile sensors can directly track the source by following the concentration gradient. However, these navigation algorithms are often tested with simplified assumptions, for example using a Gaussian plume model (Montes et al. 2014). Confronting these algorithms to more realistic plumes, using instantaneous predictions of our LES model (Fig VI.1c), shows the need to further improve their robustness in order to deal with the turbulence of the ABL.

In the following, we imagine two different but complementary ways of taking advantage of these new data provided by drones, to enrich our DA system for microscale atmospheric dispersion.

i) Adaptive sampling for data assimilation

The pioneering study of Patrikar et al. (2020) has opened the way for the use of mobile sensors in DA for microscale applications by demonstrating that measurements from an anemometer carried by a UAV can be assimilated to improve CFD wind field predictions in urban environments (Fig. VI.1b). However, this study is limited to the assimilation of measurements alongside a predefined trajectory. In our opinion, it could be pushed even further by adapting UAV navigation to minimize errors as much as possible in the DA scheme.



(a) **(b)** **(c)**

Figure VI.1: *(a) Experimental setup by Neumann et al. (2013) using an Airrobot AR100-B micro-drone carrying a gas detector device (Dräger X-am 5600). (b) Real-world experiment from Patrikar et al. (2020) in which UAV measurements from an onboard anemometer (magenta arrows) are assimilated using a particle filter to correct inlet boundary conditions and improve CFD wind field prediction (blue arrows). (c) One snapshot of a prediction of instantaneous propylene concentration at $z = 4\,m$ from the LES model presented in Chapter II. Synthetic measurements are obtained by coupling these predictions with a drone flight simulator and then used to track the pollutant source with an adapted version of the navigation algorithm proposed by Montes et al. (2014). The resulting trajectory starting from the blue marker is represented by the dashed white line.*

Sensor position has indeed a very important effect on DA performance since when sensors are positioned in areas that are not very sensitive to the control vector, the assimilation problem becomes ill-posed (Mons et al. 2017; Sousa et al. 2018). This motivates the search for sensor networks that optimize DA accuracy (Dur et al. 2020; Deng et al. 2021b; Mons and Marquet 2021). Using mobile sensors would allow us to go even further by targeting measurement locations in real time, based on DA uncertainty updates. The development of such dynamic DA systems raises numerous issues including: i) the estimation of the information that will be obtained a priori by assimilating measurements from a given location (Choi and How 2011; Peng et al. 2014), ii) the choice of the sensor positions to maximize the total information, which explodes in complexity with the number of sensors to be placed (Qian and Claudel 2020), and iii) the trajectory planning of the fleet of UAVs to reach optimal positions in a efficient way (Reymann et al. 2018).

These are research topics on their own, and assembling them into a complete DA framework based on a CFD model and applied to a realistic case requires considerable effort, the fruits of which are undoubtedly innovative.

Possible limitations could arise from the uncertainty on drone positions, especially in windy conditions. Representing position error in a DA system represents an important issue (Rochoux et al. 2018) which should not be overlooked considering the strong gradients in a pollutant plume. Moreover, the short duration of the MUST experiments and the relatively large size of the container array limit the distance that can cover drones to provide representative measurements. We should therefore look for more appropriate experimental trials to conduct preliminary synthetic experiments.

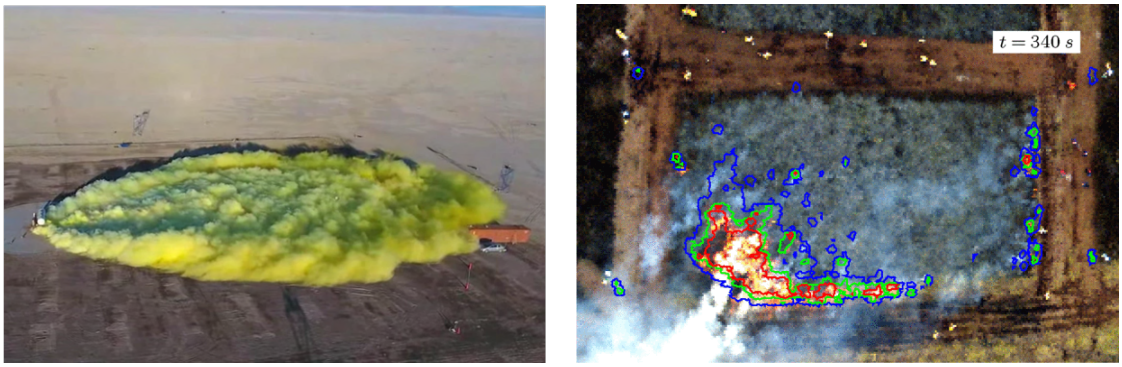
In addition to these technical constraints, we emphasize that the interaction between mobile sensors and the observed phenomenon – typically the pollutant plume deformation caused by the air flow induced by a rotorcraft – is an essential question to answer before developing such dynamic DA systems. Simple parameterizations exist for correcting the measurements of an anemometer carried by a UAV (Bruschi et al. 2016; Thorpe et al. 2018). However, we believe that further efforts are required before it can be accurately taken into account in the DA system presented in this thesis. For instance, LES carried out with AVBP, alongside real drone validation experiments, could enable fine characterization of the effect of rotorcraft blades on the local flow and plume, drawing on the analogy that can be made with wind turbine modeling in ABL simulations (Joulin et al. 2020; Dabas et al. 2022).

ii) Assimilation of plume front shape from airborne images

In some cases, UAVs can also be used to provide non-intrusive observational data by flying over the phenomenon and monitoring it with onboard sensors (visual or infrared cameras, and lidars). The data collected then offer macroscopic information on the phenomenon, such as its size, propagation speed, surface roughness, etc. Examples of airborne images from the Jack Rabbit II pollutant dispersion experiment, and from a field-scale wildland fire experiment are shown in Fig. VI.2. Note that for large-scale phenomena such as wildfires, rotorcrafts can be replaced by fixed-wing UAVs. This approach solves the problem of interaction with observed phenomena. For pollutant dispersion studies, it is however limited to species that are traceable at a distance.

This kind of observational data is very promising for DA. Although not directly measuring the quantity of interest complicates the observation operator and introduces new uncertainties, these global measurements can be more informative than precise local measurements. This is all the more the case in contexts where position errors are important, such as pollutant dispersion and fire front propagation (Rochoux et al. 2014b; Zhang et al. 2019).

To carry out a first numerical experiment of such system in a dispersion context, we imagine the following action plan: i) generate a sequence of synthetic images from an LES model and an inverse observation operator to represent the link between the pollutant concentration and the remotely measured quantity, ii) reconstruct the plume shape using a contour detection algorithm (Xin-Yi et al. 2018; Chu et al. 2022), this task basically consists in drawing the contour of the plume in Fig. VI.2a, and iii) apply a shape-oriented DA algorithm (Rochoux et al. 2018) to correct the predicted plume contour with the observed one. Once in place, this system could be validated with real



(a) **(b)**

Figure VI.2: *Aerial pictures of environmental field experiments of interest. (a) Picture of the chlorine plume corresponding to the trial #7 of the Jack Rabbit II field experiment (2016, US). Source: drone footage captured by Utah Valley University Emergency Services. (b) Picture of a heather controlled-burn experiment in Northumberland (March 2010, UK). The blue, green and red lines indicate the contours of the 500-K, 600-K and 700-K iso-temperatures respectively, as identified by Paugam et al. (2013) using a helicopter-borne infrared thermal camera.*

onboard visual measurements such as the ones from the Jack Rabbit II field campaign (Fig. VI.2a).

Ultimately, these two types of observational data provided by UAVs – aerial images, and direct in-situ measurements – could be assimilated jointly by our system. This has the potential to better address the uncertainty in LES models by correcting both positional and intensity errors, and thus further enrich the accuracy of microscale dispersion predictions.

Conclusion et perspectives

Contributions et enseignements

Cette thèse constitue une preuve de concept de la capacité d'un système d'assimilation de données (AD) à coût réduit à améliorer les prévisions de dispersion à micro-échelle dans un environnement urbain idéalisé par la simulation aux grandes échelles de tourbillons (LES). Ce système quantifie et réduit l'incertitude liée au forçage météorologique à grande échelle en utilisant des informations provenant de mesures in situ pour résoudre un problème inverse. Il s'appuie sur un modèle d'ordre réduit pour accélérer considérablement les estimations. Cela se fait tout en tenant compte de la variabilité interne de la couche limite atmosphérique (CLA), ce qui rend le système d'AD plus réaliste et plus robuste.

Les principales étapes de la construction de ce système ont été les suivantes :

i) le développement d'un modèle LES d'un essai de la campagne expérimentale MUST correspondant à une condition de stratification neutre de l'atmosphère,

ii) la validation robuste de ce modèle en comparant ses prévisions avec les mesures expérimentales, tout en quantifiant ses principales sources d'incertitude,

iii) la construction d'une base de données d'entraînement de 200 prévisions LES, afin d'entraîner et d'optimiser un modèle d'ordre réduit qui combine décomposition orthogonale aux valeurs propres (POD) et régression par processus gaussiens (GPR) pour émuler efficacement la réponse du modèle LES aux changements des conditions aux limites météorologiques,

iv) l'assemblage et le test du système d'assimilation de données (AD) à coût réduit, d'abord par le biais d'expériences jumelles, puis en utilisant les mesures de terrain de la campagne MUST.

Il est démontré que le modèle LES, son modèle d'ordre réduit et le système d'assimilation de données associé atteignent tous de hauts niveaux de performance. Les estimations statistiques LES peuvent être utilisées pour améliorer notre compréhension des schémas d'écoulement complexes induits par l'interaction entre la CLA et l'environnement bâti, et leur effet sur la dispersion à micro-échelle. Une fois entraîné pour une configuration urbaine spécifique, le modèle d'ordre réduit POD–GPRs fournit des prévisions de dispersion probabilistes précises en moins d'un dixième de seconde, qui peuvent être utilisées pour

accélérer le système d'AD tout en tenant compte d'une partie des incertitudes en jeu. Enfin, le système d'AD parvient à corriger des biais dans les conditions météorologiques à grande échelle et fournit des prévisions d'incertitude robustes et réalistes.

Au cours de cette thèse, une attention particulière est portée à la variabilité interne de la CLA, qui induit deux formes d'incertitude dans les estimations des modèles LES : i) une incertitude sur les conditions aux limites du modèle due aux fluctuations à grande échelle de l'atmosphère, et ii) une incertitude aléatoire et irréductible correspondant à l'effet des fluctuations à micro-échelle. L'effet de l'incertitude sur les conditions limites à grande échelle sur les prévisions d'écoulements à micro-échelle a déjà été étudié (Wise et al. 2018; García-Sanchez and Gorlé 2018) et peut être réduit grâce à l'AD (Sousa et al. 2018; Sousa and Gorlé 2019; Defforge et al. 2021). Cependant, à notre connaissance, l'incertitude aléatoire inhérente à la variabilité micro-échelle a été moins bien étudiée et il n'existe pas de méthodes prêtes à l'emploi pour la quantifier. Pour combler cette lacune, nous avons développé et validé une approche basée sur l'algorithme de bootstrap stationnaire de Politis and Romano (1994). Il s'agit d'une méthode prometteuse, simple et néanmoins efficace, qui ne nécessite pas de longues acquisitions et repose sur des hypothèses statistiques minimales.

En utilisant cette méthode de bootstrap, nous démontrons que la variabilité interne à micro-échelle affecte significativement non seulement les prévisions LES mais aussi les observations de terrain. Nous expliquons également comment elle peut être prise en compte dans la validation des modèles, répondant ainsi à un besoin exprimé par la communauté scientifique (Schatzmann and Leitl 2011; Dauxois et al. 2021). La possibilité d'estimer l'effet de la variabilité interne s'est avérée essentielle tout au long de la thèse, puisque nous l'avons utilisée pour :

- démontrer que les incertitudes structurelles (par exemple liées à la modélisation de la turbulence et aux schémas numériques) ne sont pas significatives, contrairement aux incertitudes liées aux conditions aux limites,
- faire un choix éclairé du nombre de modes utilisés par le modèle POD–GPRs afin d'éviter de reproduire les structures liées à l'incertitude aléatoire associée à la variabilité interne lors de l'apprentissage, et ainsi optimiser la précision des prévisions POD–GPRs,
- prendre en compte cette incertitude dans le système d'AD afin de fournir des modèles d'erreur réalistes pour les observations, l'ébauche et le modèle. Ces améliorations méthodologiques sont apportées de manière élégante et sans générer de lourdeur de mise en œuvre ou de calcul. Elles améliorent la robustesse de notre système d'AD en l'empêchant de sous-estimer les incertitudes, comblant ainsi une lacune exprimée par la communauté (Mons et al. 2017; Defforge et al. 2021).

D'un point de vue plus large, cette thèse souligne le potentiel d'amélioration des chaînes de modélisation de la dispersion à micro-échelle en accélérant de manière significative les prévisions d'un modèle CFD coûteux sans compromettre sa précision, tout en évaluant et en réduisant les incertitudes de bout en bout grâce aux données d'observation. De telles avancées revêtent une importance significative pour les applications de prévision des risques.

Améliorations possibles

À court et moyen terme, diverses modifications peuvent être apportées pour améliorer et étendre le champ d'application du système d'AD à coût réduit construit dans cette thèse. Dans cette section, nous résumons les principales pistes d'amélioration de chacune des composantes du système

Le modèle de dispersion à micro-échelle LES peut être encore amélioré en termes de précision, comme illustré par le niveau limité de concordance avec les mesures expérimentales constaté dans le Chapitre III. Bien que nous ayons expliqué certaines de ces différences par la variabilité interne de la CLA ou par le manque d'étendue verticale du domaine (Annexe A.2), il existe encore des différences qui n'ont pas pu être corrigées par notre système d'AD. Cela signifie que les erreurs du modèle ne s'expliquent pas uniquement par l'incertitude des conditions aux limites météorologiques et motive l'exploration de nouvelles pistes. Nous pourrions par exemple augmenter le niveau de détail de la représentation de la géométrie urbaine, comme l'ont fait Santiago et al. (2010). Nous pourrions également améliorer la représentativité des conditions aux limites météorologiques, en commençant par adopter des profils verticaux de direction du vent et de fluctuations turbulentes plus réalistes, avant de passer à la réduction d'échelle à partir de simulations à méso-échelle (García-Sanchez and Gorlé 2018; Nagel et al. 2022). La comparaison avec d'autres codes est également une perspective intéressante, qui pourrait déterminer si l'erreur dans notre modèle LES est structurelle ou non.

Il convient de noter que l'ensemble LES calculé dans cette thèse a été principalement utilisé pour construire le modèle d'ordre réduit permettant d'accélérer les prévisions de concentration moyenne de polluant. Ceci ne nécessite qu'une petite fraction des données collectées lors de ces simulations à haute résolution. La quantité substantielle de données disponibles mérite d'être explorée davantage afin d'améliorer notre compréhension des processus physiques impliqués dans la dispersion à micro-échelle.

La quantification de la variabilité interne est principalement limitée dans cette thèse aux fluctuations à micro-échelle puisque notre modèle LES limite la taille des plus grands tourbillons. Une perspective assez directe et prometteuse est donc de quantifier la variabilité à méso-échelle, en particulier pour des périodes d'analyse plus longues. Pour ce faire, une solution serait d'utiliser un modèle LES multi-échelle avec des conditions aux limites d'entrée définies par descente d'échelle et qui changent dynamiquement (Wiersema et al. 2020; Nagel et al. 2022). L'amélioration de la représentation des fluctuations atmosphériques à grande échelle devrait permettre de combler l'écart actuel entre la variabilité interne prévue et observée (voir le Chapitre III).

La représentation de la variabilité interne dans les prévisions du modèle d'ordre réduit POD–GPRs pourrait également être améliorée. En effet, après optimisation, le paramètre de variance du bruit des processus gaussiens correspond à la variabilité maximale observée dans la base de données d'apprentissage, comme démontré dans le Chapitre IV. Cependant, la variance liée à la variabilité interne n'est pas homogène dans l'espace des paramètres d'entrée, puisqu'elle augmente lorsque la vitesse de frottement diminue. Pour

rendre compte de cette hétéroscédasticité de la surface de réponse, il serait intéressant de prendre en compte la dépendance de la variance du bruit aux paramètres d'entrée, comme le proposent Miyagusuku et al. (2015).

L'approche de réduction de modèle POD–GPRs que nous avons adoptée reste limitée en termes de précision en raison de l'erreur due à l'étape de réduction de la dimension. À notre avis, cela est principalement dû au fait qu'il est difficile de représenter fidèlement l'intégralité des échelles de concentration avec une méthode de réduction linéaire telle que la POD. Il nous semble donc très pertinent de réfléchir à remplacer la POD par des techniques de réduction non linéaires telles que les auto-encodeurs convolutifs (Murata et al. 2020; Nony 2023). Cependant, en utilisant des auto-encodeurs, nous pourrions perdre l'aspect hiérarchique de la décomposition qui est utile pour filtrer l'effet de la variabilité interne.

Il est également possible d'améliorer l'étape de prétraitement des champs, car nous avons constaté qu'elle affecte de manière significative l'erreur de projection POD. En particulier, l'utilisation d'un prétraitement basé sur la transformation logarithmique de la concentration moyenne améliore la précision globale du modèle réduit mais la détériore à proximité de la source. La recherche d'un prétraitement optimal pour la construction du POD pourrait donc constituer une autre moyen d'améliorer l'étape de réduction. Dans l'Annexe B.3, nous montrons également qu'il est possible de combiner les prévisions des POD–GPR construits avec différents types de prétraitement pour obtenir le meilleur de chaque prétraitement.

Asimilation de données Les expériences que nous avons menées dans le Chapitre V pour valider le système d'AD présentent plusieurs limitations :

- nous assimilons uniquement des mesures de concentration et il serait intéressant d'étudier si le gain de performance obtenu par Mons et al. (2017) et Defforge et al. (2021) lors de l'assimilation des données relatives à la vitesse et à la direction du vent peut également être obtenu dans notre cas,
- nous ne corrigeons qu'un biais choisi arbitrairement sur les paramètres des conditions aux limites météorologiques, mais dans les applications réelles, ce biais est inconnu, il serait donc pertinent d'évaluer la robustesse de notre système d'AD pour une large gamme de biais,
- nous montrons que les prévisions de l'AD sont fortement influencées par le seuil choisi pour l'anamorphose de la concentration. Des efforts supplémentaires pourraient être menés pour étudier cette sensibilité afin de comprendre comment faire un choix éclairé de la valeur du seuil,
- il conviendrait d'augmenter la variance d'erreur d'ébauche pour la direction du vent car elle est sous-estimée par rapport aux observations. Toutefois, l'hypothèse de distribution d'erreur gaussienne ne serait alors plus vérifiée, ce qui souligne la nécessité de modifier le filtre de Kalman d'ensemble (EnKF) pour les variables directionnelles (Jammalamadaka and SenGupta 2001).

Plus important encore, lorsqu'il s'agit d'observations réelles, notre système d'AD n'est pas en mesure d'améliorer parfaitement les prévisions de concentration car il surcorrige les paramètres de contrôle pour compenser les biais du modèle. Il s'agit d'une limitation intrinsèque de notre système, qui peut être surmontée en modifiant la définition du vecteur de contrôle. Dans un premier temps, nous pourrions discrétiser les profils de conditions aux limites afin d'augmenter la dimension de l'espace des paramètres et donc le nombre de degrés de liberté pour résoudre le problème d'AD (Defforge et al. 2021). Dans un second temps, nous pourrions envisager d'utiliser des algorithmes d'AD d'estimation jointe état-paramètres (Evensen 2009; Smith et al. 2013; Ruckstuhl and Janjić 2020). Cette perspective prometteuse pourrait permettre de corriger les biais du modèle (Zhang et al. 2019), mais nécessite une prudence accrue dans la définition de la matrice de covariance d'erreur d'ébauche afin de garantir la cohérence physique des estimations. Compte tenu de la taille considérable du vecteur d'état ($N \approx 10^6$), cette perspective peut nécessiter la combinaison de méthodes de réduction, telles que la POD que nous utilisons déjà, avec l'algorithme d'AD (Tissot et al. 2013; Arcucci et al. 2019; Bauweraerts and Meyers 2021).

Enfin, les récents progrès dans l'utilisation de techniques d'apprentissage automatique pour l'AD (Cheng et al. 2023) pourraient apporter des améliorations significatives à notre système, en particulier pour le traitement des erreurs du modèle (Farchi et al. 2021).

Perspectives

Dans cette section, nous détaillons les perspectives à long terme découlant des travaux présentés dans cette thèse. Ces perspectives pourraient améliorer de manière significative notre système d'AD à coût réduit. Dans un premier temps, nous examinons comment nous pourrions étendre le champ d'application de ce système. C'est l'occasion de réfléchir aux défis scientifiques associés à ces nouvelles applications, comme l'intégration de la dimension temporelle pour le suivi des phénomènes instationnaires. Dans un second temps, nous explorons la question spécifique mais ouverte de l'assimilation de données provenant de capteurs mobiles pour enrichir le système d'AD présenté dans cette thèse.

Expansion des capacités du système d'assimilation de données

Vers de nouvelles applications Pour élargir le champ d'application de notre système d'AD à coût réduit, il conviandra tout d'abord d'étendre les capacités du modèle LES, pour : i) traiter tous les différents scénarios de stratification thermique de l'atmosphère, afin de gagner en généralité, et ii) considérer la flottabilité des polluants, afin de pouvoir traiter des espèces polluantes denses ou volatiles. De plus, l'utilisation du solveur AVBP rend possible la modélisation d'espèces chimiquement réactives et des effets dynamiques et thermodynamiques de source, qui se produisent par exemple lorsque des gaz pressurisés/liquifiés fuient (Spicer and Tickle 2021). Ces développements sont essentiels pour étudier la dispersion d'espèces largement utilisées dans l'industrie, comme l'ammoniac, et d'autres qui suscitent un intérêt croissant, notamment l'hydrogène en tant que vecteur d'énergie (Guilbert 2021) et le dioxyde de carbone en vue de sa capture afin d'atténuer le changement climatique (Delprat-Jannaud 2022). À plus long terme, le choix d'utiliser

AVBP ouvre également la voie à de nouvelles applications dans le domaine de la sécurité, telles que la combustion de l'hydrogène (Boivin et al. 2012), les explosions (Vermorel et al. 2017), les feux de forêt (Rochoux et al. 2014b), ainsi que la dispersion des fumées qui y sont associées.

Ces nouvelles applications nécessiteront des expériences de validation bien instrumentées. Nous pourrons d'abord utiliser d'autres essais MUST pour évaluer la capacité de notre modèle à simuler des conditions instables et stables (Biltoft 2001). La campagne de dispersion JOINT URBAN 2003 (Allwine et al. 2004) à Oklahoma City (États-Unis) ou les expériences de dispersion en soufflerie reproduisant la raffinerie de Feyzin (France) menées par Vendel (2011) sont également des études de cas pertinentes pour tenir compte d'une canopée urbaine plus réaliste et pour étudier des rejets de polluants intermittents. La campagne de dispersion de chlore Jack Rabbit II (Fox et al. 2022), illustrée Fig. VI.4a, pourrait également être particulièrement adaptée à la prise en compte des effets de flottabilité et de source. Enfin, nous notons que des efforts expérimentaux importants sont menés pour fournir des mesures détaillées de feux de forêt réels, par exemple avec l'expérience à l'échelle du terrain illustrée Fig. VI.4b ou la campagne FASMEE plus récente (Liu et al. 2019).

Ces nouvelles applications demanderont également des développements en matière de réduction de modèle et d'assimilation de données. Comme nous avons opté pour une approche de réduction de modèle non intrusive, notre système d'AD peut être adapté à de nouveaux cas d'étude pour intégrer une physique plus riche et plus complexe, à condition d'enrichir sa base d'apprentissage. Etant donné le coût de calcul de la plupart des modèles numériques, limiter le nombre de simulations nécessaires pour entraîner le modèle d'ordre réduit est l'une de nos principales préoccupations. Grâce à l'expérience de cette thèse, nous saurons comment exploiter la théorie de similitude physique de la dispersion afin de réduire le nombre de simulations nécessaires pour émuler la réponse du modèle LES pour différents débits d'émission de polluants et différentes conditions de stratification thermique de l'atmosphère. Le système d'AD étant maintenant établi, nous serons également en mesure de minimiser le nombre de simulations d'entraînement en construisant de manière incrémentale le modèle d'ordre réduit et en nous arrêtant dès que nous atteignons un plateau de précision. En ce sens, l'utilisation de méthodes d'échantillonnage adaptatif (Picheny et al. 2010; Braconnier et al. 2011) pourrait être pertinente pour optimiser le gain de précision de chaque nouvel échantillon en donnant la priorité aux zones critiques de l'espace des paramètres. En ce qui concerne l'algorithme d'AD, le choix de l'EnKF restera pratique grâce à sa simplicité et à sa capacité à traiter des modèles non linéaires et des espaces de contrôle de grande dimension.

Dans ce qui suit, nous examinons deux problématiques spécifiques qui découlent de ces potentielles nouvelles applications.

La prise en compte de l'incertitude liée à l'emplacement de la source permettrait à notre système d'AD d'estimer l'emplacement des sources de polluants en plus des conditions météorologiques. Cela a une application directe pour la surveillance des sites industriels tels que les raffineries, où les polluants peuvent s'échapper de plusieurs

endroits différents. Toutefois, la nature de l'incertitude relative à la localisation de la source soulève des questions méthodologiques spécifiques.

Tout d'abord, il faudrait un très grand ensemble de simulations pour représenter toutes les combinaisons possibles de conditions de vent et d'emplacement de source lors de l'apprentissage du modèle d'ordre réduit. Pour relever ce défi, Nony (2023) montre que l'adoption d'un système multi-fidélité peut permettre d'économiser des ressources de calcul importantes en découplant les formes d'incertitude. Cela repose sur le fait que les équations de Navier-Stokes sont indépendantes de l'équation de transport des polluants lorsque l'on considère des espèces passives. Nous pourrions ainsi conserver l'approche de réduction de modèle adoptée dans cette thèse pour prévoir le champ d'écoulement du vent pour diverses conditions météorologiques, puis utiliser ces prévisions comme forçage pour l'équation de transport découplée afin de prévoir le champ de concentration de polluant. Cette dernière pourrait alors être résolue pour tout emplacement de source possible en utilisant, par exemple, un modèle lagrangien hautement parallélisable.

Ensuite, il conviendra de prendre en compte que le problème inverse de la recherche de la position de la source est fortement mal posé (Hutchinson et al. 2017). De plus, les hypothèses de l'EnKF ne sont pas bien adaptées pour décrire l'ensemble des positions possibles de la source. Nous devrions donc envisager d'utiliser des algorithmes d'AD plus directs tels que les méthodes de Monte-Carlo par chaînes de Markov (Gilks et al. 1995) pour résoudre le problème inverse. Ce genre d'algorithme est très gourmand en évaluations du modèle et nécessitera donc un modèle d'ordre réduit multi-fidélité très efficace.

Le suivi des phénomènes instationnaires, tels que les émissions intermittentes, la propagation des fronts de feux de forêt ou bien des conditions météorologiques fortement transitoires, nécessiterait une amélioration significative de notre système d'AD à coût réduit. L'approche de modélisation LES que nous utilisons fournit déjà une représentation instantanée de l'écoulement atmosphérique, et l'EnKF est naturellement adapté à l'estimation séquentielle, mais nécesiterait d'inclure les conditions initiales dans le vecteur de contrôle. À notre avis, le plus grand défi résiderait surtout dans la construction d'un modèle d'ordre réduit capable de prévoir des séries temporelles physiques dans des contextes hautement incertains. Il s'agit d'un sujet de recherche actif en mécanique des fluides (Xie et al. 2018; Kim et al. 2019a; Deng et al. 2021a), mais aucune solution idéale avec des besoins limités en matière de base de données d'apprentissage n'a encore émergé à notre connaissance.

De plus, nous soulignons que les phénomènes instationnaires, de par leurs échelles de temps caractéristiques, sont susceptibles d'être très sensibles à la variabilité interne de la CLA. Cependant, la méthode de quantification de la variabilité interne que nous proposons dans cette thèse repose sur une hypothèse de stationnarité. Pour les applications instationnaires, l'approche standard consiste à effectuer plusieurs prévisions indépendantes correspondant à différents états turbulents de la CLA (Harms et al. 2011; Costes et al. 2021), ce qui est très exigeant en termes de calcul avec LES. Le développement méthodologique visant à accélérer la quantification de la variabilité interne dans des contextes instationnaires est donc d'un grand intérêt.

Assimilation d'observations provenant de capteurs mobiles

Ces dernières années, les véhicules aériens sans humain à bord (UAV), plus communément appelés drones, ont connu des avancées technologiques remarquables et sont de plus en plus utilisés pour la surveillance de l'environnement (Manfreda et al. 2018), dans des applications telles que l'agriculture (Kim et al. 2019b) et la surveillance des crues des rivières (Perks et al. 2016) et des feux de forêt (Bailon-Ruiz and Lacroix 2020; Hu et al. 2022). En embarquant différents types de capteurs, les drones permettent i) d'échantillonner une large zone avec un budget capteur limité, ii) de prendre des images à haute altitude d'un phénomène (voir par exemple Fig. VI.4a) avec une résolution plus élevée que les moyens d'observation conventionnels (avions, satellites) et à moindre coût, et iii) de cibler la zone à mesurer pour maximiser la collecte d'informations.

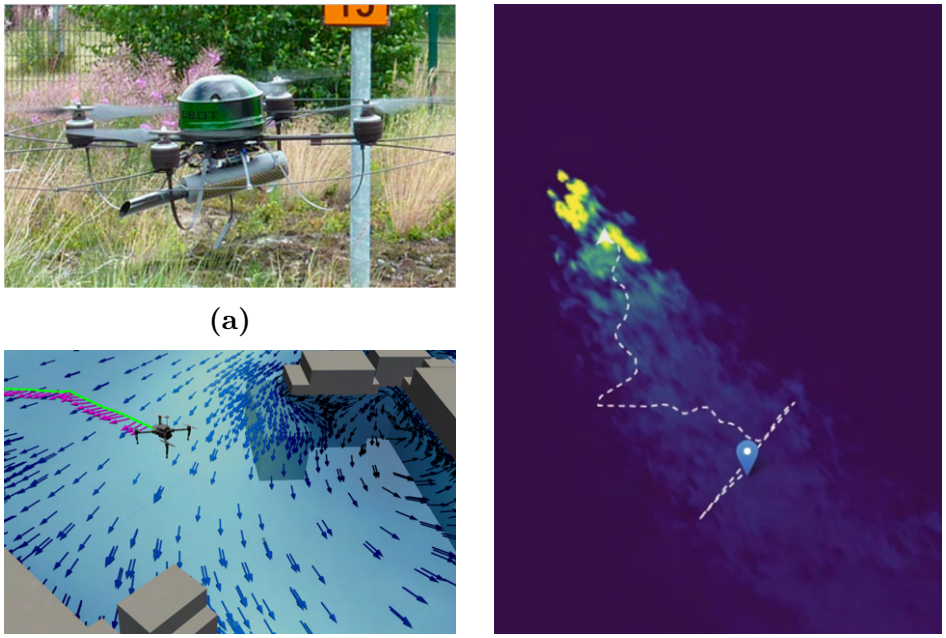
Pour ces raisons, les drones ont également un grand potentiel pour l'étude de la dispersion à micro-échelle (Villa et al. 2016a). Jusqu'à présent, de nombreuses études se sont concentrées sur le suivi des panaches et la localisation des sources de polluants (Brink and Pebesma 2014; Montes et al. 2014; Letheren et al. 2016), ce qui a nécessité le développement de capteurs de concentration en gaz embarqués (Neumann et al. 2013; Villa et al. 2016b), comme illustré Fig. VI.3a. Au lieu de résoudre un problème inverse mal posé pour identifier l'emplacement de la source des polluants, les capteurs mobiles peuvent permettre de directement remonter à la source en suivant le gradient de concentration. Cependant, ces algorithmes de navigation sont souvent testés avec des hypothèses simplifiées, par exemple en utilisant un modèle de panache gaussien (Montes et al. 2014). La confrontation de ces algorithmes à des panaches plus réalistes, en utilisant les prévisions d'instantanés de notre modèle LES (Fig VI.3c), montre la nécessité d'améliorer encore leur robustesse pour faire face à la turbulence de la CLA.

Dans la suite, nous imaginons deux manières différentes mais complémentaires de tirer parti de ces nouvelles données fournies par les drones, pour enrichir notre système d'AD pour la dispersion atmosphérique à micro-échelle.

i) Échantillonnage adaptatif pour l'assimilation de données

L'étude pionnière de Patrikar et al. (2020) a ouvert la voie de l'assimilation de données de capteurs mobiles pour des applications à micro-échelle en démontrant que les mesures d'un anémomètre transporté par un drone peuvent être assimilées pour améliorer les estimations d'un modèle CFD du champ de vent dans les environnements urbains (Fig. VI.3b). Cependant, cette étude se limite à l'assimilation de mesures le long d'une trajectoire prédéfinie. À notre avis, cette approche pourrait être poussée encore plus loin en adaptant la navigation des drones dans l'objectif de minimiser l'erreur d'estimation du système d'AD.

La position des capteurs a en effet un effet très important sur la performance de l'AD. En effet, lorsque les capteurs sont positionnés dans des zones qui ne sont pas très sensibles au vecteur de contrôle, le problème d'assimilation devient mal posé (Mons et al. 2017; Sousa et al. 2018). Cela motive la recherche de réseaux de capteurs qui optimisent la précision de l'AD (Dur et al. 2020; Deng et al. 2021b; Mons and Marquet 2021). L'utilisation de capteurs mobiles nous permettrait d'aller encore plus loin en ciblant les emplacements



(a) **(b)** **(c)**

Figure VI.3: *(a) Dispositif expérimental mis au point par Neumann et al. (2013) comportant un micro-drone Airrobot AR100-B qui transporte un détecteur de gaz (Dräger X-am 5600). (b) Expérience réelle réalisée par Patrikar et al. (2020) dans laquelle les mesures d'un drone provenant d'un anémomètre embarqué (flèches magenta) sont assimilées à l'aide d'un filtre particulaire pour corriger les conditions limites d'un modèle CFD et améliorer ses prévisions du vent local (flèches bleues). (c) Champ de concentration instantanée de propène à $z = 4\,m$ estimé par le modèle LES présenté dans le Chapitre II. Des mesures synthétiques sont obtenues en couplant ces estimations avec un simulateur de vol de drone, puis utilisées pour remonter à la source de polluant avec une version adaptée de l'algorithme de navigation proposé par Montes et al. (2014). La trajectoire résultante partant du marqueur bleu est représentée par la ligne blanche en pointillés.*

de mesure en temps réel, sur la base des mises à jour de l'incertitude estimée par l'AD. Le développement de tels systèmes d'AD dynamiques soulève de nombreux challenges, notamment : i) l'estimation de l'information qui sera obtenue a priori en assimilant les mesures d'un emplacement donné (Choi and How 2011; Peng et al. 2014), ii) le choix des positions des capteurs pour maximiser l'information totale, dont la complexité explose avec le nombre de capteurs à placer (Qian and Claudel 2020), et iii) la planification de la trajectoire de la flotte de drones pour atteindre les positions optimales de manière efficace (Reymann et al. 2018).

Il s'agit là de sujets de recherche à part entière et leur assemblage dans un système d'AD dynamique complet, basé sur un modèle CFD et appliqué à un cas réaliste, nécessiterait un effort considérable, mais serait sans aucun doute très novateur.

Des limitations éventuelles pourraient résulter de l'incertitude liée à la position des drones, en particulier dans des conditions venteuses. La représentation des erreurs de

position dans un système d'AD constitue une question importante (Rochoux et al. 2018) qui ne doit pas être négligée compte tenu des forts gradients dans un panache de polluants. En outre, la courte durée des expériences MUST et la taille relativement importante de la canopée urbaine simplifiée limitent la distance qui peuvent couvrir les drones pour fournir des mesures représentatives. Nous devrions donc rechercher des essais expérimentaux plus appropriés pour mener des expériences synthétiques préliminaires.

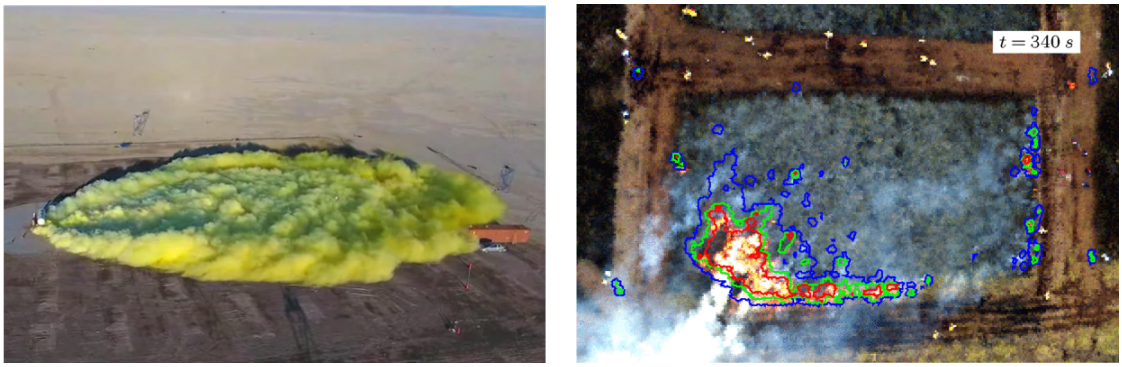
En plus de ces contraintes techniques, nous soulignons que l'interaction entre les capteurs mobiles et le phénomène observé – typiquement la déformation du panache de polluants causée par l'écoulement d'air induit par un giravion – est une question essentielle à laquelle il faut répondre avant de développer de tels systèmes d'AD dynamiques. Des paramétrisations simples existent pour corriger les mesures d'un anémomètre transporté par un drone (Bruschi et al. 2016; Thorpe et al. 2018). Cependant, nous pensons que des efforts supplémentaires sont nécessaires afin de pouvoir prendre en compte avec précision ce genre d'interactions dans le système d'AD présenté dans cette thèse. Par exemple, des LES réalisées avec AVBP, parallèlement à des campagnes de mesures de drones, pourraient permettre de caractériser finement l'effet des pales d'un giravion sur l'écoulement local et le panache, en s'inspirant de l'analogie qui peut être faite avec la modélisation des éoliennes dans les simulations de CLA (Joulin et al. 2020; Dabas et al. 2022).

ii) Assimilation de données de front de panache à partir d'images aéroportées

Dans certains cas, les drones peuvent également être utilisés pour fournir des données d'observation non intrusives en survolant le phénomène et en le surveillant à l'aide de capteurs embarqués (caméras visuelles ou infrarouges, ou bien lidars). Les données collectées offrent alors des informations macroscopiques sur le phénomène, telles que sa taille, sa vitesse de propagation, la rugosité de sa surface, etc. Des exemples d'images aériennes issues de l'expérience de dispersion de polluants Jack Rabbit II et d'une expérience de feux de forêt sont présentés dans la Fig. VI.4. Cette approche résout le problème de l'interaction avec les phénomènes observés. Pour les études de dispersion des polluants, elle est cependant limitée aux espèces traçables à distance.

Ce type de données d'observation est très prometteur pour l'AD. En effet, bien que le fait de ne pas mesurer directement la quantité d'intérêt complexifie l'opérateur d'observation et introduit de nouvelles incertitudes, ces mesures globales peuvent être plus informatives que des mesures locales précises. Ceci est d'autant plus vrai dans des contextes où les erreurs de position sont importantes, comme la dispersion des polluants et la propagation des fronts de feux (Rochoux et al. 2014b; Zhang et al. 2019).

Pour réaliser une première expérience numérique d'un tel système dans un contexte de dispersion, nous imaginons le plan d'action suivant : i) générer une séquence d'images synthétiques à partir d'un modèle LES et d'un opérateur d'observation inverse pour représenter le lien entre la concentration de polluant et la quantité mesurée à distance, ii) reconstruire la forme du panache à l'aide d'un algorithme de détection de contour (Xin-Yi et al. 2018; Chu et al. 2022), cette tâche consiste essentiellement à dessiner le contour du panache sur la Fig. VI.4a, et iii) appliquer un algorithme d'assimilation de



(a) **(b)**

Figure VI.4: *Images aériennes d'expériences environnementales de terrain pertinentes. (a) Panache de chlore correspondant à l'essai #7 de l'expérience de terrain Jack Rabbit II (2016, US). Source : caméra embarquée par un drone opéré par les services d'urgence de l'université de Utah Valley. (b) Photo d'une expérience de brûlage contrôlé de bruyère dans le Northumberland (mars 2010, Royaume-Uni). Les lignes bleues, vertes et rouges indiquent les contours des iso-températures 500 K, 600 K et 700 K respectivement, telles qu'identifiées par Paugam et al. (2013) à l'aide d'une caméra thermique infrarouge héliportée.*

contours, comme celui proposé par Rochoux et al. (2018), pour corriger le contour estimé du panache avec celui observé. Une fois mis en place, ce système pourrait être validé par des mesures visuelles réelles à bord, comme celles de la campagne de terrain Jack Rabbit II (Fig. VI.4a).

En fin de compte, ces deux types de données d'observation fournies par les drones – images aériennes et mesures directes in situ – pourraient être assimilées conjointement par notre système d'AD. Cela pourrait permettre de mieux prendre en compte l'incertitude des modèles LES en corrigeant les erreurs de position et d'intensité, et donc d'améliorer la précision des prévisions de dispersion à micro-échelle.

Appendix A

Additional sensitivity tests of the LES model

The general aim of this appendix is to present further tests carried out to assess the validity, the robustness but also the limitations of the LES model used in the thesis. In particular, we investigate the convergence of the LES predictions with the mesh resolution in Sect. A.1. Then, we assess the model sensitivities to the computational domain height (Sect. A.2) and to the addition of turbulence injection (Sect. A.3).

A.1 LES model mesh convergence

When using an LES model, it is of primary importance to validate the LES solution convergence with the spatial discretization resolution (Piomelli 1999). Indeed the coarser the mesh, the less the proportion of turbulence resolved (Fig. II.1, page 50), and so the more room for model inaccuracy, as a larger part of the turbulent scales is modeled and not simulated. To address this issue, it is recommended to perform a mesh sensitivity analysis which consists of using several meshes with increasing spatial resolution and establishing from which resolution the model estimates become independent of the mesh resolution (Franke et al. 2007; Schatzmann et al. 2010; Blocken 2015).

To assess the mesh convergence of the LES model of the MUST case presented in Chapter II, simulations are performed on two additional meshes: one coarser than the reference one defined in Sect. II.4.1, page 61, and one with a more refined spatial resolution (see Table A.1). These two meshes are obtained by rescaling the cells of the non-structured mesh by a uniform and isotropic factor λ_{mesh}, using the Mmg[1] library.

Figure. A.1 shows the vertical profiles of three quantities of interest at tower T (located in the canopy) for the three mesh resolutions: mean horizontal wind velocity, horizontal turbulent kinetic energy and mean concentration. Note that tower T is located within the canopy. The profiles do not overlap perfectly, especially for the horizontal turbulent kinetic energy and the mean concentration. However, most of the differences are of the order of magnitude of the internal variability error, as represented by the blue shaded areas

[1]See https://www.mmgtools.org/.

Table A.1: *Characteristics of the meshes used to assess the mesh convergence of the LES MUST model with different refinement factors λ_{mesh}. N_{cells}, and N_{nodes} are the numbers of tetrahedral cells and nodes in million; h_{min} (resp. h_{max}) is the minimum (resp. maximum) cell edge length; z_1 is the averaged height of the center of the first cell above the ground.*

	λ_{mesh}	N_{cells} ($\times 10^6$)	N_{nodes} ($\times 10^6$)	h_{min} (m)	h_{max} (m)	z_1 (m)
Coarse mesh	1.4	49	8.9	0.089	12.6	0.17
Reference mesh	1.0	91	17	0.118	10.7	0.12
Refined mesh	0.8	161	28	0.064	8.68	0.10

in Figure. A.1. Note at high altitudes the profiles associated with the tested meshes are outside of the internal variability confidence intervals, but internal variability is probably under-estimated in altitude (Fig. III.12, page 106).

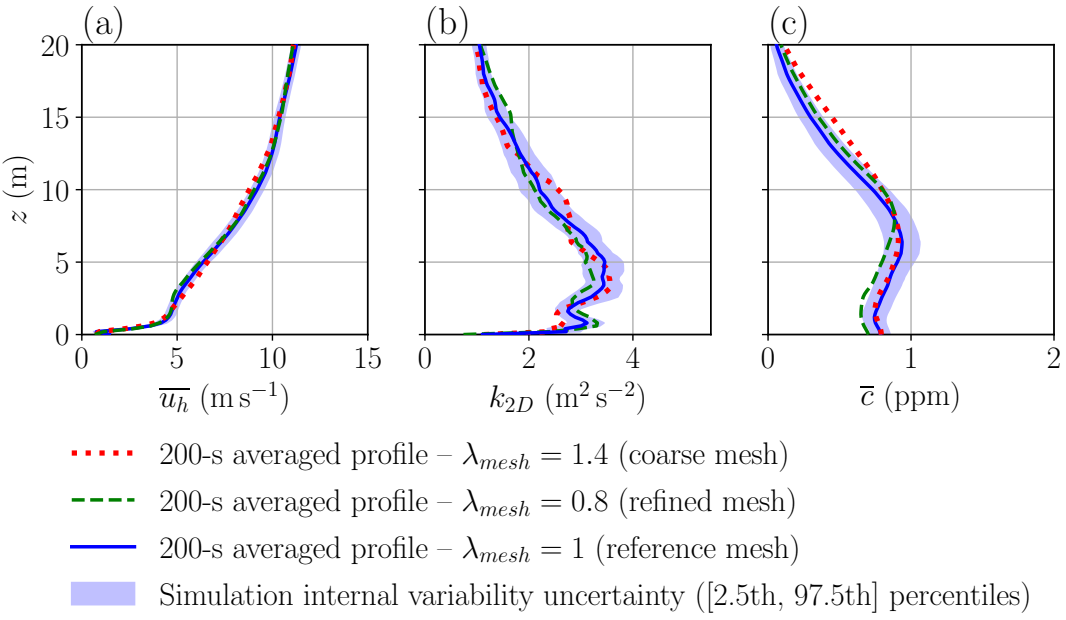


Figure A.1: *Vertical profiles of mean horizontal velocity (a), horizontal turbulent kinetic energy (b), and mean concentration (c) at tower T. Results are given for the simulations performed with the reference mesh, the coarse mesh, and the refined mesh (Table A.1), as blue solid lines, dotted red lines and dashed green lines, respectively. Internal variability uncertainty is represented with the [2.5th, 97.5th] percentile intervals estimated by stationary bootstrap presented in Sect. III.2.4, page 83 as shaded blue areas for the reference simulation.*

In conclusion, given the inherent uncertainty of LES predictions over time-finite periods, we can consider the LES estimates are not significantly impacted by the mesh resolution. We can conclude that the reference mesh is sufficiently well refined to reach spatial discretization convergence in such an uncertain context.

A.2 LES model sensitivity to computational domain height

For CFD modeling of atmospheric flows, it is important to have a domain high enough to limit the artificial impact from the top boundary condition on the area of interest and to limit flow blockage. As described in Sect. II.4, page 61, the baseline computational domain used in this study matches the usual recommendations for the domain height (Tominaga et al. 2008; Franke et al. 2007). Nevertheless, during the validation of the current LES model (Sect. III.4.1, page 92), it was found that the turbulent kinetic energy was underestimated near the top of the domain (Fig. III.6, page 93). We investigate in this section, the effect of the domain height on the turbulent fluctuations predicted over the canopy.

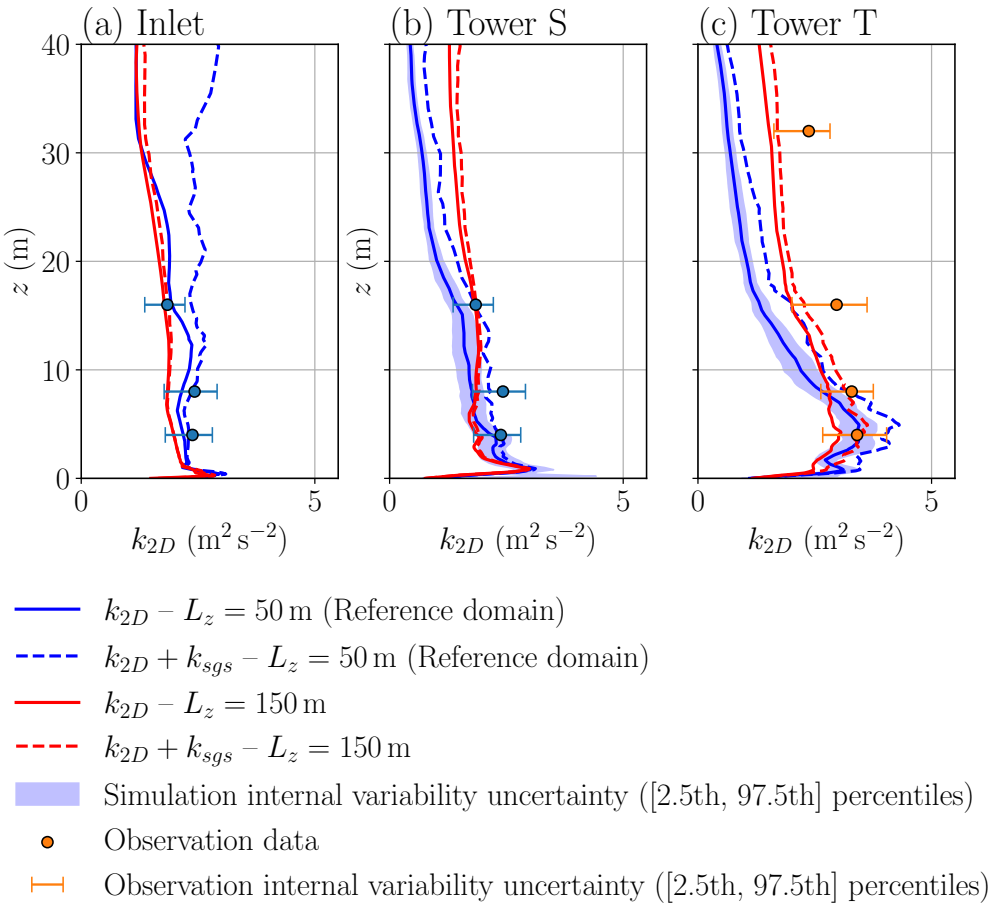


Figure A.2: *Vertical profiles of horizontal turbulent kinetic energy at the inlet, tower S and tower T for the reference domain height (50 m) and the one with extended height (150 m). The dotted lines represent the subgrid-scale turbulent kinetic energy estimated with Eq. A.1. Circle symbols correspond to the observations. Internal variability uncertainty is represented with the [2.5th, 97.5th] percentile intervals estimated by stationary bootstrap as shaded blue areas for the reference simulation and as orange bars for the observations.*

For this purpose, the computational domain height is increased from 50 m to 150 m. A new mesh, with the same properties as the reference one (see Sect. II.4.1, page 61) is generated to discretize the new domain. In this new domain, the turbulent injection parameter λ_e, i.e., the most energetic length of the turbulent spectrum, is multiplied by three as it was previously limited by the domain height. A new simulation is carried out with this setup. Figure A.2 shows that increasing the domain height improves the overall model accuracy at high altitudes. In particular, it solves the problem of the turbulent kinetic energy dissipation for heights between 20 m and 40 m. This demonstrates that it is the outflow boundary condition used for the top of the domain that dissipates turbulence. We also find that, by reducing the error on the turbulent kinetic energy profile and enabling the injection of larger structures, the simulation with increased domain height better estimates the internal variability in altitude compared to the baseline simulation.

Note that the effect of domain height on the mean concentration is not significant when compared to internal variability errors (not shown here). This is because most of the plume is located near the ground.

Note also that, for the same MUST trial and using a different LES code (Meso-NH, (Lac et al. 2018)), Nagel et al. (2022) found no significant differences in the vertical profiles predicted by two simulations with different domain heights of 40 m and 3000 m, respectively. This is because the top boundary condition treatment they use is a free-slip boundary condition. In contrast, we use an outflow boundary condition that imposes zero vertical turbulent transport and thus has more effect on the turbulent statistics within the domain, as suggested by Calaf et al. (2011).

To analyze the effect on the turbulent kinetic energy in more detail, we also take into account the subgrid-scale turbulent kinetic energy k_{sgs}, which represents the part of the turbulent energy that is modeled and not resolved by the LES (Fig. II.1, page 50). We estimate it from the subgrid-scale turbulent viscosity ν_t as in Quillatre's thesis (2014):

$$k_{sgs} = \left(\frac{\nu_t}{C_m \Delta} \right)^2, \tag{A.1}$$

with Δ the cell size, and $C_m = 0.091$. Figure. A.2 shows the k_{sgs} profiles added to the horizontal turbulent kinetic energy. Interestingly, increasing the domain height reduces the estimated subgrid-scale turbulence. Since the mesh resolution is the same for both domains, it implies that the reference simulation overestimates subgrid-scale viscosity.

In conclusion, the finite vertical extent of the domain interferes locally with the predictions of turbulent statistics because of the outflow boundary condition used at the top of the domain. In particular, the top boundary condition leads to a dissipation of turbulent kinetic energy down to approximately $z = 15$ m. As in Calaf et al. (2011), we find that the height of the domain should be at least three to five times the height of the area of interest. Note that this is not an issue for computational cost as the upper part of the domain contains much fewer cells when using unstructured mesh or non-conforming structured mesh with refinement. Unfortunately, we could not change the model configuration for the thesis because the ensemble of LES (Sect. IV.4, page 137) was already computed with a domain height of 50 m. Therefore we should be careful with the model predictions above the canopy knowing that they are imperfect. Note this is not a ma-

jor issue for dispersion analyses since all the concentration sensors are located near the ground (under 6 m).

To check whether similar problems or horizontal flow blockage occur with lateral boundary conditions, we also carried out a new simulation with an increased horizontal extent (with a distance between the first obstacle and the lateral boundaries increased from 80 m to 920 m). Results show that increasing the domain horizontal extent has no significant effect on the wind flow statistics predicted by the LES model compared with the internal variability involved. We conclude that our baseline LES model is not artificially affected by the lateral boundary conditions.

A.3 Impact of adding turbulence injection on LES predictions

One of the modeling contributions of this thesis is the implementation of a turbulence injection method to improve the physical representativeness of the AVBP model (Sect. II.4.3, page 64). In this section, we assess to what extent this addition influences model predictions.

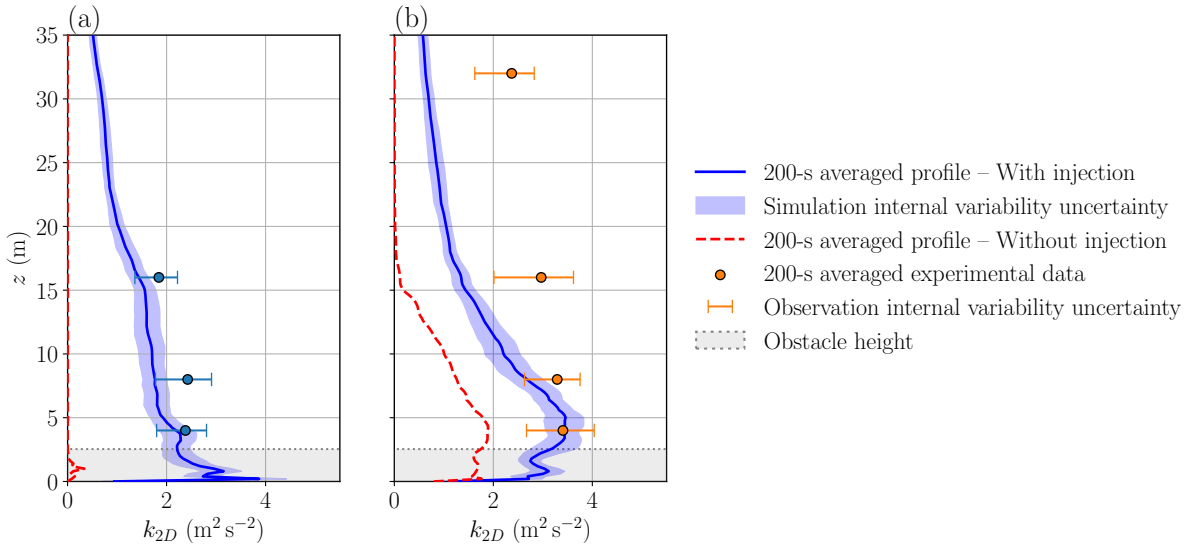


Figure A.3: *Vertical profiles of the horizontal turbulent kinetic energy at tower S (a) and T (b). Simulation time-averaged profiles are given with and without turbulence injection, as blue and red lines respectively. Colored circles correspond to observations. Internal variability uncertainty is represented with the* [2.5*th*, 97.5*th*] *percentile intervals estimated by the stationary bootstrap method presented in Sect. III.2.4, page 83, as shaded blue areas for the reference simulation (with injection), and orange error bars for the observations.*

This is done by comparing two simulations: one with turbulence injection and one without. As expected, not using turbulence injection results in a drastic underestimation of the horizontal turbulent kinetic energy k_{2D} compared with both observations and

simulation with turbulence injection (Fig. A.3). For tower S, located upstream of the containers, there is no turbulence at all over the entire vertical (red line Fig. A.3a). While at tower T, positioned after the fifth line of containers, some of the turbulence induced by the containers is captured by the LES model without injection up to 15 m above the ground. Still, the total turbulent kinetic energy is underestimated compared with observations (Fig. A.3b), even in the obstacle region where one could expect a lesser impact of the incoming atmospheric turbulence. This shows that the turbulence brought by the atmospheric boundary layer is not negligible in this part of the simplified urban canopy, at least for the studied trial.

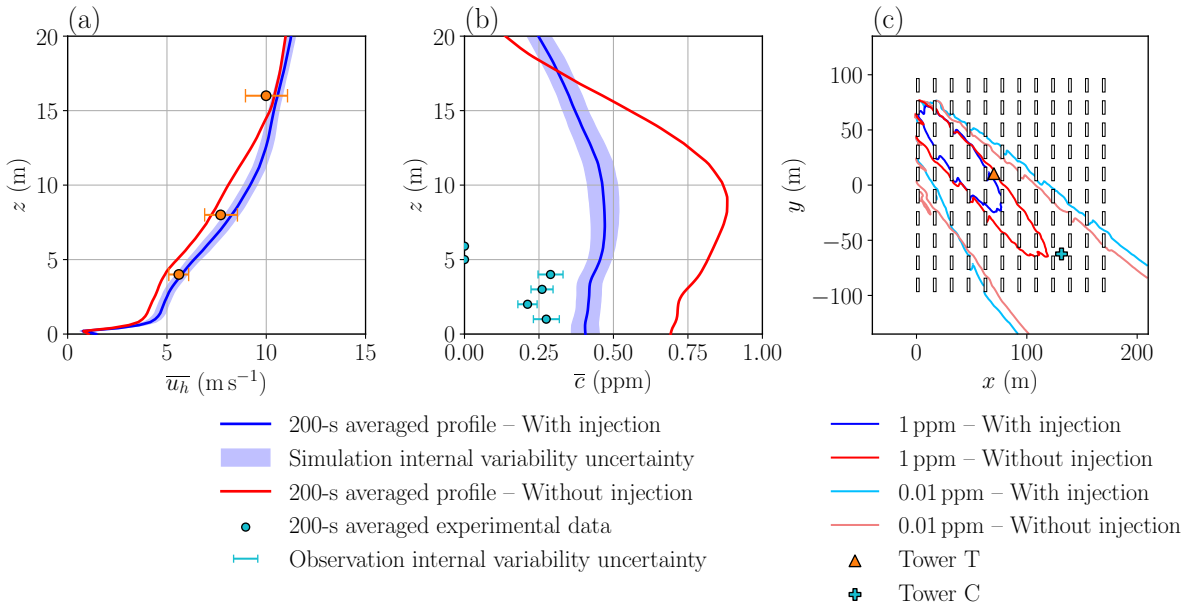


Figure A.4: *Vertical profiles of mean horizontal wind velocity at tower T (a) and mean concentration at tower C (b). Mean concentration isolines on the horizontal plane at $z = 1.6m$ (a). Simulation time-averaged profiles are given with and without turbulence injection, as blue and red lines respectively. Colored circles correspond to observations. Internal variability uncertainty is represented with the [2.5th, 97.5th] percentile intervals estimated by stationary bootstrap presented in Sect. III.2.4, page 83, as shaded blue areas for the reference simulation (with injection), and error bars for the observations.*

This underestimation of turbulent energy causes the flow in the canopy to slow down (Fig. A.4a), which reduces tracer mixing and thereby increases concentration levels. This is evidenced by the longer plume in the streamwise direction obtained without injection (Fig. A.4b). Compared to field measurements, the simulation without turbulence injection underestimates mean horizontal wind velocity at tower T (Fig. A.4a), and overestimates mean concentration, as shown by the vertical profiles of mean concentration at tower C (Fig. A.4b). Note that towers T and C are located within the canopy (Fig. A.4c). In addition, we highlight that these changes in the LES predictions are not caused by internal variability as the mean profiles obtained without injection are outside of the 95% confidence intervals representing the internal variability uncertainty (Fig. A.3–A.4).

In conclusion, this sensitivity test demonstrates that turbulence injection has a preponderant effect on LES predictions of wind flow and pollutant concentration statistics, even after a few lines of obstacles. In particular, not using turbulence injection deteriorates the agreement between LES and field measurements, even inside the urban canopy. This supports the use of turbulence injection for LES of microscale atmospheric flows, as already suggested by Breuer (2007) and Vasaturo et al. (2018).

Appendix B

Reduced-order model additional applications

This appendix presents additional applications of the POD–GPRs (Proper Orthogonal Decomposition and Gaussian Process Regressors) reduced-order model designed and validated in Chapter IV for predicting mean concentration fields for varying inlet wind direction and friction velocity. In Sect. B.1, we demonstrate that POD–GPRs can be used to accurately predict other LES fields such as mean horizontal velocity, turbulent kinetic energy and maximum concentration. In complement, Section B.2 illustrates how to use the reduced-order model for global sensitivity analysis. Finally, we investigate in Sect. B.3 an approach called mixture-of-experts that combines different POD–GPRs models to improve the accuracy of mean concentration predictions.

B.1 Prediction of other fields

In Chapter IV, the construction and validation of the POD-GPRs model are focused on mean concentration prediction. Here, we show that the POD-GPRs approach can be easily transposed to the prediction of other quantities.

We consider the problem of predicting three additional fields: the mean horizontal velocity u_h, the horizontal turbulent kinetic energy $k_{2\mathrm{D}}$, and the maximum concentration $c_{\max}$ over a 200-s analysis period. The first two could be typically used for data assimilation as shown by Defforge et al. (2021). In addition, emulating the turbulent kinetic energy can be used in promising multi-fidelity approaches for dispersion (Nony 2023). Finally, maximum concentration can be useful in an operational context for maximal exposure mapping. Concerning the preprocessing, we use a log-prescaling (Eq. IV.22) for $c_{\max}$ and linear centering (Eq. IV.21) for the other two fields. In all cases, we normalize fields by the friction velocity, as in Eq. IV.3a, b, for $c_{\max}$ and u_h, and by dividing by u_*^2 for $k_{2\mathrm{D}}$.

For these other fields, the POD–GPRs predictions are in very good agreement with the LES test fields, as shown by the horizontal cuts at $z = 1.6\,\mathrm{m}$ for the reference sample (Fig. B.1). The POD–GPRs model can capture the general appearance of the fields, namely the flow slowdown through the canopy (Fig. B.1a, b), the peak of turbulent

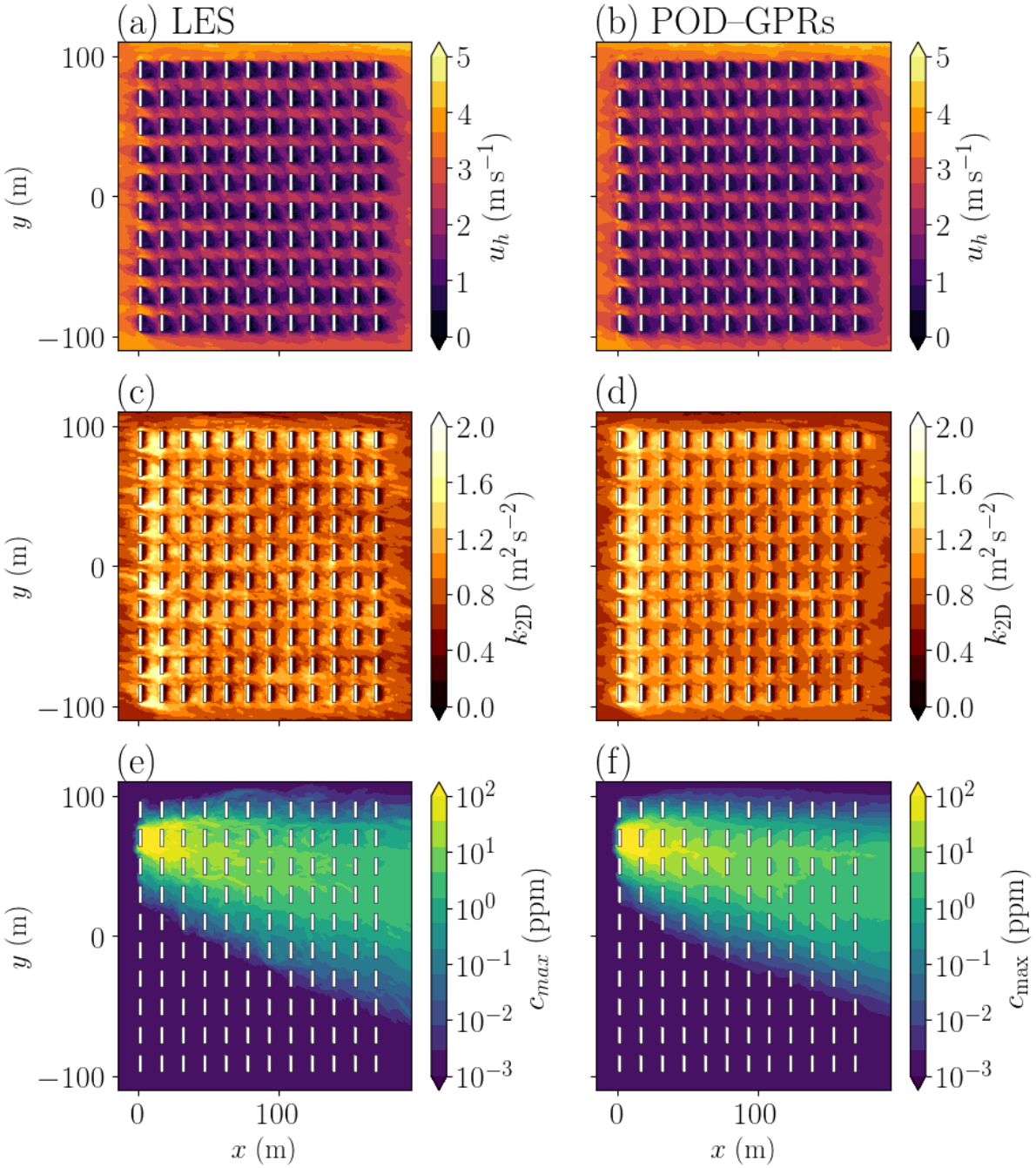


Figure B.1: *Horizontal cuts at $z = 1.6\,m$ of the mean horizontal velocity u_h (a, b), the horizontal turbulent kinetic energy k_{2D} (c, d), and the maximum concentration c_{max} over a 200-s analysis period (e, f). The first and second columns correspond respectively to the LES and POD–GPRs predictions. Results are given for the test sample #91 with $(\alpha_{inlet}^{(91)}, u_*^{(91)}) = (-8.9\,°, 0.38\,m\,s^{-1})$. The preprocessing used for each field is described in Sect. B.1.*

kinetic energy behind the first line of containers (Fig. B.1c, d), and the less elongated form of c_{max} isolines (Fig. B.1e, f). Careful examination shows that POD–GPRs may

however miss some local features of the reference fields, for example, local increase of k_{2D} within the obstacles array or the puff of concentration that goes outside of the array in Fig. B.1e. We presume that these local features are closely related to random fluctuations of the atmospheric boundary layer and therefore not generalizable.

Table B.1: *POD–GPRs prediction error for the mean horizontal velocity u_h, the horizontal turbulent kinetic energy k_{2D}, and the maximum concentration c_{max} over a 200-s analysis period. Errors are quantified using Q2 metric (Eq. IV.36), Mean Absolute Error MAE, and the air quality metrics presented in Sect. III.3.2, page 90. The metrics scores are then averaged over the test set.*

	Q2	MAE	FB	NMSE	FAC2	MG	VG	FMS(1 ppm)	FMS(0.01 ppm)
c_{max}	0.67	0.50 ppm	0.01	8.18	0.86	0.96	1.34	0.88	0.95
u_h	0.99	0.11 $m\,s^{-1}$	/	/	/	/	/	/	/
k_{2D}	0.97	0.07 $m^2\,s^{-2}$	/	/	/	/	/	/	/

We also proceed to a quantitative evaluation of the POD–GPRs accuracy over the test set. First, we verify that POD–GPRs reproduce well the variance of the test set as shown by the Q2 scores in Table B.1. Note that the less good Q2 score for c_{max} is mainly due to the log-preprocessing that limits the ability of the POD to capture variance on high values, as demonstrated in Sect. IV.5.2. As Q2 scores may fail at detecting non-physical predictions (see Table IV.2, page 148), we also evaluate the POD–GPRs accuracy using the Mean Absolute Error (MAE) and the standard air quality metrics for c_{max}. Results show a very fine level of accuracy for u_h and k_{2D}. Concerning, c_{max} the validation scores are slightly less good than for the mean concentration but are still in very good agreement, especially for the shape of the maximum concentration plume. As for the mean concentration, the model has difficulty predicting high maximum concentrations, as illustrated by the average NMSE in Table B.1. Note that these already very fine results are suboptimal since we use the same model design as for the mean concentration, i.e. the choice of the model preprocessing, the number of modes, and the kernel function. However, there is no a priori reason why these choices should remain optimal when predicting u_h, k_{2D}, and c_{max}. If the prediction of other fields is of particular interest, we recommend repeating the analyses presented in Sect. IV.2, page 122, to choose these parameters.

Finally, note that the reduced-order models predicting a given field are built independently. Building a single model, that could learn the underlying relationships between the fields predicted by the LES model, is an interesting prospect but has not been considered in this study.

B.2 Application to global sensitivity analysis

In this section, we give the results of a global sensitivity analysis of the LES dispersion model of the MUST case presented in Chapter II. These results are a continuation of the preliminary One-At-a-Time sensitivity analysis presented in Chapter III. In this first step, we have identified the parameters to which the LES model is most sensitive: the inlet wind direction α_{inlet} and the friction velocity u_*. In the current study, we go further by assessing in detail the global sensitivities of the LES model to these two parameters. To take into account non-linearities in the model surface response and investigate the coupled effect of the parameters, we use a stochastic approach based on the computation of the Sobol' indices presented in Sect. B.2.1. Since it requires a large number of model integrations we use the POD–GPRs reduced-order model built in Chapter IV as a surrogate for the LES model. Finally, results on the sensitivity of the mean concentration field are given in Sect. B.2.2.

B.2.1 Sobol' indices for sensitivity analysis

The calculation of Sobol' indices is a stochastic sensitivity analysis method that has the advantage of taking into account both the non-linearities of the model and the effects of interactions between parameters (Sobol' 1990). To quantify the sensitivity of a model output $\mathbf{y}$ to a certain number d of input variables $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^{d}$, the Sobol' indices are based on the decomposition of the total observed variance $V = \mathbb{V}(\mathbf{y})$:

$$V = \sum_{i=1}^{p} V_i + \sum_{1\leq i<j\leq p} V_{ij} + \ \ldots \ + V_{1\ldots p}, \tag{B.1}$$

$$\begin{aligned}
\text{with} \qquad V_i &= \mathbb{V}\left(\mathbb{E}(\mathbf{y}|\theta_i)\right), \\
V_{ij} &= \mathbb{V}\left(\mathbb{E}(\mathbf{y}|\theta_i,\theta_j)\right) - V_i - V_j, \\
V_{ijk} &= \mathbb{V}\left(\mathbb{E}(\mathbf{y}|\theta_i,\theta_j,\theta_k)\right) - V_i - V_j - V_k - V_{ij} - V_{ik} - V_{jk}, \\
&\vdots \\
V_{1\ldots p} &= V - \sum_{i=1}^{p} V_i - \sum_{1\leq i<j\leq p} V_{i,j} - \ \ldots \ - \sum_{1\leq i_1<\cdots<i_{p-1}\leq p} V_{i_1\ldots i_{p-1}}.
\end{aligned}$$

The first-order terms V_i quantifies the variance of $\mathbf{y}$ if it were conditioned solely by the input variable θ_i. Using the useful formula $V_i = \mathbb{V}(Y) - \mathbb{E}(\mathbb{V}(Y|X_i))$ it can also be interpreted as the expected decrease in the variance of Y when the variable θ_i is fixed. Second-order terms V_ij quantify the variance of $\mathbf{y}$ if it were conditioned solely by θ_i and θ_j after removing the associated first-order contributions. Higher-order terms are defined similarly.

Sobol' sensitivity indices are then simply defined by normalizing Eq. B.1 by the total variance:

$$1 = \sum_{i=1}^{p} S_i + \sum_{1\leq i\leq j\leq p} S_{i,j} + ... + S_{1,2,...,p}, \tag{B.2a}$$

$$\text{so that } S_i = \frac{V_i}{V}, \quad S_{ij} = \frac{V_{ij}}{V}, \ \ldots, \quad S_{1\ldots p} = \frac{V_{1\ldots p}}{V}. \tag{B.2b}$$

The indices S_i are referred to as first-order Sobol' indices, S_{ij} as second-order, and so on. In addition, the total sensitivity indices ST_i are introduced to quantify all the effects of a variable θ_i on the variance of the output, i.e. the direct effect and all cross-effects of order greater than or equal to 2. They are defined as the sum of all sensitivity indices relating to the variable θ_i:

$$ST_i = S_i + \sum_{j \neq i} S_{i,j} + \sum_{\substack{j \neq i \\ k \neq i,j}} S_{i,j,k} + ... + S_{1,2,..p} \tag{B.3}$$

In practice, these indices are computed using Monte-Carlo techniques to estimate the conditional variances in Eq. B.1. In the original method proposed by Sobol' (1990), this is done using a pseudo-random uniform sequence of samples of inputs $\boldsymbol{\theta}$ defined so that each sample shares k inputs with another sample, k being equal to the order of the indices to estimate. Other methods exist (McKay 1995; Saltelli et al. 2010) but they all rely on Monte-Carlo estimation. This implies that the convergence of the Sobol' indices estimates is in $\mathcal{O}(\sqrt{N})$ with N the number of model evaluations, which is unsuitable for the LES model because of its computational cost (see Chapter II). To circumvent this issue we use the reduced-order model built in Chapter IV as a surrogate for the LES model. Note that this is a classic strategy in sensitivity analysis (Sudret 2008; Cheng et al. 2020).

To write the variance decomposition as in Eq. B.1, the input variables are assumed to be decorrelated. This is not necessarily true in reality: for example, high friction velocities can be strongly correlated with certain wind directions. As this section is mainly an example of how to apply the reduced model, we do not investigate this aspect further. Note that if the variables were found to be correlated, we would have to use another sensitivity quantification approach, for example using the indices from Shapley (1953) instead of the Sobol' indices, as shown by Iooss and Prieur (2019).

B.2.2 Global sensitivity of the mean concentration predictions to wind conditions

In this section, we estimate the Sobol' indices of the mean concentration field $\overline{\mathbf{c}}$ of the LES model of the MUST field experiment (see Chapter II), for each node of the domain. To do so, we use the SALib[1] implementation of the algorithm proposed by Saltelli et al. (2010). In this method, the total number of input samples for which the model must be evaluated is equal to $N \times (2d + 2)$ with N the number of terms in the Monte-Carlo estimator and d the number of input variables. Here, we only consider the $d = 2$ most important input parameters according to the preliminary sensitivity analysis presented in Chapter III: the inlet wind direction α_{inlet} and the friction velocity u_*. It implies that it is not required to compute the indices of second order as they can simply be derived from Eq. B.2a, thus reducing the number of model evaluations to $4N$. In our case, a

[1]See `https://salib.readthedocs.io/en/latest/`

convergence study shows that $N = 2048$ is enough to reach convergence of the Sobol' indices estimates.

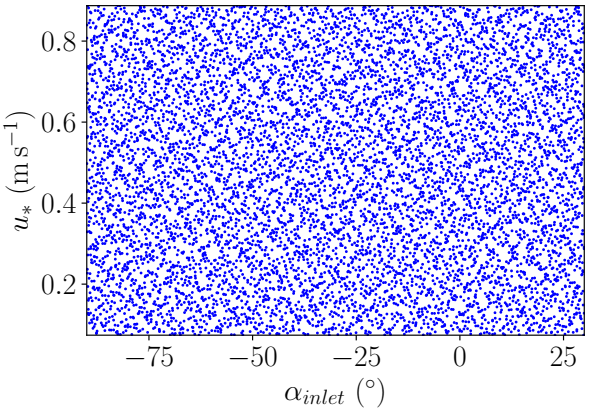


Figure B.2: *Input parameter space sampling obtained using the scheme from Saltelli et al. (2010), each point is a pair of parameters for which we evaluate the reduced-order model.*

To be able to carry out this large number of queries, we use the POD–GPRs reduced-order model presented in Chapter IV as a surrogate for the LES model. We use log-prescaling $\mathcal{T}_{\log}$ of the mean concentration field (see Sect. IV.2.4, page 128). However, we do not normalize the fields by the friction velocity as it would force the structure of dependency to this parameter. In opposition to the local One-At-a-Time sensitivity analysis presented in Chapter III, we now study the model response surface for the whole input parameter space $\Omega_{\boldsymbol{\theta}}$, as defined for the construction of the reduced-order model (Eq. IV.38, page 138). The sampling of $\Omega_{\boldsymbol{\theta}}$ obtained by the Saltelli sequence is illustrated in Fig. B.2. Note that the 2048 first samples of the sequence are skipped following recommendations from Campolongo et al. (2011) and Owen (2020).

Figure B.3 shows the spatial distribution of the Sobol' indices of the first and second order of the mean concentration at $z = 1.6\,\mathrm{m}$ within the array of obstacles. It appears that the variance of the model prediction is mainly explained by the inlet wind direction (Fig B.3a). Conversely, the first-order Sobol' indices relative to the friction velocity are overall very close to zero, except near the source and along one inter-obstacles axis (Fig B.3b). On the other hand, the second-order indices reach significant levels ($S_{\alpha_{inlet},u_*} \geq 0.5$), especially in a zone of about 5 obstacles by 5 obstacles around the source (Fig B.3c). This demonstrates that there is an important interaction effect between α_{inlet} and u_*. It is explained by the fact the mean concentration at one given location can be either inversely proportional to friction speed or independent of u_*, depending on whether or not the plume passes through this location. Hence the wind direction conditions the dependence on friction velocity.

At a given location, when the majority of the concentration samples from the Saltelli sequence (Fig. B.2) are close to zero, the computation of the Sobol' indices becomes ill-posed. The part of the domain where 99% of the concentration samples are under a threshold $c_t = 10^{-4}\,\mathrm{ppm}$ is highlighted by blue hatches in Fig. B.3. Note that the estimated Sobol' indices should also be treated with caution in the vicinity of the source,

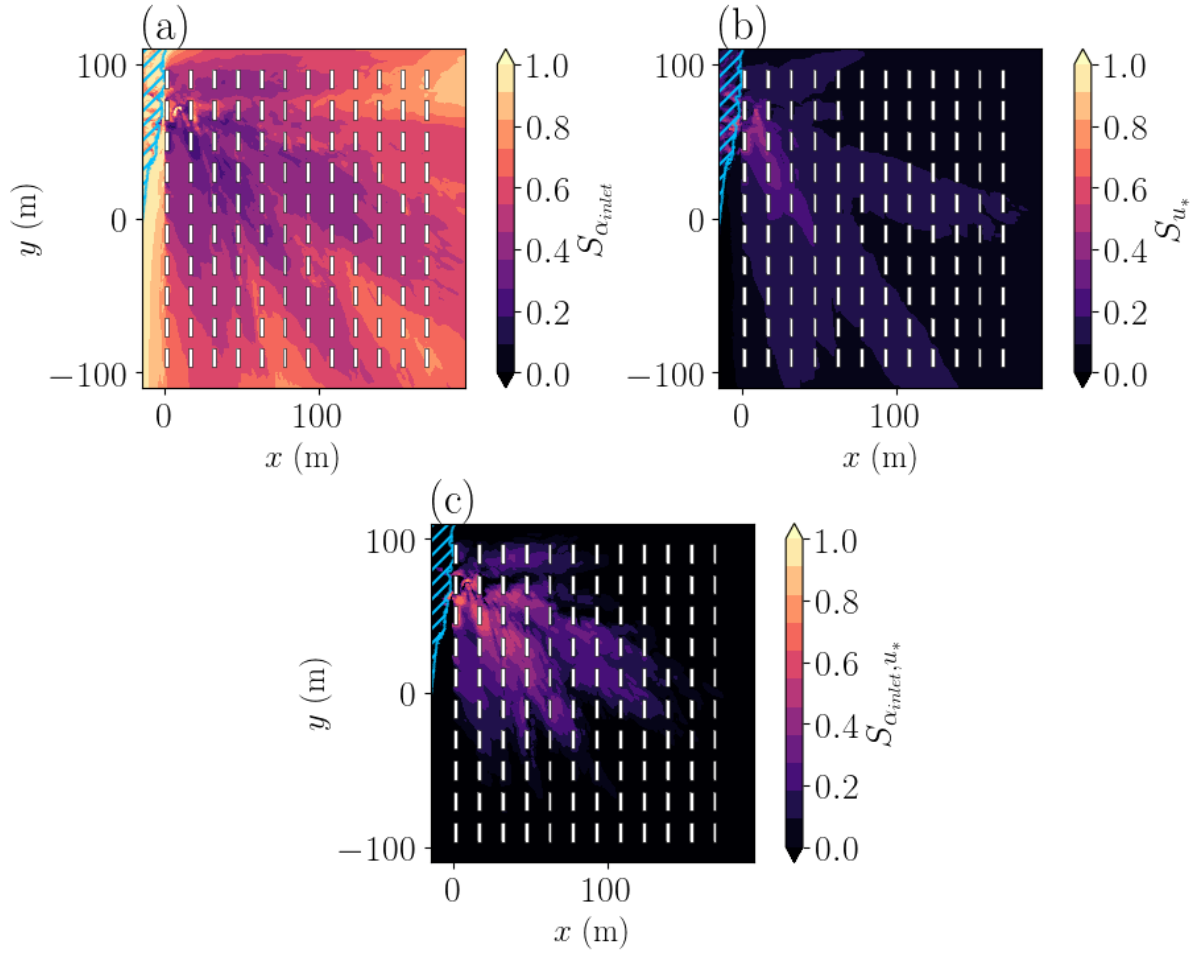


Figure B.3: *Horizontal cuts at $z = 1.6\,m$ of the maps of the Sobol' first-order indices quantifying the sensitivity of the mean concentration field to the wind direction α_{inlet} (a) and to the friction velocity u_* (b). The effect of interactions between these two parameters is quantified by the second-order Sobol' indices (c). The blue hatched area corresponds to mesh nodes for wich 99% of the concentration samples (Fig. B.2) are under the threshold $c_t = 10^{-4}\,ppm$.*

as we saw in Chapter IV that the POD–GPRs model lacks precision in this area.

Table B.2: *Global Sobol' indices quantifying the sensitivity of the mean concentration field to two wind condition parameters: the inlet wind direction α_{inlet}, and the friction velocity u_*. Results given in this table are spatial averages of the Sobol' indices of first order S_i, second order S_{ij} and total order ST_i, computed for each parameter.*

	S_i	S_{ij}	ST_i
α_{inlet}	0.69	0.25	0.94
u_*	0.06	//	0.31

Finally, to give an overview of the model's sensitivities, we give in Table B.2 the first,

second, and total order Sobol' indices averaged over the whole domain. Nodes for which the computation of the Sobol' indices is ill-posed are not taken into account in these scores. The aggregated indices thus obtained are consistent with observations from Fig. B.3. Indeed, the inlet wind direction α_{inlet} is the predominant input parameter with a very high global total-order indice $ST_{\alpha_{inlet}}$. And the mean concentration dependence on u_* is still mostly conditioned by α_{inlet} as $S_{\alpha_{inlet},u_*} > S_{u_*}$. We note that these results on the global sensitivity of the model surface response are coherent with those of the preliminary One-At-a-Time sensitivity analysis presented in Sect. III.5.2, page 109.

We highlight that Sobol' indices maps (Fig. B.3) can be used to guide the positioning of sensors during the design of experimental campaigns or observation networks. It is particularly useful in the context of data assimilation for parameter estimation as shown by Mons et al. (2017) in a very similar context. Indeed, these maps describe the effect of a perturbation of each input parameter on the observed quantity at each point in the domain. In practice for the current example, it tells us that to improve the chances of correctly inferring friction speed, it is best to place sensors close to the source. Conversely, to infer the inlet wind direction it is better to position sensors further away from the source, on the edges of the obstacle array. Finally, we point out that placing sensors only according to the Sobol' indices can be problematic, as areas with zero concentrations in most scenarios can obtain very high first-order indices, as illustrated by the hatched area in blue Fig. B.3. We therefore suggest using, in addition to Sobol' indices, a criterion on the probability of actually detecting the tracer.

B.3 Towards reduced-order modeling based on a mixture of experts

To overcome the poor performances of the POD–GPRs at predicting high mean concentrations (see Sect. IV.6.1), we investigate the use of a Mixture-Of-Experts (MOE) method. This method, inspired by the works of El Garroussi et al. (2020), broadly consists of using a mix of predictions of different reduced-order models to optimize the overall prediction accuracy.

In our case, we adopt a very pragmatic approach to assembly the predictions of two POD–GPRs models, one with linear fields preprocessing $\mathcal{T}$ (Eq. IV.21) and the other with log-transformation $\mathcal{T}_{log}$ (Eq. IV.22). This takes advantage of the fact that the POD built with $\mathcal{T}$ better represents the high concentration while using $\mathcal{T}_{log}$ better captures the shape of the plume and low concentrations, as shown in Table IV.2. The mix between the two predictions $\mathbf{c}_{\mathcal{T}}$ is $\mathbf{c}_{\mathcal{T}_{log}}$ is simply achieved by using the first for high concentrations and the second for low concentrations. In practice, this translates into a spatial juxtaposition of the two fields:

$$\mathbf{c}_{\mathrm{MOE}}(\mathbf{x}_k) = \begin{cases} \mathbf{c}_{\mathcal{T}}(\mathbf{x}_k), \text{ if } \mathbf{c}_{\mathcal{T}}(\mathbf{x}_k) > c_{mix}, \\ \mathbf{c}_{\mathcal{T}_{log}}(\mathbf{x}_k), \text{ else,} \end{cases} \qquad 1 \leq k \leq N, \tag{B.4}$$

with c_{mix} the mixture concentration level, which separates the two predictions. To choose c_{mix} we compute the averaged FMS (Eq. III.18, page 91) scores obtained by both POD–GPRs models for increasing levels of isoconcentration, and we select the level of concentration for which the prediction with linear scaling becomes better than with log scaling. This gives $c_{mix} = 5\,\mathrm{ppm}$.

Figure B.4 illustrates the principle of the MOE reduced-order model with mean concentration prediction errors for the reference test sample #91 (represented as a yellow star in Fig. IV.5). We can see that the MOE approach achieves to reduce the estimation error for high concentrations close to the source compared with the POD–GPRs model with $\mathcal{T}_{log}$ (Fig. B.4a, c, e). Note that the error magnitude near the source remains important. In our opinion, this is because this zone is highly uncertain by nature, due to the strong concentration gradients and the high internal variability close to the edges of the obstacles array. On the other hand, MOE uses the POD–GPRs prediction with $\mathcal{T}_{log}$ for the rest of the domain which prevents it from predicting the large unphysical concentration features obtained with $\mathcal{T}$ (Fig. B.4b, d, f).

The overall results of the MOE, with and without fields normalization by the friction velocity, are given in Table B.3. In both cases, the MOE model achieves the highest scores by capturing the best of both worlds, i.e. NMSE scores aligned with the ones of the POD–GPRs models using linear preprocessing ($\mathcal{T}$ and $\mathcal{T}_{1\mathrm{D}}$), and the scores obtained with log-transformation ($\mathcal{T}_{log}$ and $\mathcal{T}_{log-1\mathrm{D}}$) a for every other metric. These results are mainly because the validation metrics are highly spatialized, as this analysis formally demonstrates. Finally, we note that when normalizing the fields by the friction velocity (last row of the Table B.3) the improvement in NMSE is relatively less important because the NMSE score obtained by the POD–GPRs model using $\mathcal{T}_{log-1\mathrm{D}}$ is better in the first place.

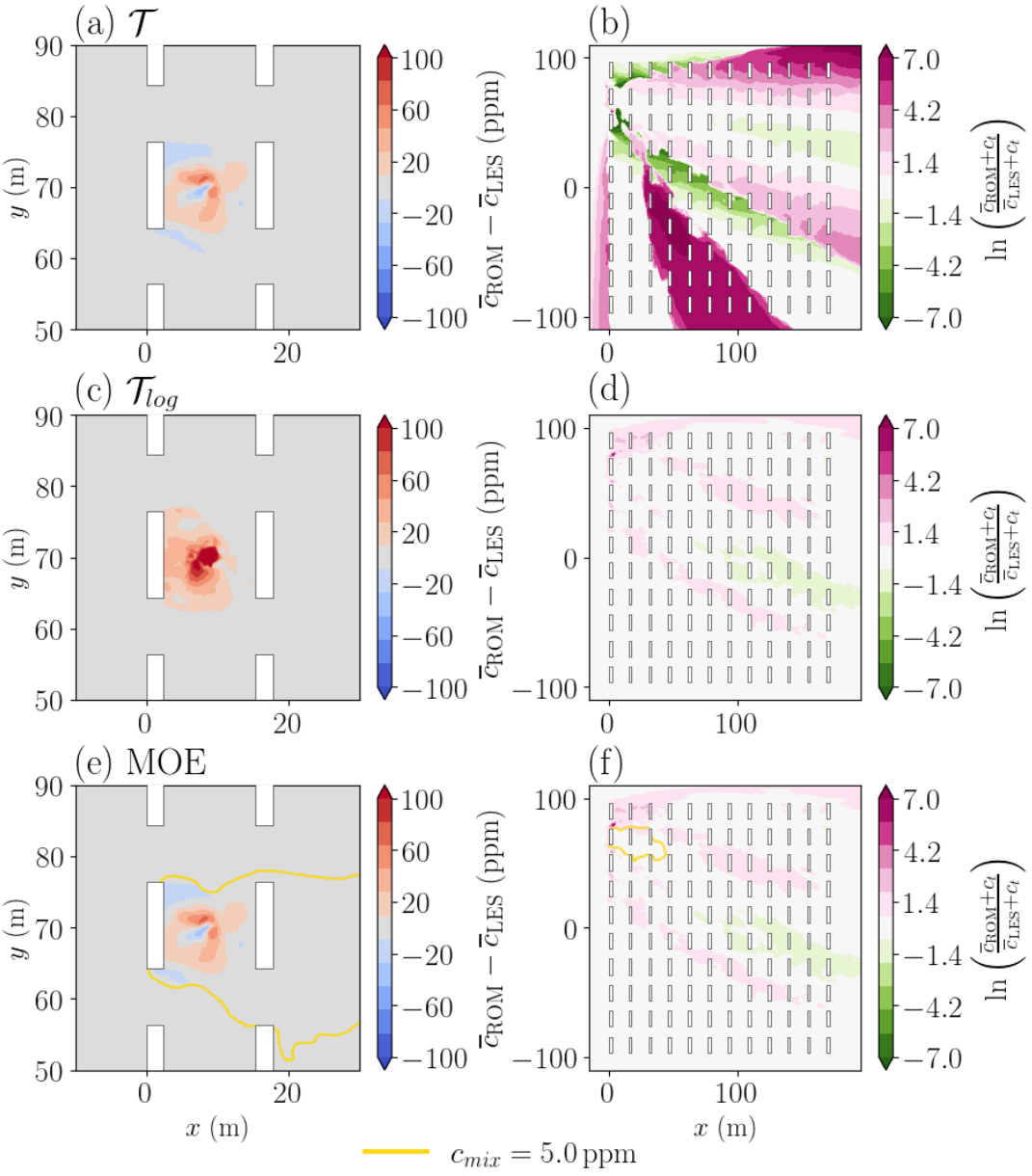


Figure B.4: *Horizontal cuts at $z = 1.6\,m$ of the difference between the LES and reduced-order model predictions of the mean concentration, for the POD–GPRs with the $\mathcal{T}$ (a, b), and $\mathcal{T}_{log}$ preprocessing (c, d), and for the MOE (e, f). The predictions are compared using absolute difference on the first column and log difference on the second column. The panels of the first column are a close-up view of the vicinity of the source. The yellow line corresponds to the 5-ppm isoline used to assemble the predictions (Eq. B.4). Results are given for the test sample #91 with $(\alpha_{inlet}^{(91)}, u_*^{(91)}) = (-8.9\,°, 0.38\,m\,s^{-1})$.*

One of the main issues with the MOE approach is that continuity both in spatial and parameter spaces is no longer ensured which results in concentration jumps at $c = c_{mix}$ (not shown here). On the contrary standard POD–GPRs models are continuous in the

Table B.3: *Comparison of the validation scores averaged over the test set for the POD–GPRs and MOE reduced-order models. Results for the POD–GPRs are given for the two fields preprocessing with and without log-scaling ($\mathcal{T}_{log}$ and $\mathcal{T}$). Definitions of the validation metrics are given in Sect. III.3.2, page 90. The second row corresponds to the mean level of error solely due to internal variability (see Sect. IV.3.4). The third and fourth rows correspond respectively to the results without and with fields normalization by the friction velocity.*

	FB	NMSE	FAC2	MG	VG	FMS (1ppm)	FMS (0.01ppm)
Internal variability	0	1.80	0.95	1.00	1.39	0.83	0.93
POD–GPRs – $\mathcal{T}$	-0.01	4.86	0.56	0.21	1.13×10^6	0.68	0.43
POD–GPRs – $\mathcal{T}_{log}$	-0.04	20.6	0.91	1.00	1.39	0.75	0.92
MOE	-0.03	4.41	0.91	1.00	1.39	0.75	0.92
POD–GPRs – $\mathcal{T}_{1D}$	-0.06	3.10	0.59	0.27	1.71×10^5	0.78	0.47
POD–GPRs – $\mathcal{T}_{log-1D}$	-0.02	4.61	0.90	0.97	1.40	0.79	0.93
MOE (1D)	-0.02	3.29	0.90	0.97	1.40	0.79	0.93

parameter space by construction since they are expressed as a linear transformation of continuous Gaussian processes predictions (Fig. IV.17) and the spatial continuity of the predicted fields is well verified in practice (Fig. IV.15c, d).

In conclusion, the MOE approach achieves to get the best out of both POD–GPR models, with and without logarithmic transformation and therefore offers a short-term solution for improving the accuracy of reduced-order models. To further improve the MOE approach, it would be pertinent to consider other reduced-order models for the choice of expert. We also believe that choosing the predicting model according to input parameters as El Garroussi et al. (2020) could improve the homogeneity of reduced-order model accuracy in parameter space (see Sect. IV.6.4). Finally, the model choice could be directly learned from the train set instead of being specified a priori in Eq. B.4. This classification learning problem could be solved using support vector machines (Platt 1999) or decision trees (Rokach and Maimon 2014).

Appendix C

Carbon footprint estimation

Introduction

Climate change is an urgent and unprecedented environmental issue that is reshaping the world as we know it. It refers to the long-term increase in the Earth's average surface temperature, primarily driven by the accumulation of greenhouse gases (GHG) in the Earth's atmosphere, such as carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O), which trap heat in the atmosphere, leading to a gradual rise in temperatures. This phenomenon has far-reaching and potentially catastrophic consequences, including the melting of polar ice caps, rising sea levels, the expansion of deserts, the intensification and increased frequency of heat waves and wildland fires, as well as the disruption of ecosystems and biodiversity. The effects of climate change are already being felt, with a 1.00 °C to 1.25 °C rise in average surface temperature over the 2013–2022 decade compared with the pre-industrial era (Forster et al. 2023), and GHG emissions related to human activities have undoubtedly played a central role in this warming (Masson-Delmotte et al. 2023).

In order to mitigate the effects of climate change and prevent the climate system from reaching tipping points, nations collectively committed, in the 2015 Paris Agreement, to keep warming well below 2 °C, and preferably as low as 1.5 °C, compared with pre-industrial levels. To limit global warming to 1.5 °C, humanity will need to cut emissions by half by 2030 and achieve near-zero emissions by 2050 (Rogelj et al. 2022). A fundamental first step in developing reduction strategies and committing human society to a low-carbon transition is the quantification of all anthropogenic emissions, as well as their classification by type of source (energy consumption, transport, industrial processes, etc). This assessment, called carbon footprint, can be carried out at different scales, so that individuals, companies and governments can make informed decisions to reduce their impact on the climate.

In this context, the world of research has also begun to think about its carbon footprint. In France, Labos 1point5, a collective of members from the academic world, has been at the forefront of estimating the carbon footprint of academic research. They notably develop and provide access to a carbon footprint estimation tool for laboratories:

GES 1point5[1], and initiated a multi-laboratory comparison (De Paepe et al. 2023). At CERFACS, a working group was formed to estimate the carbon footprint of the laboratory. During my thesis, I actively participated in the group's activities, including data collection and processing, as well as presenting our estimates. Figure C.1 gives an overview of CERFACS' carbon footprints for the years 2019, 2021 and 2022.

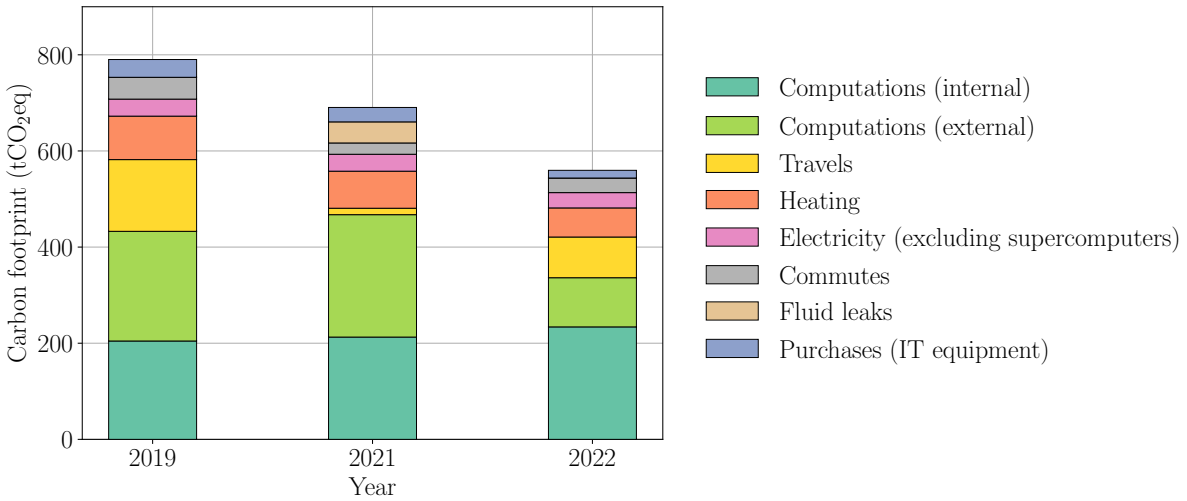


Figure C.1: *CERFACS' carbon footprints for the years 2019, 2021 and 2022. Each emission source is represented by a different color.*

The aim of this appendix is to adopt a finer degree of emissions estimation in order to assess the carbon footprint of the present thesis.

Emissions estimation methodology

To begin with, we need to define the scope of the thesis's carbon footprint. To do so, we decide to take into account the main emitting sources identified in CERFACS' carbon footprints (Fig. C.1):

1. computations emissions, which account for the use and life cycle of supercomputers,
2. emissions from travel to conferences, summer schools, etc.,
3. heating-related emissions, i.e. GHGs emitted directly by the gas-fired boiler used to heat CERFACS,
4. electricity emissions, which correspond to the emissions caused by the production and distribution of the energy consumed at CERFACS. In mainland France, the associated emission factor is estimated at $0.052\,\mathrm{kgCO_2eq\,kWh^{-1}}\,\pm 10\,\%$ in 2022[2],
5. commuting emissions,
6. fluid leaks, i.e. GHGs released directly in the atmosphere as a result of leaks, for example from the fire-fighting circuit in CERFACS' machine room. These leaks can have a considerable impact on the carbon footprint, for example, we estimate that

[1] Available at: https://apps.labos1point5.org/ges-1point5
[2] According to ADEME estimates, see https://base-empreinte.ademe.fr/

a leak of 400 kg of FM200 in 2013 had the same impact on climate change as the emission of 1500 t of CO_2,

7. purchases, which mainly take into account IT equipment in the scope of CERFACS' carbon footprint.

This scope includes both direct GHG emissions linked to heating or fluid leaks, and indirect emissions considered to be the most significant (emissions linked to electricity production, travels, commutes and purchases). Other indirect emissions, for example, related to catering, are not taken into account in the estimate presented here. Note that the comparison of the carbon footprints of a hundred French research laboratories proposed by De Paepe et al. (2023) highlights the same sources of emissions, but with a greater contribution from purchasing-related emissions.

In the following, we describe the approach adopted to estimate the GHG emissions associated with each of these items. Particular emphasis is placed on computation-related emissions, given their predominant impact on CERFACS' carbon footprint and their role in the thesis. Carbon impacts are expressed in CO_2 equivalent, which corresponds to the quantity of CO_2 that would cause the same cumulative radiative forcing over a given period of time.

1 - Computations

Estimation of computation emission factor Numerical simulation is at the core of the thesis project, and estimating the associated GHG emissions is therefore of primary importance. As part of CERFACS' carbon footprint assessment, a major effort was made to quantify the emissions of in-house computation. Thanks to a precise accounting of electrical and cooling consumption and of all the computing hours performed, we have estimated the emission factor f_{use} to be of the order of 1.1 gCO_2eq per core hour of calculation at CERFACS. Information was also gathered from the major computing centers we work with, in order to ascertain their emission factors. In cases where we did not receive a reply, we took an average emission factor corresponding to supercomputers of similar size and generation.

However, this emission factor only covers supercomputer energy consumption, and it is important to also take into account the emissions linked to the supercomputer life cycle, i.e. to include manufacturing, transport and recycling, in the emission factor. This accounting is difficult because i) each supercomputer is custom-built and made up of thousands of components, and ii) it requires that all the actors in the supercomputer's life cycle have also estimated their emissions and are willing to share them. Once estimated, the total emissions can be divided by the expected lifetime of the machines, typically around 6 years, then divided each year by the number of core hours of computation performed. This makes it possible to express an emission factor $f_{indirect}$ analogous to that for supercalculator energy consumption f_{use} and to distribute life-cycle emissions to users. We did not have the resources to carry out this study for CERFACS supercomputers, and we found only two estimates of such emissions in France: for a modest-sized supercalculator (Berthoud et al. 2020), and for one partition of a national computing center (private communication). Given these results and the large uncertainties involved,

we estimate that the indirect emissions linked to the life cycle of supercomputers are of the same order of magnitude as those linked to their use: $f_{indirect} \approx f_{use}$. In the end, the total emission factor of computation at CERFACS is estimated at

$$f_{total} = f_{use} + f_{indirect} \approx 2.2\,\mathrm{gCO_2eq\,h_{CPU}}^{-1}. \tag{C.1}$$

Using this emission factor, we can estimate by proportionality the GHG emissions of the simulations carried out in this thesis.

Accounting of computations made during the thesis Soon after the start of the thesis, we began to track all the simulations carried out with AVBP[3] as part of the thesis, allowing us to a posteriori estimate the emissions associated with each simulation, based on their durations and on the supercomputer used. It should be noted that the emission factors of national supercomputers (Occigen, Irene and Belenos) are around twice as low as that estimated by CERFACS (Eq. C.1), due to the economies of scale achieved on these larger supercomputers. Figure C.2 shows the history of core hours consumed on the various supercomputers used during the thesis. The generation of the LES ensemble used to build the reduced-order model presented in Chapter IV, which was carried out in two stages, in July 2022 and December 2022, appears to be the most important part of the computation carried out during the thesis.

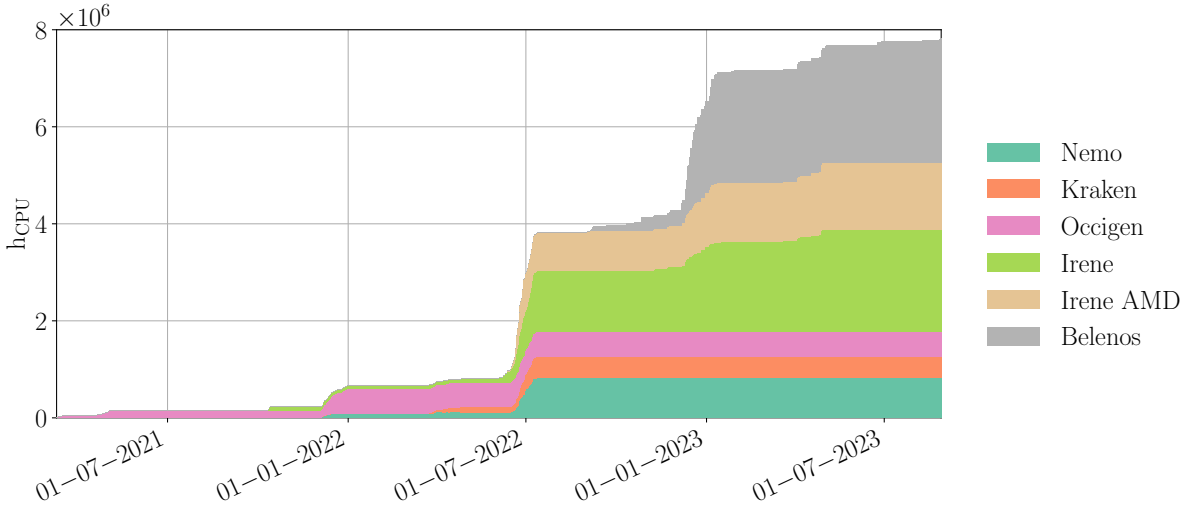


Figure C.2: *History of core hours consumed by AVBP simulations carried out during the thesis (from 10–03–2021). Each color corresponds to a different supercomputer: Nemo and Kraken from CERFACS, Occigen from CINES (Centre Informatique National de l'Enseignement Supérieur), Irene from TTGC (Très Grand Centre de Calcul), and Belenos from Météo-France. Irene AMD designates a specific partition of the Irene supercomputer (c.f. Table IV.1, page 142). Access to Occigen and Irene resources was granted by GENCI (Grand Équipement National de Calcul Intensif) as part of the DARI project A0062A10822, 2020–2022.*

[3]See footnote 1, page 47.

Using the `sacct`[4] Slurm command, we estimate that around 1 million of core hours have been spent on CERFACS' computers in addition to the simulations presented in Fig. C.2. This substantial amount includes simulations carried out during model familiarization before the start of simulation tracking, simulations performed with other codes not presented in this thesis, and the majority of post-processing and data exploitation work, including bootstrap applications (Chapter III), data assimilation (Chapter V) and sensitivity analysis (Appendix B).

The total computation-related GHG emissions of the thesis are given in Table C.3. It should be noted that this estimate does not include emissions related to data storage because i) it requires significantly less energy than computation (in the context of this thesis), and ii) it is much more difficult to attribute emissions from storage servers to each user. Emissions linked to data transfer are considered negligible based on the study of Ficher et al. (2021), which estimates the emissions associated with transferring 1 Go of data over the RENATER[5] fiber optic network between Orsay (France) and Montpellier (France) at 1.4 gC02eq.

2 - Purchases (IT equipment)

Concerning purchases related to this thesis, we only take into account purchases of IT equipment as i) the thesis did not require any specific experimental resources, ii) emissions related to the manufacture of supercomputers are included in computation-related emissions, and iii) we estimate that emissions related to other supplies (notebooks, pencils, etc.) are negligible. Computer equipment purchased during the thesis period and the related emissions are listed in Table C.1.

Table C.1: *GHG emissions linked to IT equipment purchased during the thesis period, in CO_2 equivalent.*

Purchase	Emissions ($kgCO_2e$)
MacBook Pro 13"	256
Dell P2419HC 24" monitor	430
LG HDR UHD 4K 27" monitor	430

These emissions are estimated using Ecodiag[6] tool developed by the EcoInfo service group of the CNRS and cover impacts linked to the manufacture and transport of equipment purchased. Both screens have the same carbon impact since they are similar in size and the emissions estimated by the tool are directly linked to screen size. We note that one screen has a greater carbon impact than a laptop computer. It is also worth noting that for the laptop, manufacturing is estimated to contribute over 80% of total GHG emissions (Ferreboeuf et al. 2018).

[4] See https://slurm.schedmd.com/sacct.html

[5] *Réseau national de télécommunications pour la technologie, l'enseignement et la recherche.*

[6] Available at https://ecoinfo.cnrs.fr/ecodiag-calcul/

3 - Travels

For travel-related emissions, only three main travels were made by the PhD student during the course of this thesis, to attend:

- the 21st International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (HARMO 21) in Aveiro (Portugal) in September 2022,
- a Lecture Series on CFD for atmospheric flows and wind engineering organized by the von Karman Institute for Fluid Dynamics near Brussels (Belgique) in May 2023,
- a drone field campaign in Lannemezan (France).

For each trip, emissions are calculated using the GES 1point5 tool, which estimates the travel distance and associated emissions based on the means of transport. In line with ADEME[7] guidelines, we decide to take into account the non-CO_2 radiative effect of contrails, which almost doubles the emission factor of air transport (Vieira Da Rocha and André 2021). The resulting emissions are given in Table. C.2.

Table C.2: *Carbon impact of travel undertaken during the thesis period, in CO_2 equivalent.*

Travel	Means of transport	Distance (km)	Emissions ($kgCO_2e$)
Toulouse ↔ Lisbon	plane	2266	425
Lisbon ↔ Aveiro	train	524	19
Toulouse ↔ Brussels	train	2004	32
Toulouse ↔ Lannemezan	car	262	57

As expected, air travel is responsible for the majority of travel-related emissions. Note that there is no fast train connection between Toulouse and Aveiro, whereas, for a similar distance, the trip to Brussels shows that, when the infrastructure is in place, emissions linked to travel can be divided by 10 by choosing the train over the plane. We should mention that travel-related emissions reported in Table C.2 do not account for travel by thesis supervisors during the course of the thesis.

4 - Commutes

We also use the tool developed by Labos 1point5 to estimate the emissions associated with daily commuting based on a typical week. Only the PhD student's commute is taken into account: 6km twice a day by bike, 5 days a week. This gives around 18 $kgCO_2eq$ after 3 years, which is negligible compared to other sources of emissions, as shown in Table C.3.

5 - Laboratory operation

For the three remaining emissions sources: heating, electricity consumption and fluid leaks, we use laboratory-level estimates divided by the number of people working at

[7] *Agence de l'environnement et de la maîtrise de l'énergie.*

CERFACS and assessed over the period October 2020 - November 2023. The thesis footprint is assigned a share of emissions, considering only that of the PhD student. Note that we do not include emissions linked to supercomputer electricity consumption, as this has already been taken into account in the computation emissions. As the year 2023 is not yet over at the time of estimating this carbon footprint, we interpolate electricity and heating consumption from previous years. Only one fluid leak occurred during the thesis period (27 kg of R407C in 2019, equivalent to 44 tCO_2eq[8]).

Emissions linked to heating, electricity consumption and fluid leaks thus estimated are given in Table C.3. These emissions are irreducible on the scale of the thesis since they correspond to the operation of the laboratory. However, a strategy to reduce emissions is currently being considered by CERFACS.

Thesis carbon footprint estimate

The total thesis carbon footprint is finally estimated to be of the order of 15 tCO_2eq. For comparison, 1 tCO_2eq represents approximately the emissions associated with a return flight from Paris to New York[9] and 10 tCO_2eq the annual footprint of a French person in 2022[10]. The annual footprint of this thesis is around 5 tCO_2eq, which is far more than the 2.3 tCO_2eq per capita in France estimated to be required by 2050 to limit global warming to 1.5 °C (Fouré et al. 2020).

Table C.3: *Carbon footprint of the thesis and details of the carbon impact of each sector considered, in CO_2 equivalent. GHG emissions related to the production of the electricity consumed by supercomputers are included in computation-related emissions.*

Source	Emissions (tCO_2eq)
Computations	10.4
Purchases (IT equipment)	1.1
Travels	0.5
Commutes	1.8×10^{-2}
Heating	1.3
Electricity consumption (excluding supercomputers)	0.6
Fluid leaks	0.3
Thesis carbon footprint	14.2

Table C.3 and Figure C.3 show the contribution of the various emission sources considered to the carbon footprint of the thesis. Computations account for by far the largest

[8] According to the emission factor used in GES 1point5 (https://apps.labos1point5.org/ges-1point5).

[9] According to the GES 1point5 tool. See https://apps.labos1point5.org/travels-simulator.

[10] Estimated by the French Ministry of Ecological Transition and Solidarity, and updated in 2022 by Carbone 4 (https://www.carbone4.com/myco2-empreinte-moyenne-evolution-methodo).

share of the thesis' GHG emissions (73%), yet this item is difficult to reduce since it is at the heart of the thesis project. Emissions linked to laboratory operations are also significant and difficult to reduce at the thesis level (15%). Emissions linked to purchases and travels, the two main areas of emissions for research laboratories in France (De Paepe et al. 2023), have been significantly limited in this thesis (1.6 tCO_2eq).

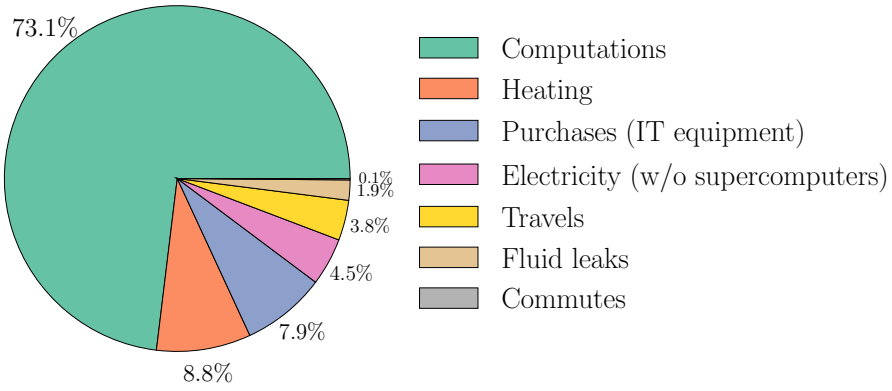


Figure C.3: *Breakdown of contributions to the thesis carbon footprint.*

Finally, it is important to note that there are various sources of uncertainty in this carbon footprint estimate. On the one hand, emissions linked to heating, fluid leaks and electricity consumption are fairly accurately estimated, the main source of error being the emission factors used. For example, the uncertainty surrounding the electricity emission factor is of the order of 10%[11]. On the other hand, for other emission sources, it is impossible to exactly quantify emissions, and we only estimate orders of magnitude, for which we sometimes have an estimate of the associated uncertainty. For example, for IT equipment, the uncertainty is estimated to represent 50% of the emissions estimate; for emissions linked to plane travel, the relative uncertainty even rises to 70% when taking into account the effect of contrails[12]. Finally, we note that our rough estimation of the emissions related to the construction, transport and recycling of the supercomputers (the same order of magnitude as the electricity and cooling consumption emissions) is highly uncertain, despite contributing more than a third to the total carbon footprint of the thesis.

[11] See footnote 2.
[12] Uncertainty estimations provided by the GES 1point5 tool.

Conclusion

In this study, we estimated the carbon footprint of the thesis to be around 14 tCO_2eq. Although there are significant uncertainties surrounding this estimate, it is highly relevant for identifying the most important sources of emissions in order to target reduction efforts. In the present case, the main sources of GHG emissions are related to computation. In particular, the generation of the LES ensemble used in Chapter IV represents the most substantial part of our emissions, highlighting the need to develop efficient sampling and learning methods to limit the carbon footprint of data-driven reduced-order models. More generally, this evaluation exercise raises awareness of the issue of carbon emissions in the research sector and raises questions about the future direction of research.

Bibliography

Afzal, A., Ansari, Z., Faizabadi, A. R., and Ramis, M. K. (2017). Parallelization Strategies for Computational Fluid Dynamics Software: State of the Art Review. *Archives of Computational Methods in Engineering*, 24(2):337–363. ISSN 1886-1784. DOI: 10.1007/s11831-016-9165-4.

Allen, C. T., Young, G. S., and Haupt, S. E. (2007). Improving pollutant source characterization by better estimating wind direction with a genetic algorithm. *Atmospheric Environment*, 41(11):2283–2289. DOI: 10.1016/j.atmosenv.2006.11.007.

Allwine, K. J., Shinn, J. H., Streit, G. E., Clawson, K. L., and Brown, M. (2002). Overview of URBAN 2000: A Multiscale Field Study of Dispersion through an Urban Environment. *Bulletin of the American Meteorological Society*, 83(4):521 – 536. DOI: 10.1175/1520-0477(2002)083<0521:OOUAMF>2.3.CO;2.

Allwine, K., Leach, M., Stockham, L., Shinn, J., Hosker, R., Bowers, J., and Pace, J. (2004). Overview of Joint Urban 2003: An atmospheric dispersion study in Oklahoma City. In *AMS Symposium on planning, nowcasting and forecasting in urban zone*, Seattle, WA. URL `https://ams.confex.com/ams/pdfpapers/74349.pdf`.

Anderson, J. L. and Anderson, S. L. (1999). A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. *Monthly Weather Review*, 127(12):2741 – 2758. DOI: 10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2.

Antonioni, G., Burkhart, S., Burman, J., Dejoan, A., Fusco, A., Gaasbeek, R., Gjesdal, T., Jäppinen, A., Riikonen, K., Morra, P., Parmhed, O., and Santiago, J. (2012). Comparison of CFD and operational dispersion models in an urban-like environment. *Atmospheric Environment*, 47:365–372. ISSN 1352-2310. DOI: 10.1016/j.atmosenv.2011.10.053.

Arcucci, R., Mottet, L., Pain, C., and Guo, Y.-K. (2019). Optimal reduced space for Variational Data Assimilation. *Journal of Computational Physics*, 379:51–69. ISSN 0021-9991. DOI: 10.1016/j.jcp.2018.10.042.

Aristodemou, E., Arcucci, R., Mottet, L., Robins, A., Pain, C., and Guo, Y.-K. (2019). Enhancing CFD-LES air pollution prediction accuracy using data assimilation. *Build. Environ.*, 165:106383. ISSN 0360-1323. DOI: 10.1016/j.buildenv.2019.106383.

Arnold, C. P. and Dey, C. H. (1986). Observing-Systems Simulation Experiments: Past, Present, and Future. *Bulletin of the American Meteorological Society*, 67(6):687 – 695. DOI: 10.1175/1520-0477(1986)067<0687:OSSEPP>2.0.CO;2.

Arthur, R. S., Mirocha, J. D., Lundquist, K. A., and Street, R. L. (2019). Using a canopy model framework to improve large-eddy simulations of the neutral atmospheric boundary layer in the weather research and forecasting model. *Monthly Weather Review*, 147 (1):31–52. DOI: 10.1175/MWR-D-18-0204.1.

Asch, M., Bocquet, M., and Nodet, M. (2016). *Data assimilation: methods, algorithms, and applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA. DOI: 10.1137/1.9781611974546.

Bahlali, M. L., Dupont, E., and Carissimo, B. (2019). Atmospheric dispersion using a Lagrangian stochastic approach: Application to an idealized urban area under neutral and stable meteorological conditions. *Journal of Wind Engineering and Industrial Aerodynamics*, 193:103976. ISSN 0167-6105. DOI: 10.1016/j.jweia.2019.103976.

Bailey, W. G., Oke, T. R., and Rouse, W. R. (1997). *Surface Climates of Canada*. McGill-Queen's University Press. ISBN 9780773509283. DOI: 10.1515/9780773563575.

Bailon-Ruiz, R. and Lacroix, S. (2020). Wildfire remote sensing with UAVs: A review from the autonomy point of view. In *2020 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 412–420. DOI: 10.1109/ICUAS48674.2020.9213986.

Basu, S. and Lacser, A. (2017). A cautionary note on the use of Monin–Obukhov similarity theory in very high-resolution large-eddy simulations. *Boundary-Layer Meteorology*, 163(2):351–355. DOI: 10.1007/s10546-016-0225-y.

Bauweraerts, P. and Meyers, J. (2021). Reconstruction of turbulent flow fields from lidar measurements using large-eddy simulation. *Journal of Fluid Mechanics*, 906:A17. DOI: 10.1017/jfm.2020.805.

Berkooz, G., Holmes, P., and Lumley, J. L. (1993). The proper orthogonal decomposition in the analysis of turbulent flows. *Annual review of fluid mechanics*, 25(1):539–575. DOI: 10.1146/annurev.fl.25.010193.002543.

Berthoud, F., Bzeznik, B., Gibelin, N., Laurens, M., Bonamy, C., Morel, M., and Schwindenhammer, X. (2020). Estimation de l'empreinte carbone d'une heure.coeur de calcul. Research report (in french), UGA - Université Grenoble Alpes, CNRS, INP Grenoble, INRIA. URL `https://hal.science/hal-02549565`. Accessed: 2023-12-01.

Beven, K. J., Almeida, S., Aspinall, W. P., Bates, P. D., Blazkova, S., Borgomeo, E., Freer, J., Goda, K., Hall, J. W., Phillips, J. C., Simpson, M., Smith, P. J., Stephenson, D. B., Wagener, T., Watson, M., and Wilkins, K. L. (2018). Epistemic uncertainties and natural hazard risk assessment – Part 1: A review of different natural hazard areas. *Natural Hazards and Earth System Sciences*, 18(10):2741–2768. DOI: 10.5194/nhess-18-2741-2018.

Bezpalcová, K. (2007). *Physical Modelling of Flow and Dispersion in Urban Canopy*. PhD thesis, Univerzita Karlova, Matematicko-fyzikální fakulta.

Biltoft, C. (1995). Surface effects on concentration fluctuation profiles. DPG Document No. JCP-95019 (U), Joint Contact Point Directorate, U.S. Army Dugway Proving Ground, Utah, USA.

Biltoft, C. (2001). Customer report for Mock Urban Setting Test. DPG Document No. WDTC-FR-01-121, West Desert Test Center, U.S. Army Dugway Proving Ground, Utah, USA.

Bishop, C. H., Etherton, B. J., and Majumdar, S. J. (2001). Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Monthly Weather Review*, 129(3):420 – 436. DOI: 10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Blocken, B. and Gualtieri, C. (2012). Ten iterative steps for model development and evaluation applied to computational fluid dynamics for environmental fluid mechanics. *Environ. Model. Softw.*, 33:1–22. DOI: 10.1016/j.envsoft.2012.02.001.

Blocken, B., Stathopoulos, T., Saathoff, P., and Wang, X. (2008). Numerical evaluation of pollutant dispersion in the built environment: Comparisons between models and experiments. *Journal of Wind Engineering and Industrial Aerodynamics*, 96(10):1817–1831. ISSN 0167-6105. DOI: 10.1016/j.jweia.2008.02.049. 4th International Symposium on Computational Wind Engineering (CWE2006).

Blocken, B., Tominaga, Y., and Stathopoulos, T. (2013). CFD simulation of micro-scale pollutant dispersion in the built environment. *Building and Environment*, 64:225–230. ISSN 0360-1323. DOI: 10.1016/j.buildenv.2013.01.001.

Blocken, B. (2014). 50 years of computational wind engineering: Past, present and future. *Journal of Wind Engineering and Industrial Aerodynamics*, 129:69–102. ISSN 0167-6105. DOI: 10.1016/j.jweia.2014.03.008.

Blocken, B. (2015). Computational Fluid Dynamics for urban physics: Importance, scales, possibilities, limitations and ten tips and tricks towards accurate and reliable simulations. *Building and Environment*, 91:219–245. ISSN 0360-1323. DOI: 10.1016/j.buildenv.2015.02.015. Fifty Year Anniversary for Building and Environment.

Blocken, B. and Carmeliet, J. (2004). Pedestrian wind environment around buildings: Literature review and practical examples. *Journal of Thermal Envelope and Building Science*, 28(2):107–159. DOI: 10.1177/1097196304044396.

Bocquet, M. and Sakov, P. (2014). An iterative ensemble Kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, 140(682):1521–1535. DOI: 10.1002/qj.2236.

Bocquet, M. (2014). Introduction to the principles and methods of data assimilation in geosciences. Lecture notes, École des Ponts ParisTech. URL `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1eed296a1003aee50809969a22acb66479c72415`. Accessed: 2023-12-01.

Boivin, P., Dauptain, A., Jiménez, C., and Cuenot, B. (2012). Simulation of a supersonic hydrogen–air autoignition-stabilized flame using reduced chemistry. *Combustion and Flame*, 159(4):1779–1790. ISSN 0010-2180. DOI: 10.1016/j.combustflame.2011.12.012.

Bonavita, M., Raynaud, L., and Isaksen, L. (2011). Estimating background-error variances with the ECMWF Ensemble of Data Assimilations system: some effects of ensemble size and day-to-day variability. *Quarterly Journal of the Royal Meteorological Society*, 137(655):423–434. DOI: 10.1002/qj.756.

Boudin, A. (2021). Turbulence injection methods for large-eddy simulations. Working note, Cerfacs – ISAE SUPAERO. URL `https://cerfacs.fr/wp-content/uploads/2021/10/Rapport_stage_A_Boudin_WN_CFD_21_228.pdf`. Accessed: 2023-12-01.

Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59(4):291–294. ISSN 1432-0770. DOI: 10.1007/BF00332918.

Bouyer, J. (2017). Les interactions ville et climat. Focus sur le phénomène d'îlot de chaleur urbain. Lecture slides, in French, Cerema. URL `https://www.pays-de-la-loire.developpement-durable.gouv.fr/IMG/pdf/2_2017-03-14_atelier-ecoq-dreal-nantes_jbouyer_vf.pdf`. Accessed: 2023-09-07.

Braconnier, T., Ferrier, M., Jouhaud, J.-C., Montagnac, M., and Sagaut, P. (2011). Towards an adaptive POD/SVD surrogate model for aeronautic design. *Computers & Fluids*, 40(1):195–209. ISSN 0045-7930. DOI: 10.1016/j.compfluid.2010.09.002.

Breuer, M. (2007). Boundary conditions for LES. In *Large-Eddy Simulation for Acoustics*. Cambridge Univ. Press. DOI: 10.1017/CBO9780511546143.009.

Brink, J. and Pebesma, E. (2014). Plume tracking with a mobile sensor based on incomplete and imprecise information. *Transactions in GIS*, 18(5):740–766. DOI: 10.1111/tgis.12063.

Britter, R. E. and Hanna, S. R. (2003). Flow and dispersion in urban areas. *Annual Review of Fluid Mechanics*, 35(1):469–496. DOI: 10.1146/annurev.fluid.35.101101.161147.

Britter, R. E., Di Sabatino, S., Caton, F., Cooke, K. M., Simmonds, P. G., and Nickless, G. (2002). Results from Three Field Tracer Experiments on the Neighbourhood Scale in the City of Birmingham UK. *Water, Air and Soil Pollution: Focus*, 2(5):79–90. ISSN 1573-2940. DOI: 10.1023/A:1021306612036.

Brunton, S. L. and Kutz, J. N. (2019). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press. DOI: 10.1017/9781108380690.

Bruschi, P., Piotto, M., Dell'Agnello, F., Ware, J., and Roy, N. (2016). Wind speed and direction detection by means of solid-state anemometers embedded on small quadcopters. *Procedia Engineering*, 168:802–805. ISSN 1877-7058. DOI:

10.1016/j.proeng.2016.11.274. Proceedings of the 30th anniversary Eurosensors Conference – Eurosensors 2016, 4-7. Sepember 2016, Budapest, Hungary.

Buehner, M., McTaggart-Cowan, R., Beaulne, A., Charette, C., Garand, L., Heilliette, S., Lapalme, E., Laroche, S., Macpherson, S. R., Morneau, J., et al. (2015). Implementation of Deterministic Weather Forecasting Systems Based on Ensemble–Variational Data Assimilation at Environment Canada. Part I: The Global System. *Monthly Weather Review*, 143(7):2532 – 2559. DOI: 10.1175/MWR-D-14-00354.1.

Burgers, G., van Leeuwen, P. J., and Evensen, G. (1998). Analysis Scheme in the Ensemble Kalman Filter. *Monthly Weather Review*, 126(6):1719 – 1724. DOI: 10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2.

Cai, J., Wu, J., Yuan, S., Kong, D., and Zhang, X. (2022). Prediction of gas leakage and dispersion in utility tunnels based on CFD-EnKF coupling model: A 3D full-scale application. *Sustainable Cities and Society*, 80:103789. ISSN 2210-6707. DOI: 10.1016/j.scs.2022.103789.

Calaf, M., Parlange, M. B., and Meneveau, C. (2011). Large eddy simulation study of scalar transport in fully developed wind-turbine array boundary layers. *Physics of Fluids*, 23(12):126603. ISSN 1070-6631. DOI: 10.1063/1.3663376.

Calhoun, R., Gouveia, F., Shinn, J., Chan, S., Stevens, D., Lee, R., and Leone, J. (2004). Flow around a complex building: Comparisons between experiments and a reynolds-averaged navier–stokes approach. *Journal of Applied Meteorology*, 43(5):696–710. DOI: 10.1175/2067.1.

Camelli, F., Lohner, R., and Hanna, S. (2005). VLES study of MUST experiment. In *43rd AIAA Aerospace Sciences Meeting and Exhibit*. DOI: 10.2514/6.2005-1279.

Campolongo, F., Saltelli, A., and Cariboni, J. (2011). From screening to quantitative sensitivity analysis. a unified approach. *Computer physics communications*, 182(4): 978–988. DOI: 10.1016/j.cpc.2010.12.039.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, 14(3):1171–1179. DOI: 10.1214/aos/1176350057.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5):e535. DOI: 10.1002/wcc.535.

Carruthers, D., Holroyd, R., Hunt, J., Weng, W., Robins, A., Apsley, D., Thompson, D., and Smith, F. (1994). UK-ADMS: A new approach to modelling dispersion in the Earth's atmospheric boundary layer. *J. Wind. Eng. Ind. Aerodyn.*, 52:139–153. DOI: 10.1016/0167-6105(94)90044-2.

Cassiani, M., Bertagni, M. B., Marro, M., and Salizzoni, P. (2020). Concentration fluctuations from localized atmospheric releases. *Boundary-Layer Meteorology*, 177(2): 461–510. ISSN 1573-1472. DOI: 10.1007/s10546-020-00547-4.

Chan, L. and Kwok, W. (2000). Vertical dispersion of suspended particulates in urban area of Hong Kong. *Atmospheric Environment*, 34(26):4403–4412. ISSN 1352-2310. DOI: 10.1016/S1352-2310(00)00181-3.

Chan, S. C., Kendon, E. J., Berthou, S., Fosser, G., Lewis, E., and Fowler, H. J. (2020). Europe-wide precipitation projections at convection permitting scale with the Unified Model. *Clim. Dyn.*, 55(3):409–428. ISSN 1432-0894. DOI: 10.1007/s00382-020-05192-8.

Chang, J. and Hanna, S. (2004). Air quality model performance evaluation. *Meteorol. Atm. Phys*, 87(1):167–196. DOI: 10.1007/s00703-003-0070-7.

Chang, J. C. and Hanna, S. R. (2005). Technical descriptions and user's guide for the BOOT statistical model evaluation software package, version 2.0. *George Mason University and Harvard School of Public Health, Fairfax, Virginia, USA*. URL `https://www.harmo.org/Kit/Download/BOOT_UG.pdf`. Accessed: 2023-12-01.

Cheng, K., Lu, Z., Ling, C., and Zhou, S. (2020). Surrogate-assisted global sensitivity analysis: an overview. *Structural and Multidisciplinary Optimization*, 61:1187–1213. DOI: 10.1007/s00158-019-02413-5.

Cheng, S., Quilodrán-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss, B., Brajard, J., Xiao, D., Janjic, T., Ding, W., Guo, Y., Carrassi, A., Bocquet, M., and Arcucci, R. (2023). Machine learning with data assimilation and uncertainty quantification for dynamical systems: A review. *IEEE/CAA Journal of Automatica Sinica*, 10(6):1361–1387. DOI: 10.1109/JAS.2023.123537.

Cheng, Y., Lien, F. S., Yee, E., and Sinclair, R. (2003). A comparison of large Eddy simulations with a standard k–ϵ Reynolds-averaged Navier–Stokes model for the prediction of a fully developed turbulent flow over a matrix of cubes. *Journal of Wind Engineering and Industrial Aerodynamics*, 91(11):1301–1328. ISSN 0167-6105. DOI: 10.1016/j.jweia.2003.08.001.

Chinesta, F., Ladeveze, P., and Cueto, E. (2011). A short review on model order reduction based on proper generalized decomposition. *Archives of Computational Methods in Engineering*, 18(4):395–404. ISSN 1886-1784. DOI: 10.1007/s11831-011-9064-7.

Choi, H. and Moin, P. (2012). Grid-point requirements for large eddy simulation: Chapman's estimates revisited. *Physics of fluids*, 24(1). DOI: 10.1063/1.3676783.

Choi, H.-L. and How, J. P. (2011). Efficient targeting of sensor networks for large-scale systems. *IEEE Transactions on Control Systems Technology*, 19(6):1569–1577. DOI: 10.1109/TCST.2010.2093134.

Chu, M., Liu, L., Zheng, Q., Franz, E., Seidel, H.-P., Theobalt, C., and Zayer, R. (2022). Physics informed neural fields for smoke reconstruction with sparse data. *ACM Transactions on Graphics (TOG)*, 41(4):1–14. DOI: 10.1145/3528223.3530169.

Cimorelli, A. J., Perry, S. G., Venkatram, A., Weil, J. C., Paine, R. J., Wilson, R. B., Lee, R. F., Peters, W. D., and Brode, R. W. (2005). AERMOD: A dispersion model for industrial source applications. Part I: General model formulation and boundary layer characterization. *J. Appl. Meteorol.*, 44(5):682–693. DOI: 10.1175/JAM2227.1.

Coceal, O., Thomas, T. G., Castro, I. P., and Belcher, S. E. (2006). Mean flow and turbulence statistics over groups of urban-like cubical obstacles. *Boundary-Layer Meteorology*, 121(3):491–519. ISSN 1573-1472. DOI: 10.1007/s10546-006-9076-2.

Colin, O. and Rudgyard, M. (2000). Development of high-order Taylor–Galerkin schemes for LES. *Journal of computational physics*, 162(2):338–371. DOI: 10.1006/jcph.2000.6538.

Cordier, L. and Bergmann, M. (2006). Réduction de dynamique par décomposition orthogonale aux valeurs propres (POD) (in French). Lecture notes, Ecole de printemps OCET. URL `https://www.math.u-bordeaux.fr/~mbergman/PDF/OuvrageSynthese/OCET06.pdf`. Accessed: 2023-12-01.

Costes, A., Rochoux, M. C., Lac, C., and Masson, V. (2021). Subgrid-scale fire front reconstruction for ensemble coupled atmosphere-fire simulations of the FireFlux I experiment. *Fire Saf. J.*, 126:103475. DOI: 10.1016/j.firesaf.2021.103475.

Dabas, J., Gicquel, L., Odier, N., and Duchaine, F. (2022). Large eddy simulations of wind turbine flows. In *Volume 11: Wind Energy*, Turbo Expo: Power for Land, Sea, and Air, page V011T38A011. DOI: 10.1115/GT2022-82096.

Dauxois, T., Peacock, T., Bauer, P., Caulfield, C. P., Cenedese, C., Gorlé, C., Haller, G., Ivey, G. N., Linden, P. F., Meiburg, E., Pinardi, N., Vriend, N. M., and Woods, A. W. (2021). Confronting grand challenges in environmental fluid mechanics. *Phys. Rev. Fluids*, 6:020501. DOI: 10.1103/PhysRevFluids.6.020501.

Davidson, M., Mylne, K., Jones, C., Phillips, J., Perkins, R., Fung, J., and Hunt, J. (1995). Plume dispersion through large groups of obstacles — A field investigation. *Atmospheric Environment*, 29(22):3245–3256. ISSN 1352-2310. DOI: 10.1016/1352-2310(95)00254-V.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. DOI: 10.1017/CBO9780511802843.

De Paepe, M., Jeanneau, L., Mariette, J., Aumont, O., and Estevez-Torres, A. (2023). Purchases dominate the carbon footprint of research laboratories. *bioRxiv*. DOI: 10.1101/2023.04.04.535626. Preprint.

Deardorff, J. W. et al. (1970). A numerical study of three-dimensional turbulent channel flow at large reynolds numbers. *J. Fluid Mech*, 41(2):453–480. DOI: 10.1017/S0022112070000691.

Defforge, C. L., Carissimo, B., Bocquet, M., Bresson, R., and Armand, P. (2021). Improving numerical dispersion modelling in built environments with data assimilation using the iterative ensemble Kalman smoother. *Boundary-Layer Meteorology*, 179(2): 209–240. ISSN 1573-1472. DOI: 10.1007/s10546-020-00588-9.

Defforge, C. (2019). *Data assimilation for micrometeorological applications with the fluid dynamics model Code_Saturne*. PhD thesis, Université Paris-Est. URL `https://theses.hal.science/tel-02318713v2/document`. Accessed: 2023-12-01.

Defforge, C. L., Carissimo, B., Bocquet, M., Bresson, R., and Armand, P. (2019). Improving CFD atmospheric simulations at local scale for wind resource assessment using the iterative ensemble Kalman smoother. *Journal of Wind Engineering and Industrial Aerodynamics*, 189:243–257. ISSN 0167-6105. DOI: 10.1016/j.jweia.2019.03.030.

Dejoan, A., Santiago, J., Martilli, A., Martin, F., and Pinelli, A. (2010). Comparison between large-eddy simulation and Reynolds-averaged Navier–Stokes computations for the MUST field experiment. Part II: Effects of incident wind angle deviation on the mean flow and plume dispersion. *Boundary-Layer Meteorology*, 135(1):133–150. DOI: 10.1007/s10546-010-9467-2.

Delprat-Jannaud, F. (2022). La capture et le stockage du carbone, comment ça marche ? (in French). *The Conversation*. URL `https://theconversation.com/la-capture-et-le-stockage-du-carbone-comment-ca-marche-192673`. Accessed: 2023-10-26.

Deng, N., Noack, B. R., Morzyński, M., and Pastur, L. R. (2021)a. Galerkin force model for transient and post-transient dynamics of the fluidic pinball. *Journal of Fluid Mechanics*, 918:A4. DOI: 10.1017/jfm.2021.299.

Deng, Z., He, C., and Liu, Y. (2021)b. Deep neural network-based strategy for optimal sensor placement in data assimilation of turbulent flow. *Physics of Fluids*, 33(2):025119. ISSN 1070-6631. DOI: 10.1063/5.0035230.

Dhamankar, N. S., Blaisdell, G. A., and Lyrintzis, A. S. (2018). Overview of turbulent inflow boundary conditions for large-eddy simulations. *Aiaa Journal*, 56(4):1317–1334.

Di Sabatino, S., Buccolieri, R., Olesen, H. R., Ketzel, M., Berkowicz, R., Franke, J., Schatzmann, M., Schlunzen, K., Leitl, B., Britter, R., Borrego, C., Costa, A., Castelli, S., Reisin, T., Hellsten, A., Saloranta, J., Moussiopoulos, N., Barmpas, F., Brzozowski, K., Goricsan, I., Balczo, M., Bartzis, J., Efthimiou, G., Santiago, J., Martilli, A., Piringer, M., Baumann-Stanzer, K., Hirtl, M., Baklanov, A., Nuterman, R., and Starchenko, A. (2011). COST 732 in practice: the MUST model evaluation exercise. *International Journal of Environment and Pollution*, 44(1-4):403–418. DOI: 10.1504/IJEP.2011.038442.

Diffenbaugh, N. S., Singh, D., Mankin, J. S., Horton, D. E., Swain, D. L., Touma, D., Charland, A., Liu, Y., Haugen, M., Tsiang, M., et al. (2017). Quantifying the influence of global warming on unprecedented extreme climate events. *Proc. Natl. Acad. Sci. U.S.A.*, 114(19):4881–4886. DOI: 10.1073/pnas.1618082114.

Donnelly, R., Lyons, T., and Flassak, T. (2009). Evaluation of results of a numerical simulation of dispersion in an idealised urban area for emergency response modelling. *Atmos. Environ.*, 43(29):4416–4423. ISSN 1352-2310. DOI: 10.1016/j.atmosenv.2009.05.038.

Duchaine, F., Jauré, S., Poitou, D., Quémerais, E., Staffelbach, G., Morel, T., and Gicquel, L. (2015). Analysis of high performance conjugate heat transfer with the OpenPALM coupler. *Computational Science & Discovery*, 8(1):015003. DOI: 10.1088/1749-4699/8/1/015003.

Dumont Le Brazidec, J., Bocquet, M., Saunier, O., and Roustan, Y. (2023). Bayesian transdimensional inverse reconstruction of the Fukushima Daiichi caesium 137 release. *Geosci. Model Dev.*, 16(3):1039–1052. DOI: 10.5194/gmd-16-1039-2023.

Dur, T. H., Arcucci, R., Mottet, L., Solana, M. M., Pain, C., and Guo, Y.-K. (2020). Weak Constraint Gaussian Processes for optimal sensor placement. *Journal of Computational Science*, 42:101110. ISSN 1877-7503. DOI: 10.1016/j.jocs.2020.101110.

Ebden, M. et al. (2008). Gaussian processes for regression: A quick introduction. Lecture notes, The Website of Robotics Research Group in Department on Engineering Science, University of Oxford. URL `https://arxiv.org/abs/1505.02965`. Accessed: 2023-12-01.

Eckerman, I. (2005). *The Bhopal saga: causes and consequences of the world's largest industrial disaster*. Universities press.

EEA. (2020). Air quality in Europe. 2020 report, European Environment Agency. URL `https://www.eea.europa.eu/publications/air-quality-in-europe-2020`.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26. DOI: 10.1007/978-1-4612-4380-9_41.

Efthimiou, G. C., Bartzis, J. G., and Koutsourakis, N. (2011). Modelling concentration fluctuations and individual exposure in complex urban environments. *Journal of Wind Engineering and Industrial Aerodynamics*, 99(4):349–356. ISSN 0167-6105. DOI: 10.1016/j.jweia.2010.12.007. The Fifth International Symposium on Computational Wind Engineering.

Ehrendorfer, M. (2007). A review of issues in ensemble-based Kalman filtering. *Meteorologische Zeitschrift*, 16(6):795–818. DOI: 10.1127/0941-2948/2007/0256.

El Garroussi, S., Ricci, S., De Lozzo, M., Goutal, N., and Lucor, D. (2020). Assessing uncertainties in flood forecasts using a mixture of generalized polynomial chaos expansions. In *2020 TELEMAC-MASCARET User Conference*. URL `https://hal.science/hal-03444227/document`. Accessed: 2023-12-01.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162. DOI: 10.1029/94JC00572.

Evensen, G. (2003). The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367. ISSN 1616-7228. DOI: 10.1007/s10236-003-0036-9.

Evensen, G. (2009). The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control Systems Magazine*, 29(3):83–104. DOI: 10.1109/MCS.2009.932223.

Farchi, A., Laloyaux, P., Bonavita, M., and Bocquet, M. (2021). Using machine learning to correct model error in data assimilation and forecast applications. *Quarterly Journal of the Royal Meteorological Society*, 147(739):3067–3084. DOI: 10.1002/qj.4116.

Favre, A. (1965). The equations of compressible turbulent gases. *Annual Summary Report AD0622097*. URL `https://apps.dtic.mil/sti/pdfs/AD0622097.pdf`. Accessed: 2023-12-01.

Ferreboeuf, H., Berthoud, F., Bihouix, P., Fabre, P., Kaplan, D., Lefèvre, L., Monnin, A., Ridoux, O., Vaija, S., Vautier, M., Verne, X., Ducass, A., Efoui-Hess, M., and Kahraman, Z. (2018). Lean ICT : pour une sobriété numérique (in French). Technical report, The Shift Project. URL `https://theshiftproject.org/wp-content/uploads/2018/11/Rapport-final-v8-WEB.pdf`. Accessed: 2023-12-01.

Ficher, M., Berthoud, F., Ligozat, A.-L., Sigonneau, P., Tebbani, B., and Wisslé, M. (2021). Évaluation de l'empreinte carbone de la transmission d'un gigaoctet de données sur le réseau RENATER (in French). Technical report, RENATER, EcoInfo. URL `https://hal.science/hal-04197870v1/document`. Accessed: 2023-12-01.

Fisher, B., Joffre, S., Kukkonen, J., Piringer, M., Rotach, M., and Schatzmann, M. (2005). Meteorology applied to urban air pollution problems. Technical report, COST European Cooperation in Science and Technology. URL `https://www2.dmu.dk/atmosphericenvironment/cost/docs/cost715-final.pdf`. Accessed: 2023-12-01.

Forkman, J., Josse, J., and Piepho, H.-P. (2019). Hypothesis tests for principal component analysis when variables are standardized. *Journal of Agricultural, Biological and Environmental Statistics*, 24:289–308. DOI: 10.1007/s13253-019-00355-5.

Forster, P. M., Smith, C. J., Walsh, T., Lamb, W. F., Lamboll, R., Hauser, M., Ribes, A., Rosen, D., Gillett, N., Palmer, M. D., Rogelj, J., von Schuckmann, K., Seneviratne, S. I., Trewin, B., Zhang, X., Allen, M., Andrew, R., Birt, A., Borger, A., Boyer, T., Broersma, J. A., Cheng, L., Dentener, F., Friedlingstein, P., Gutiérrez, J. M., Gütschow, J., Hall, B., Ishii, M., Jenkins, S., Lan, X., Lee, J.-Y., Morice, C., Kadow, C., Kennedy, J., Killick, R., Minx, J. C., Naik, V., Peters, G. P., Pirani, A., Pongratz, J., Schleussner, C.-F., Szopa, S., Thorne, P., Rohde, R., Rojas Corradi, M., Schumacher,

D., Vose, R., Zickfeld, K., Masson-Delmotte, V., and Zhai, P. (2023). Indicators of global climate change 2022: annual update of large-scale indicators of the state of the climate system and human influence. *Earth System Science Data*, 15(6):2295–2327. DOI: 10.5194/essd-15-2295-2023.

Fouré, J., Martin, S., Berry, A., Fontan, O., Ferrat, M., Tamokoué Kamga, P.-H., and Sgambati, E. (2020). Maitriser l'empreinte carbone de la France (in French). Technical report, Haut Conseil pour le Climat. URL `https://www.hautconseilclimat.fr/wp-content/uploads/2020/10/hcc_rapport_maitriser-lempreinte-carbone-de-la-france-1.pdf`. Accessed: 2023-12-01.

Fox, S., Hanna, S., Mazzola, T., Spicer, T., Chang, J., and Gant, S. (2022). Overview of the Jack Rabbit II (JR II) field experiments and summary of the methods used in the dispersion model comparisons. *Atmos. Environ.*, 269:118783. DOI: 10.1016/j.atmosenv.2021.118783.

Fox, S. B. and Storwold, D. (2011). Project Jack Rabbit: Field tests. Technical report, Chemical Security and Analysis Center, Science and Technology Directorate, US Department of Homeland Security, CSAC. URL `https://www.dhs.gov/sites/default/files/publications/csac-11-006_r1_project_jack_rabbit_field_tests_pr-508c.pdf`. Accessed: 2023-12-01.

Francis, P. N., Cooke, M. C., and Saunders, R. W. (2012). Retrieval of physical properties of volcanic ash using Meteosat: A case study from the 2010 Eyjafjallajökull eruption. *Journal of Geophysical Research: Atmospheres*, 117(D20). DOI: 10.1029/2011JD016788.

Franke, J., Hellsten, A., Schlünzen, H., and Carissimo, B. (2007). Best practice guideline for the CFD simulation of flows in the urban environmen. Technical report, COST European Cooperation in Science and Technology. URL `https://hal.science/hal-04181390`. Accessed: 2023-12-01.

Gao, X., Wang, Y., Overton, N., Zupanski, M., and Tu, X. (2017). Data-assimilated computational fluid dynamics modeling of convection-diffusion-reaction problems. *Journal of Computational Science*, 21:38–59. ISSN 1877-7503. DOI: 10.1016/j.jocs.2017.05.014.

Garbero, V. (2008). *Pollutant dispersion in urban canopy study of the plume behaviour through an obstacle array*. PhD thesis, École Centrale de Lyon, France.

Garbero, V., Salizzoni, P., and Soulhac, L. (2010). Experimental study of pollutant dispersion within a network of streets. *Boundary-Layer Meteorology*, 136(3):457–487. ISSN 1573-1472. DOI: 10.1007/s10546-010-9511-2.

García-Sanchez, C. and Gorlé, C. (2018). Uncertainty quantification for microscale CFD simulations based on input from mesoscale codes. *J. Wind. Eng. Ind. Aerodyn.*, 176: 87–97. DOI: 10.1016/j.jweia.2018.03.011.

García-Sanchez, C., van Beeck, J., and Gorlé, C. (2018). Predictive large eddy simulations for urban flows: Challenges and opportunities. *Build. Environ.*, 139:146–156. ISSN 0360-1323. DOI: 10.1016/j.buildenv.2018.05.007.

García-Sánchez, C., Philips, D., and Gorlé, C. (2014). Quantifying inflow uncertainties for CFD simulations of the flow in downtown Oklahoma City. *Build. Environ.*, 78: 118–129. ISSN 0360-1323. DOI: 10.1016/j.buildenv.2014.04.013.

García-Sánchez, C., Van Tendeloo, G., and Gorlé, C. (2017). Quantifying inflow uncertainties in RANS simulations of urban pollutant dispersion. *Atmospheric Environment*, 161:263–273. ISSN 1352-2310. DOI: 10.1016/j.atmosenv.2017.04.019.

Germano, M., Piomelli, U., Moin, P., and Cabot, W. H. (1991). A dynamic subgrid-scale eddy viscosity model. *Physics of Fluids A: Fluid Dynamics*, 3(7):1760–1765. DOI: 10.1063/1.857955.

Gicquel, L. Y., Gourdain, N., Boussuge, J.-F., Deniau, H., Staffelbach, G., Wolf, P., and Poinsot, T. (2011). High performance parallel computing of flows in complex geometries. *Comptes Rendus Mécanique*, 339(2):104–124. ISSN 1631-0721. DOI: 10.1016/j.crme.2010.11.006. High Performance Computing.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140:107–113(6). ISSN 0956-375X. DOI: 10.1049/ip-f-2.1993.0015.

Gorlé, C. and Iaccarino, G. (2013). A framework for epistemic uncertainty quantification of turbulent scalar flux models for Reynolds-averaged Navier-Stokes simulations. *Physics of Fluids*, 25(5):055105. ISSN 1070-6631. DOI: 10.1063/1.4807067.

Gorlé, C., Garcia-Sanchez, C., and Iaccarino, G. (2015). Quantifying inflow and RANS turbulence model form uncertainties for wind engineering flows. *Journal of Wind Engineering and Industrial Aerodynamics*, 144:202–212. DOI: 10.1016/j.jweia.2015.03.025.

Gousseau, P., Blocken, B., Stathopoulos, T., and van Heijst, G. (2011). CFD simulation of near-field pollutant dispersion on a high-resolution grid: A case study by LES and RANS for a building group in downtown Montreal. *Atmos. Environ.*, 45(2):428–438. ISSN 1352-2310. DOI: 10.1016/j.atmosenv.2010.09.065.

Gromke, C., Jamarkattel, N., and Ruck, B. (2016). Influence of roadside hedgerows on air quality in urban street canyons. *Atmospheric Environment*, 139:75–86. ISSN 1352-2310. DOI: 10.1016/j.atmosenv.2016.05.014.

Gronskis, A., Heitz, D., and Mémin, E. (2013). Inflow and initial conditions for direct numerical simulation based on adjoint data assimilation. *Journal of Computational Physics*, 242:480–497. ISSN 0021-9991. DOI: 10.1016/j.jcp.2013.01.051.

Grylls, T., Cornec, C. M. L., Salizzoni, P., Soulhac, L., Stettler, M. E., and van Reeuwijk, M. (2019). Evaluation of an operational air quality model using large-eddy simulation. *Atmos. Environ.*, 3:100041. ISSN 2590–1621. DOI: 10.1016/j.aeaoa.2019.100041.

Gryning, S.-E. and Lyck, E. (1984). Atmospheric dispersion from elevated sources in an urban area: Comparison between tracer experiments and model calculations. *Journal of Applied Meteorology and Climatology*, 23(4):651 – 660. DOI: 10.1175/1520-0450(1984)023<0651:ADFESI>2.0.CO;2.

Guilbert, D. (2021). Comment l'hydrogène peut contribuer à stocker l'électricité à grande échelle (in French). *The Conversation*. URL `https://theconversation.com/comment-lhydrogene-peut-contribuer-a-stocker-lelectricite-a-grande-echelle-160307`. Accessed: 2023-10-26.

Hajra, B. and Stathopoulos, T. (2012). A wind tunnel study of the effect of downstream buildings on near-field pollutant dispersion. *Building and Environment*, 52:19–31. ISSN 0360-1323. DOI: 10.1016/j.buildenv.2011.12.021.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2009). Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. Technical report, California Institute of Technology. URL `https://authors.library.caltech.edu/27187/1/Caltech_ACM_TR_2009_05.pdf`. Accessed: 2023-12-01.

Halliwell, G. R., Srinivasan, A., Kourafalou, V., Yang, H., Willey, D., Le Hénaff, M., and Atlas, R. (2014). Rigorous evaluation of a fraternal twin ocean osse system for the open gulf of mexico. *Journal of Atmospheric and Oceanic Technology*, 31(1):105 – 130. DOI: 10.1175/JTECH-D-13-00011.1.

Halton, J. H. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, 7(12):701–702. DOI: 10.1145/355588.365104.

Hamby, D. (1995). A comparison of sensitivity analysis techniques. *Health physics*, 68 (2):195–204. DOI: 10.1097/00004032-199502000-00005.

Hamill, T. M. and Snyder, C. (2000). A hybrid ensemble Kalman filter-3D variational analysis scheme. *Monthly Weather Review*, 128(8):2905 – 2919. DOI: 10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2.

Hamill, T. M., Whitaker, J. S., and Snyder, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, 129(11):2776 – 2790. DOI: 10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2.

Hanna, S. (1989)a. Plume dispersion and concentration fluctuations in the atmosphere. Encyclopedia of environmental control technology. *Air pollution control*, 2:547–582.

Hanna, S., Tehranian, S., Carissimo, B., Macdonald, R., and Lohner, R. (2002). Comparisons of model simulations with observations of mean flow and turbulence within simple obstacle arrays. *Atmos. Environ.*, 36(32):5067–5079. ISSN 1352-2310. DOI: 10.1016/S1352-2310(02)00566-6.

Hanna, S. R. (1989)b. Confidence limits for air quality model evaluations, as estimated by bootstrap and jackknife resampling methods. *Atmos. Environ.*, 23(6):1385–1398. ISSN 0004-6981. DOI: 10.1016/0004-6981(89)90161-3.

Hanna, S. R. and Chang, J. C. (2001). Use of the Kit Fox field data to analyze dense gas dispersion modeling issues. *Atmospheric Environment*, 35(13):2231–2242. ISSN 1352-2310. DOI: 10.1016/S1352-2310(00)00481-7.

Hanna, S. R., Briggs, G. A., and Hosker Jr, R. P. (1982). Handbook on atmospheric diffusion. Technical report, National Oceanic and Atmospheric Administration, Oak Ridge TN, USA. URL `https://www.osti.gov/servlets/purl/5591108`. Accessed: 2023-12-01.

Hanna, S. R., Britter, R., and Franzese, P. (2003). A baseline urban dispersion model evaluated with Salt Lake City and Los Angeles tracer data. *Atmospheric Environment*, 37(36):5069–5082. ISSN 1352-2310. DOI: 10.1016/j.atmosenv.2003.08.014.

Hanna, S. R., Hansen, O. R., and Dharmavaram, S. (2004). FLACS CFD air quality model performance evaluation with Kit Fox, MUST, Prairie Grass, and EMU observations. *Atmos. Environ.*, 38(28):4675–4687. DOI: 10.1016/j.atmosenv.2004.05.041.

Harms, F., Leitl, B., Schatzmann, M., and Patnaik, G. (2011). Validating LES-based flow and dispersion models. *J. Wind. Eng. Ind. Aerodyn.*, 99(2):289–295. DOI: 10.1016/j.jweia.2011.01.007.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer. DOI: 10.1007/978-0-387-21606-5.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049. DOI: 10.1002/qj.3803.

Hertwig, D., Soulhac, L., Fuka, V., Auerswald, T., Carpentieri, M., Hayden, P., Robins, A., Xie, Z.-T., and Coceal, O. (2018). Evaluation of fast atmospheric dispersion models in a regular street network. *Environ. Fluid Mech.*, 18(4):1007–1044. DOI: 10.1007/s10652-018-9587-7.

Hilderman, T., Chong, R., and Kiel, D. (2008). A laboratory study of momentum and passive scalar transport and diffusion within and above a model urban canopy. Technical report, DRDC Suffield CR. URL https://apps.dtic.mil/sti/pdfs/AD1003970.pdf. Accessed: 2023-12-01.

Hinton, G. E. and Zemel, R. (1993). Autoencoders, minimum description length and Helmholtz free energy. In Cowan, J., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann. URL https://proceedings.neurips.cc/paper_files/paper/1993/file/9e3cfc48eccf81a0d57663e129aef3cb-Paper.pdf. Accessed: 2023-12-01.

Hoffman, R. N. and Atlas, R. (2016). Future observing system simulation experiments. *Bulletin of the American Meteorological Society*, 97(9):1601 – 1616. DOI: 10.1175/BAMS-D-15-00200.1.

Holmes, N. and Morawska, L. (2006). A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available. *Atmos. Environ.*, 40(30):5902–5928. DOI: 10.1016/j.atmosenv.2006.06.003.

Houtekamer, P. L. and Mitchell, H. L. (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, 126(3):796 – 811. DOI: 10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

Houtekamer, P. L. and Zhang, F. (2016). Review of the ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 144(12):4489 – 4532. DOI: 10.1175/MWR-D-15-0440.1.

Hsieh, K.-J., Lien, F.-S., and Yee, E. (2007). Numerical modeling of passive scalar dispersion in an urban canopy layer. *Journal of Wind Engineering and Industrial Aerodynamics*, 95(12):1611–1636. ISSN 0167-6105. DOI: 10.1016/j.jweia.2007.02.028.

Hu, J., Niu, H., Carrasco, J., Lennox, B., and Arvin, F. (2022). Fault-tolerant cooperative navigation of networked UAV swarms for forest fire monitoring. *Aerospace Science and Technology*, 123:107494. ISSN 1270-9638. DOI: 10.1016/j.ast.2022.107494.

Hutchinson, M., Oh, H., and Chen, W.-H. (2017). A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors. *Information Fusion*, 36:130–148. ISSN 1566-2535. DOI: 10.1016/j.inffus.2016.11.010.

Huybers, P., McKinnon, K. A., Rhines, A., and Tingley, M. (2014). US daily temperatures: The meaning of extremes in the context of nonnormality. *J. Clim.*, 27(19):7368–7384. DOI: 10.1175/JCLI-D-14-00216.1.

Hwang, Y. and Gorlé, C. (2023). Large-eddy simulations to define building-specific similarity relationships for natural ventilation flow rates. *Flow*, 3:E10. DOI: 10.1017/flo.2023.4.

Idczak, M., Mestayer, P., Rosant, J.-M., Sini, J.-F., and Violleau, M. (2007). Micrometeorological measurements in a street canyon during the joint ATREUS-PICADA experiment. *Boundary-Layer Meteorology*, 124(1):25–41. ISSN 1573-1472. DOI: 10.1007/s10546-006-9095-z.

Ide, K., Courtier, P., Ghil, M., and Lorenc, A. C. (1997). Unified notation for data assimilation : Operational, sequential and variational. *Journal of the Meteorological Society of Japan. Ser. II*, 75(1B):181–189. DOI: 10.2151/jmsj1965.75.1B_181.

Iooss, B. and Prieur, C. (2019). Shapley effects for sensitivity analysis with correlated inputs: comparisons with sobol'indices, numerical estimation and applications. *International Journal for Uncertainty Quantification*, 9(5). DOI: 10.1615/Int.J.UncertaintyQuantification.2019028372.

Jahn, W., Rein, G., and Torero, J. (2012). Forecasting fire dynamics using inverse computational fluid dynamics and tangent linearisation. *Advances in Engineering Software*, 47(1):114–126. ISSN 0965-9978. DOI: 10.1016/j.advengsoft.2011.12.005.

Jammalamadaka, S. R. and SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific. DOI: 10.1142/4031.

Jicha, M., Pospisil, J., and Katolicky, J. (2000). Dispersion of pollutants in street canyon under traffic induced flow and turbulence. *Environmental Monitoring and Assessment*, 65(1):343–351. ISSN 1573-2959. DOI: 10.1023/A:1006452422885.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202. DOI: 10.1098/rsta.2015.0202.

Jones, W. P. and B.E, L. (1972). The prediction of laminarization with a two-equation model of turbulence. *International Journal of Heat and Mass Transfer*, 15(2):301–314. ISSN 0017-9310. DOI: 10.1016/0017-9310(72)90076-2.

Joulin, P.-A., Mayol, M. L., Masson, V., Blondel, F., Rodier, Q., Cathelain, M., and Lac, C. (2020). The actuator line method in the meteorological LES model Meso-NH to analyze the Horns Rev 1 wind farm photo case. *Frontiers in Earth Science*, 7. ISSN 2296-6463. DOI: 10.3389/feart.2019.00350.

Kastner-Klein, P., Berkowicz, R., and Britter, R. (2004). The influence of street architecture on flow and dispersion in street canyons. *Meteorology and Atmospheric Physics*, 87(1):121–131. ISSN 1436-5065. DOI: 10.1007/s00703-003-0065-4.

Kato, H., Yoshizawa, A., Ueno, G., and Obayashi, S. (2015). A data assimilation methodology for reconstructing turbulent flows around aircraft. *Journal of Computational Physics*, 283:559–581. ISSN 0021-9991. DOI: 10.1016/j.jcp.2014.12.013.

Keating, A., Piomelli, U., Balaras, E., and Kaltenbach, H.-J. (2004). A priori and a posteriori tests of inflow conditions for large-eddy simulation. *Phys. Fluids*, 16(12): 4696–4712. DOI: 10.1063/1.1811672.

Keats, A., Yee, E., and Lien, F.-S. (2007). Bayesian inference for source determination with applications to a complex urban environment. *Atmospheric Environment*, 41(3): 465–479. ISSN 1352-2310. DOI: 10.1016/j.atmosenv.2006.08.044.

Kessy, A., Lewin, A., and Strimmer, K. (2018). Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314. DOI: 10.1080/00031305.2016.1277159.

Kim, B., Azevedo, V. C., Thuerey, N., Kim, T., Gross, M., and Solenthaler, B. (2019)a. Deep fluids: A generative network for parameterized fluid simulations. *Computer Graphics Forum*, 38(2):59–70. DOI: 10.1111/cgf.13619.

Kim, J., Kim, S., Ju, C., and Son, H. I. (2019)b. Unmanned aerial vehicles in agriculture: A review of perspective of platform, control, and applications. *IEEE Access*, 7:105100–105115. DOI: 10.1109/ACCESS.2019.2932119.

Kiureghian, A. D. and Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. ISSN 0167-4730. DOI: 10.1016/j.strusafe.2008.06.020. Risk Acceptance and Risk Communication.

König, M. (2014). *Large-eddy simulation modelling for urban scale*. PhD thesis, University of Leipzig. URL `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=baab9d7b41623099c1b6d840c11821b8e31fac9b`.

Koutsourakis, N., Bartzis, J. G., and Markatos, N. C. (2012). Evaluation of Reynolds stress, k-ϵ and RNG k-ϵ turbulence models in street canyon flows using various experimental datasets. *Environ. Fluid Mech.*, 12(4):379–403. DOI: 10.1007/s10652-012-9240-9.

Kraichnan, R. H. (1970). Diffusion by a random velocity field. *Phys. Fluids*, 13(1):22–31. DOI: 10.1063/1.1692799.

Kumar, N., Kerhervé, F., and Cordier, L. (2019). Dynamic reconstruction of a numerical 2D cylinder wake flow using data assimilation. In *5th Symposium on Fluid-Structure-Sound Interactions and Control (FSSIC2019)*, Crete, Greece. URL `https://hal.science/hal-02411731`. Accessed: 2023-12-01.

Kumar, P., Feiz, A.-A., Ngae, P., Singh, S. K., and Issartel, J.-P. (2015). CFD simulation of short-range plume dispersion from a point release in an urban like environment. *Atmos. Environ.*, 122:645–656. DOI: 10.1016/j.atmosenv.2015.10.027.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Stat.*, pages 1217–1241. DOI: 10.1214/aos/1176347265.

Labahn, J. W., Wu, H., Coriton, B., Frank, J. H., and Ihme, M. (2019). Data assimilation using high-speed measurements and les to examine local extinction events in turbulent flames. *Proceedings of the Combustion Institute*, 37(2):2259–2266. ISSN 1540-7489. DOI: 10.1016/j.proci.2018.06.043.

Lac, C., Chaboureau, J.-P., Masson, V., Pinty, J.-P., Tulet, P., Escobar, J., Leriche, M., Barthe, C., Aouizerats, B., Augros, C., Aumond, P., Auguste, F., Bechtold, P., Berthet, S., Bielli, S., Bosseur, F., Caumont, O., Cohard, J.-M., Colin, J., Couvreux, F., Cuxart, J., Delautier, G., Dauhut, T., Ducrocq, V., Filippi, J.-B., Gazen, D., Geoffroy, O., Gheusi, F., Honnert, R., Lafore, J.-P., Lebeaupin Brossier, C., Libois, Q., Lunet, T., Mari, C., Maric, T., Mascart, P., Mogé, M., Molinié, G., Nuissier, O., Pantillon, F., Peyrillé, P., Pergaud, J., Perraud, E., Pianezze, J., Redelsperger, J.-L., Ricard, D., Richard, E., Riette, S., Rodier, Q., Schoetter, R., Seyfried, L., Stein, J., Suhre, K., Taufour, M., Thouron, O., Turner, S., Verrelle, A., Vié, B., Visentin, F., Vionnet, V., and Wautelet, P. (2018). Overview of the Meso-NH model version 5.4 and its applications. *Geoscientific Model Development*, 11(5):1929–1969. DOI: 10.5194/gmd-11-1929-2018.

Lamberti, G. and Gorlé, C. (2021). A multi-fidelity machine learning framework to predict wind loads on buildings. *Journal of Wind Engineering and Industrial Aerodynamics*, 214:104647. ISSN 0167-6105. DOI: 10.1016/j.jweia.2021.104647.

Largeron, C., Dumont, M., Morin, S., Boone, A., Lafaysse, M., Metref, S., Cosme, E., Jonas, T., Winstral, A., and Margulis, S. A. (2020). Toward snow cover estimation in mountainous areas using modern data assimilation methods: A review. *Frontiers in Earth Science*, 8. ISSN 2296-6463. DOI: 10.3389/feart.2020.00325.

Larsson, J., Kawai, S., Bodart, J., and Bermejo-Moreno, I. (2016). Large eddy simulation with modeled wall-stress: recent progress and future directions. *Mechanical Engineering Reviews*, 3(1):15–00418–15–00418. DOI: 10.1299/mer.15-00418.

Lassila, T., Manzoni, A., Quarteroni, A., and Rozza, G. (2014). Model order reduction in fluid dynamics: challenges and perspectives. *Reduced Order Methods for modeling and computational reduction*, pages 235–273. DOI: 10.1007/978-3-319-02090-7_9.

Leelőssy, A., Molnár, F., Izsák, F., Havasi, A., Lagzi, I., and Mészáros, R. (2014). Dispersion modeling of air pollutants in the atmosphere: a review. *Cent. Eur. J. Geosci.*, 6(3):257–278. DOI: 10.2478/s13533-012-0188-6.

Letheren, B., Montes, G., Villa, T., and Gonzalez, F. (2016). Design and flight testing of a bio-inspired plume tracking algorithm for unmanned aerial vehicles. In *2016 IEEE Aerospace Conference*, pages 1–9. DOI: 10.1109/AERO.2016.7500614.

Li, W.-W. and Meroney, R. N. (1983)a. Gas dispersion near a cubical model building. Part I. Mean concentration measurements. *Journal of Wind Engineering and Industrial Aerodynamics*, 12(1):15–33. ISSN 0167-6105. DOI: 10.1016/0167-6105(83)90078-8.

Li, W.-W. and Meroney, R. N. (1983)b. Gas dispersion near a cubical model building. Part II. Concentration fluctuation measurements. *Journal of Wind Engineering and Industrial Aerodynamics*, 12(1):35–47. ISSN 0167-6105. DOI: 10.1016/0167-6105(83)90079-X.

Lilly, D. K. (1967). The representation of small-scale turbulence in numerical simulation experiments. In *Proc. IBM Sci. Comput. Symp. on Environmental Science*, pages 195–210. DOI: 10.5065/D62R3PMM.

Lilly, D. K. (1992). A proposed modification of the Germano subgrid-scale closure method. *Physics of Fluids A: Fluid Dynamics*, 4(3):633–635. ISSN 0899-8213. DOI: 10.1063/1.858280.

Liu, C., Xiao, Q., and Wang, B. (2008). An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test. *Monthly Weather Review*, 136(9):3363 – 3373. DOI: 10.1175/2008MWR2312.1.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528. DOI: 10.1007/BF01589116.

Liu, Y., Kochanski, A., Baker, K. R., Mell, W., Linn, R., Paugam, R., Mandel, J., Fournier, A., Jenkins, M. A., Goodrick, S., Achtemeier, G., Zhao, F., Ottmar, R., French, N. H. F., Larkin, N., Brown, T., Hudak, A., Dickinson, M., Potter, B., Clements, C., Urbanski, S., Prichard, S., Watts, A., and McNamara, D. (2019). Fire behaviour and smoke modelling: model improvement and measurement needs for next-generation smoke research and forecasting systems. *International Journal of Wildland Fire*, 28(8):570–588. DOI: 10.1071/WF18204.

Liu, Y., Haussaire, J.-M., Bocquet, M., Roustan, Y., Saunier, O., and Mathieu, A. (2017). Uncertainty quantification of pollutant source retrieval: comparison of Bayesian methods with application to the Chernobyl and Fukushima Daiichi accidental releases of radionuclides. *Quarterly Journal of the Royal Meteorological Society*, 143(708):2886–2901. DOI: 10.1002/qj.3138.

Lucas, D. D., Gowardhan, A., Cameron-Smith, P., and Baskett, R. L. (2016). Impact of meteorological inflow uncertainty on tracer transport and source estimation in urban atmospheres. *Atmospheric Environment*, 143:120–132. ISSN 1352-2310. DOI: 10.1016/j.atmosenv.2016.08.019.

Lumet, E., Jaravel, T., Rochoux, M. C., Vermorel, O., and Lacroix, S. (2024). Assessing the internal variability of Large-Eddy Simulations for microscale pollutant dispersion prediction in an idealized urban environment. *Boundary-Layer Meteorology*, 190(2):9. ISSN 1573-1472. DOI: 10.1007/s10546-023-00853-7.

Macdonald, R. and Ejim, C. (2002). Flow and dispersion data from a hydraulic simulation of the MUST array. Technical report, Department of Mechanical Engineering, University of Waterloo.

Manfreda, S., McCabe, M. F., Miller, P. E., Lucas, R., Pajuelo Madrigal, V., Mallinis, G., Ben Dor, E., Helman, D., Estes, L., Ciraolo, G., Müllerová, J., Tauro, F., De Lima, M. I., De Lima, J. L. M. P., Maltese, A., Frances, F., Caylor, K., Kohv, M., Perks, M., Ruiz-Pérez, G., Su, Z., Vico, G., and Toth, B. (2018). On the use of unmanned aerial systems for environmental monitoring. *Remote Sensing*, 10(4). ISSN 2072-4292. DOI: 10.3390/rs10040641.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., and Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, 8. ISSN 2296-2565. DOI: 10.3389/fpubh.2020.00014.

Mardia, K. V. and Zemroch, P. J. (1975). Algorithm as 86: The von mises distribution function. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2): 268–272. ISSN 00359254, 14679876. URL `http://www.jstor.org/stable/2346578`. Accessed: 2023-09-20.

Margheri, L. and Sagaut, P. (2016). A hybrid anchored-ANOVA – POD/Kriging method for uncertainty quantification in unsteady high-fidelity CFD simulations. *Journal of Computational Physics*, 324:137–173. ISSN 0021-9991. DOI: 10.1016/j.jcp.2016.07.036.

Maronga, B., Knigge, C., and Raasch, S. (2020). An improved surface boundary condition for large-eddy simulations based on Monin–Obukhov similarity theory: Evaluation and consequences for grid convergence in neutral and stable conditions. *Boundary-Layer Meteorology*, 174:297–325. DOI: 10.1007/s10546-019-00485-w.

Martin, M. J., Balmaseda, M., Bertino, L., Brasseur, P., Brassington, G., Cummings, J., Fujii, Y., Lea, D., Lellouche, J.-M., Mogensen, K., and other. (2015). Status and future of data assimilation in operational oceanography. *Journal of Operational Oceanography*, 8(sup1):s28–s48. DOI: 10.1080/1755876X.2015.1022055.

Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78. DOI: 10.1080/01621459.1951.10500769.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., C., P., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B. (2023). *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. DOI: 10.1017/9781009157896.

McKay, M. D. (1995). Evaluating prediction uncertainty. Technical report, Nuclear Regulatory Commission. URL `https://inis.iaea.org/collection/NCLCollectionStore/_Public/26/051/26051087.pdf`. Accessed: 2023-12-01.

Meroney, R. N., Leitl, B. M., Rafailidis, S., and Schatzmann, M. (1999). Wind-tunnel and numerical modeling of flow and dispersion about several building shapes. *J. Wind. Eng. Ind. Aerodyn.*, 81(1):333–345. DOI: 10.1016/S0167-6105(99)00028-8.

Meyers, J., Geurts, B., and Sagaut, P. (2008). *Quality and reliability of large-eddy simulations*, volume 12. Springer Science & Business Media. DOI: 10.1007/978-1-4020-8578-9.

Milano, M. and Koumoutsakos, P. (2002). Neural network modeling for near wall turbulent flow. *Journal of Computational Physics*, 182(1):1–26. ISSN 0021-9991. DOI: 10.1006/jcph.2002.7146.

Milliez, M. (2006). *Micrometeorological modelling in urban areas: polluant dispersion and radiative effects modelling*. PhD thesis, École des Ponts ParisTech. URL `https://theses.hal.science/pastel-00004042/`. Accessed: 2023-12-01.

Milliez, M. and Carissimo, B. (2007). Numerical simulations of pollutant dispersion in an idealized urban area, for different meteorological conditions. *Boundary-Layer Meteorology*, 122(2):321–342. DOI: 10.1007/s10546-006-9110-4.

Misaka, T., Ogasawara, T., Obayashi, S., Yamada, I., and Okuno, Y. (2008). Assimilation experiment of lidar measurements for wake turbulence. *Journal of Fluid Science and Technology*, 3(4):512–518. DOI: 10.1299/jfst.3.512.

Miyagusuku, R., Yamashita, A., and Asama, H. (2015). Gaussian processes with input-dependent noise variance for wireless signal strength-based localization. In *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 1–6. DOI: 10.1109/SSRR.2015.7442993.

Monin, A. and Obukhov, A. (1954). Basic laws of turbulent mixing in the surface layer of the atmosphere. *Contrib. Geophys. Inst. Acad. Sci. USSR*, 24(151):163–187. URL `https://moodle2.units.it/pluginfile.php/267453/mod_resource/content/1/ABL_lecture_13.pdf`. Accessed: 2023-12-01.

Mons, V., Margheri, L., Chassaing, J.-C., and Sagaut, P. (2017). Data assimilation-based reconstruction of urban pollutant release characteristics. *Journal of Wind Engineering and Industrial Aerodynamics*, 169:232–250. ISSN 0167-6105. DOI: 10.1016/j.jweia.2017.07.007.

Mons, V. and Marquet, O. (2021). Linear and nonlinear sensor placement strategies for mean-flow reconstruction via data assimilation. *Journal of Fluid Mechanics*, 923. DOI: 10.1017/jfm.2021.488.

Montazeri, H. and Blocken, B. (2013). CFD simulation of wind-induced pressure coefficients on buildings with and without balconies: Validation and sensitivity analysis. *Build Environ.*, 60:137–149. ISSN 0360-1323. DOI: 10.1016/j.buildenv.2012.11.012.

Montes, G., Letheren, B., Villa, T. F., and Gonzalez, F. (2014). Bio-inspired plume tracking algorithm for uavs. In *Proceedings of the 16th Australasian Conference on Robotics and Automation 2014*, pages 1–8. Australian Robotics and Automation Association (ARAA), Australia. URL `https://eprints.qut.edu.au/81610/`. Accessed: 2023-12-01.

Moonen, P., Dorer, V., and Carmeliet, J. (2012). Effect of flow unsteadiness on the mean wind flow pattern in an idealized urban environment. *J. Wind. Eng. Ind. Aerodyn.*, 104:389–396. DOI: 10.1016/j.jweia.2012.01.007.

Muñoz-Esparza, D., Kosović, B., Mirocha, J., and van Beeck, J. (2014). Bridging the transition from mesoscale to microscale turbulence in numerical weather prediction models. *Boundary-Layer Meteorology*, 153(3):409–440. DOI: 10.1007/s10546-014-9956-9.

Muñoz-Esparza, D., Shin, H. H., Sauer, J. A., Steiner, M., Hawbecker, P., Boehnert, J., Pinto, J. O., Kosović, B., and Sharman, R. D. (2021). Efficient graphics processing unit modeling of street-scale weather effects in support of aerial operations in the urban environment. *AGU Advances*, 2(2):e2021AV000432. DOI: 10.1029/2021AV000432.

Munters, W., Meneveau, C., and Meyers, J. (2016). Turbulent inflow precursor method with time-varying direction for large-eddy simulations and applications to wind farms. *Boundary-Layer Meteorology*, 159(2): 305–328. DOI: 10.1007/s10546-016-0127-z.

Murata, T., Fukami, K., and Fukagata, K. (2020). Nonlinear mode decomposition with convolutional neural networks for fluid dynamics. *Journal of Fluid Mechanics*, 882: A13. DOI: 10.1017/jfm.2019.822.

Nagel, T., Schoetter, R., Masson, V., Lac, C., and Carissimo, B. (2022). Numerical analysis of the atmospheric boundary-layer turbulence influence on microscale transport of pollutant in an idealized urban environment. *Boundary-Layer Meteorology*, 184(1): 113–141. DOI: 10.1007/s10546-022-00697-7.

Nagel, T., Schoetter, R., Bourgin, V., Masson, V., and Onofri, E. (2023). Drag coefficient and turbulence mixing length of local climate zone-based urban morphologies derived using obstacle-resolving modelling. *Boundary-Layer Meteorology*, 186(3):737–769. DOI: 10.1007/s10546-022-00780-z.

Namdeo, A., Colls, J., and Baker, C. (1999). Dispersion and re-suspension of fine and coarse particulates in an urban street canyon. *Science of The Total Environment*, 235 (1):3–13. ISSN 0048-9697. DOI: 10.1016/S0048-9697(99)00185-0.

Nazarian, N., Krayenhoff, E. S., and Martilli, A. (2020). A one-dimensional model of turbulent flow through "urban" canopies (mlucm v2.0): updates based on large-eddy simulation. *Geosci. Model Dev.*, 13(3):937–953. DOI: 10.5194/gmd-13-937-2020.

Neophytou, M., Gowardhan, A., and Brown, M. (2011). An inter-comparison of three urban wind models using Oklahoma City Joint Urban 2003 wind field measurements. *Journal of Wind Engineering and Industrial Aerodynamics*, 99(4):357–368. DOI: 10.1016/j.jweia.2011.01.010.

Neumann, P. P., Bennetts, V. H., Lilienthal, A. J., Bartholmai, M., and Schiller, J. H. (2013). Gas source localization with a micro-drone using bio-inspired

and particle filter-based algorithms. *Advanced Robotics*, 27(9):725–738. DOI: 10.1080/01691864.2013.779052.

Nicoud, F. and Ducros, F. (1999). Subgrid-scale stress modelling based on the square of the velocity gradient tensor. *Flow Turbul. Combust.*, 62(3):183–200. DOI: 10.1023/A:1009995426001.

Nony, B. X., Rochoux, M. C., Jaravel, T., and Lucor, D. (2023)a. Reduced-order modeling for parameterized large-eddy simulations of atmospheric pollutant dispersion. *Stoch. Environ. Res. Risk Assess.*, 37(6):2117–2144. ISSN 1436-3259. DOI: 10.1007/s00477-023-02383-7.

Nony, B. X. (2023). *Reduced-order models under uncertainties for microscale atmospheric pollutant dispersion in urban areas: exploring learning algorithms for high-fidelity model emulation*. Phd thesis, Université de Toulouse, France.

Nony, B. X., Rochoux, M. C., Jaravel, T., and Lucor, D. (2023)b. Reduced-order model for microscale atmospheric dispersion combining multi-fidelity LES and RANS data. In *UNCECOMP 2023 5th ECCOMAS Thematic Conference on ncertainty Quantification in Computational Sciences and Engineering*. URL `https://perso.limsi.fr/lucor/pdf/proc_uncecomp2023.pdf`. Accessed: 2023-12-01.

Oke, T. R. (1987). *Boundary layer climates*. Routledge. DOI: 10.4324/9780203407219.

Olesen, H. R., Baklanov, A., Bartzis, J., Barmpas, F., Berkowicz, R., Brzozowski, K., Buccolieri, R., Carissimo, B., Costa, A., Di Sabatino, S., et al. (2008). The MUST model evaluation exercise: Patterns in model performance. Technical Report 43/1, Hrvatsko meteorološko društvo. URL `https://hrcak.srce.hr/file/96396`. Accessed: 2023-12-01.

Orsi, M., Soulhac, L., Feraco, F., Marro, M., Rosenberg, D., Marino, R., Boffadossi, M., and Salizzoni, P. (2021). Scalar mixing in homogeneous isotropic turbulence: A numerical study. *Phys. Rev. Fluids*, 6:034502. DOI: 10.1103/PhysRevFluids.6.034502.

Owen, A. B. (2020). On dropping the first Sobol' point. In *International conference on Monte Carlo and quasi-Monte Carlo methods in scientific computing*, pages 71–86. Springer. DOI: 10.1007/978-3-030-98319-2_4.

Paoli, R., Poubeau, A., and Cariolle, D. (2020). Large-eddy simulations of a reactive solid rocket motor plume. *AIAA Journal*, 58(4):1639–1656. DOI: 10.2514/1.J058601.

Pasquier, M., Jay, S., Jacob, J., and Sagaut, P. (2023). A Lattice-Boltzmann-based modelling chain for traffic-related atmospheric pollutant dispersion at the local urban scale. *Building and Environment*, 242:110562. ISSN 0360-1323. DOI: 10.1016/j.buildenv.2023.110562.

Passot, T. and Pouquet, A. (1987). Numerical simulation of compressible homogeneous flows in the turbulent regime. *J. Fluid Mech.*, 181:441–466. DOI: 10.1017/S0022112087002167.

Patrikar, J., Moon, B. G., and Scherer, S. (2020). Wind and the city: Utilizing UAV-based in-situ measurements for estimating urban wind fields. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1254–1260. DOI: 10.1109/IROS45743.2020.9340812.

Paugam, R., Wooster, M. J., and Roberts, G. (2013). Use of handheld thermal imager data for airborne mapping of fire radiative power and energy and flame front rate of spread. *IEEE Transactions on Geoscience and Remote Sensing*, 51(6):3385–3399. DOI: 10.1109/TGRS.2012.2220368.

Pavageau, M. and Schatzmann, M. (1999). Wind tunnel measurements of concentration fluctuations in an urban street canyon. *Atmospheric Environment*, 33(24):3961–3971. ISSN 1352-2310. DOI: 10.1016/S1352-2310(99)00138-7.

Peng, L., Lipinski, D., and Mohseni, K. (2014). Dynamic data driven application system for plume estimation using UAVs. *Journal of Intelligent & Robotic Systems*, 74(1): 421–436. ISSN 1573-0409. DOI: 10.1007/s10846-013-9964-x.

Pérez Arroyo, C., Dombard, J., Duchaine, F., Gicquel, L., Odier, N., Exilard, G., Richard, S., Buffaz, N., and Démolis, J. (2020). Large-eddy simulation of an integrated high-pressure compressor and combustion chamber of a typical turbine engine architecture. In *Volume 2C: Turbomachinery*, Turbo Expo: Power for Land, Sea, and Air, page V02CT35A058. DOI: 10.1115/GT2020-16288.

Perks, M. T., Russell, A. J., and Large, A. R. G. (2016). Technical Note: Advances in flash flood monitoring using unmanned aerial vehicles (UAVs). *Hydrology and Earth System Sciences*, 20(10):4005–4015. DOI: 10.5194/hess-20-4005-2016.

Perry, S. G., Cimorelli, A. J., Paine, R. J., Brode, R. W., Weil, J. C., Venkatram, A., Wilson, R. B., Lee, R. F., and Peters, W. D. (2005). AERMOD: A dispersion model for industrial source applications. Part II: Model performance against 17 field study databases. *J. Appl. Meteorol.*, 44(5):694–708. DOI: 10.1175/JAM2228.1.

Pfeffer, H.-U., Friesel, J., Elbers, G., Beier, R., and Ellermann, K. (1995). Air pollution monitoring in street canyons in North Rhine-Westphalia, Germany. *Science of The Total Environment*, 169(1):7–15. ISSN 0048-9697. DOI: 10.1016/0048-9697(95)04627-D.

Philips, D. A., Rossi, R., and Iaccarino, G. (2013). Large-eddy simulation of passive scalar dispersion in an urban-like canopy. *Journal of Fluid Mechanics*, 723:404–428. DOI: 10.1017/jfm.2013.135.

Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R. T., and Kim, N.-H. (2010). Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design*, 132(7):071008. ISSN 1050-0472. DOI: 10.1115/1.4001873.

Pimont, F., Dupuy, J.-L., Linn, R. R., Parsons, R., and Martin-StPaul, N. (2017). Representativeness of wind measurements in fire experiments: Lessons learned from large-eddy simulations in a homogeneous forest. *Agricultural and Forest Meteorology*, 232:479–488. ISSN 0168-1923. DOI: 10.1016/j.agrformet.2016.10.002.

Piomelli, U. (1999). Large-eddy simulation: achievements and challenges. *Prog. Aerosp. Sci.*, 35(4):335–362. ISSN 0376-0421. DOI: 10.1016/S0376-0421(98)00014-1.

Piomelli, U. (2020). Large-eddy simulation of turbulent flows. part 1: Introduction. von Kármán Institute for Fluid Dynamics Lecture Series Notes.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10 (3):61–74. URL `https://home.cs.colorado.edu/~mozer/Teaching/syllabi/6622/papers/Platt1999.pdf`. Accessed: 2023-12-01.

Poinsot, T. and Lele, S. (1992). Boundary conditions for direct simulations of compressible viscous flows. *Journal of Computational Physics*, 101(1):104–129. ISSN 0021-9991. DOI: 10.1016/0021-9991(92)90046-2.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *J. Am. Stat. Assoc.*, 89(428):1303–1313. DOI: 10.1080/01621459.1994.10476870.

Pope, S. B. (2000). *Turbulent Flows*. Cambridge University Press. DOI: 10.1017/CBO9780511840531.

Poubeau, A., Paoli, R., and Cariolle, D. (2016). Evaluation of afterburning chemistry in solid-rocket motor jets using an off-line model. *J. Spacecr. Rockets*, 53(2):380–388. DOI: 10.2514/1.A33311.

Qian, E., Kramer, B., Peherstorfer, B., and Willcox, K. (2020). Lift & Learn: Physics-informed machine learning for large-scale nonlinear dynamical systems. *Physica D: Nonlinear Phenomena*, 406:132401. ISSN 0167-2789. DOI: 10.1016/j.physd.2020.132401.

Qian, K. and Claudel, C. (2020). Real-time mobile sensor management framework for city-scale environmental monitoring. *Journal of Computational Science*, 45:101205. ISSN 1877-7503. DOI: 10.1016/j.jocs.2020.101205.

Quillatre, P. (2014). *Simulation aux grandes échelles d'explosions en domaine semi-confiné*. PhD thesis, Université de Toulouse, France. URL `https://oatao.univ-toulouse.fr/11851/1/quillatre.pdf`. Accessed: 2023-12-01.

Rabier, F. (2005). Overview of global data assimilation developments in numerical weather-prediction centres. *Quarterly Journal of the Royal Meteorological Society*, 131 (613):3215–3233. DOI: 10.1256/qj.05.129.

Rai, R., Rajput, M., Agrawal, M., and Agrawal, S. (2011). Gaseous air pollutants: a review on current and future trends of emissions and impact on agriculture. *Journal of Scientific Research*, 55(771):1. URL `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=39485a602b7c7201e2c70a4a1ff79605606712ac`. Accessed: 2023-12-01.

Ramshaw, J., O'Rourke, P., and Amsden, A. (1986). Acoustic damping for explicit calculations of fluid flow at low Mach number. Technical report no. LA–10641-MS, Los Alamos National Laboratories, USA. URL `https://inis.iaea.org/collection/NCLCollectionStore/_Public/17/074/17074782.pdf`. Accessed: 2023-12-01.

Rasmussen, C. E., Williams, C. K., et al. (2006). *Gaussian processes for machine learning*, volume 1. Springer. DOI: 10.7551/mitpress/3206.001.0001.

Raupach, M. R., Antonia, R. A., and Rajagopalan, S. (1991). Rough-wall turbulent boundary layers. *Applied Mechanics Reviews*, 44(1):1–25. ISSN 0003-6900. DOI: 10.1115/1.3119492.

Raynaud, L., Berre, L., and Desroziers, G. (2011). An extended specification of flow-dependent background error variances in the Météo-France global 4D-Var system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):607–619. DOI: 10.1002/qj.795.

Refloch, A., Courbet, B., Murrone, A., Villedieu, P., Laurent, C., Gilbank, P., Troyes, J., Tessé, L., Chaineray, G., Dargaud, J., Quémerais, E., and Vuillot, F. (2011). CEDRE Software. *Aerospace Lab*. URL `https://hal.science/hal-01182463`. Accessed: 2023-12-01.

Reidmiller, D. R., Avery, C. W., Easterling, D. R., Kunkel, K. E., Lewis, K. L. M., Maycock, T. K., and Stewart, B. C. (2017). *Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, Volume II*. US global change research program. DOI: 10.7930/NCA4.2018.

Reymann, C., Renzaglia, A., Lamraoui, F., Bronz, M., and Lacroix, S. (2018). Adaptive sampling of cumulus clouds with a fleet of UAVs. *Autonomous robots*, 42(2):1–22. ISSN 1573-7527. DOI: 10.1007/s10514-017-9625-1.

Reynolds, W. C. (1990). The potential and limitations of direct and large eddy simulations. In *Whither Turbulence? Turbulence at the Crossroads: Proceedings of a Workshop Held at Cornell University, Ithaca, NY, March 22–24, 1989*, pages 313–343. DOI: 10.1007/3-540-52535-1_52.

Richards, P. and Hoxey, R. (1993). Appropriate boundary conditions for computational wind engineering models using the k–ϵ turbulence model. *Journal of Wind Engineering and Industrial Aerodynamics*, 46-47:145–153. ISSN 0167-6105. DOI: 10.1016/0167-6105(93)90124-7. Proceedings of the 1st International on Computational Wind Engineering.

Risser, M. D., Paciorek, C. J., Wehner, M. F., O'Brien, T. A., and Collins, W. D. (2019). A probabilistic gridded product for daily precipitation extremes over the United States. *Clim. Dyn.*, 53(5):2517–2538. ISSN 1432-0894. DOI: 10.1007/s00382-019-04636-0.

Ritchie, H. and Roser, M. (2017). Air pollution. URL `https://ourworldindata.org/air-pollution`. Website. Accessed: 2023-10-03.

Rochoux, M. C., Ricci, S., Lucor, D., Cuenot, B., and Trouvé, A. (2014)a. Towards predictive data-driven simulations of wildfire spread – Part I: Reduced-cost Ensemble Kalman Filter based on a Polynomial Chaos surrogate model for parameter estimation. *Natural Hazards and Earth System Sciences*, 14(11):2951–2973. DOI: 10.5194/nhess-14-2951-2014.

Rochoux, M., Collin, A., Zhang, C., Trouvé, A., Lucor, D., and Moireau, P. (2018). Front shape similarity measure for shape-oriented sensitivity analysis and data assimilation for eikonal equation. *ESAIM: ProcS*, 63:258–279. DOI: 10.1051/proc/201863258.

Rochoux, M., Lumet, E., Thouron, L., Rea, G., Auguste, F., Jaravel, T., and Vermorel, O. (2021). Large-eddy simulation multi-model comparison of the MUST trial 2681829. Technical Report TR-CMGC-21-72, CERFACS, Toulouse, France.

Rochoux, M. et al. (2014)b. *Vers une meilleure prévision de la propagation d'incendies de forêt: evaluation de modèles et assimilation de données*. PhD thesis, École Centrale de Paris. URL `https://www.theses.fr/2014ECAP0009/abes`. Accessed: 2023-12-01.

Rogelj, J., Jiang, K., and Shindell, D. (2022). Mitigation pathways compatible with 1.5°C in the context of sustainable development. In *Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*, page 93–174. Cambridge University Press. DOI: 10.1017/9781009157940.004.

Rokach, L. and Maimon, O. (2014). *Data Mining with Decision Trees*. World Scientific, 2nd edition. DOI: 10.1142/9097.

Rossi, R., Philips, D., and Iaccarino, G. (2010). A numerical study of scalar dispersion downstream of a wall-mounted cube using direct simulations and algebraic flux models. *International Journal of Heat and Fluid Flow*, 31(5):805–819. ISSN 0142-727X. DOI: 10.1016/j.ijheatfluidflow.2010.05.006. Sixth International Symposium on Turbulence, Heat and Mass Transfer, Rome, Italy, 14-18 September 2009.

Rotach, M. W., Gryning, S.-E., Batchvarova, E., Christen, A., and Vogt, R. (2004). Pollutant dispersion close to an urban surface – the BUBBLE tracer experiment. *Meteorology and Atmospheric Physics*, 87(1):39–56. DOI: 10.1007/s00703-003-0060-9.

Ruckstuhl, Y. and Janjić, T. (2020). Combined state-parameter estimation with the LETKF for convective-scale weather forecasting. *Monthly Weather Review*, 148(4): 1607 – 1628. DOI: 10.1175/MWR-D-19-0233.1.

Sagaut, P. (2005). *Large eddy simulation for incompressible flows: an introduction*. Springer Science & Business Media. DOI: 10.1007/b137536.

Salim, M. S., Buccolieri, R., Chan, A., and Di Sabatino, S. (2011). Numerical simulation of atmospheric pollutant dispersion in an urban street canyon: Comparison between RANS and LES. *Journal of Wind Engineering and Industrial Aerodynamics*, 99(2): 103–113. ISSN 0167-6105. DOI: 10.1016/j.jweia.2010.12.002.

Saltelli, A. and Annoni, P. (2010). How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software*, 25(12):1508–1517. DOI: 10.1016/j.envsoft.2010.04.012.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. (2010). Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer physics communications*, 181(2):259–270. DOI: 10.1016/j.cpc.2009.09.018.

Santiago, J. L., Dejoan, A., Martilli, A., Martin, F., and Pinelli, A. (2010). Comparison between large-eddy simulation and Reynolds-Averaged Navier–Stokes computations for the MUST field experiment. Part I: Study of the flow for an incident wind directed perpendicularly to the front array of containers. *Boundary-Layer Meteorology*, 135(1): 109–132. DOI: 10.1007/s10546-010-9466-3.

Saunier, O., Didier, D., Mathieu, A., Masson, O., and Brazidec, J. D. L. (2019). Atmospheric modeling and source reconstruction of radioactive ruthenium from an undeclared major release in 2017. *Proceedings of the National Academy of Sciences*, 116 (50):24991–25000. DOI: 10.1073/pnas.1907823116.

Scaperdas, A.-S. (2000). *Modelling air flow and pollutant dispersion at urban canyon intersections.* PhD thesis, Imperial College London (University of London). URL `https://spiral.imperial.ac.uk/bitstream/10044/1/8803/1/Athena-Sophia_Scaperdas-2000-PhD-Thesis.pdf`. Accessed: 2023-12-01.

Schatzmann, M. and Leitl, B. (2011). Issues with validation of urban flow and dispersion CFD models. *J. Wind Eng. Ind. Aerodyn.*, 99(4):169–186. ISSN 0167-6105. DOI: 10.1016/j.jweia.2011.01.005. The Fifth International Symposium on Computational Wind Engineering.

Schatzmann, M., Olesen, H., and Franke, J. (2010). COST 732 model evaluation case studies: approach and results. Technical report, University of Hamburg, Meteorological Institute. URL https://www.researchgate.net/profile/George-Efthimiou-3/post/Has-fluent-been-compared-to-starccm/attachment/59d6585379197b80779ae4bd/AS%3A538043318628353%401505290931380/download/5th_Docu_May_10.pdf. Accessed: 2023-12-01.

Scheiner, S. M. and Gurevitch, J. (2001). *Design and analysis of ecological experiments*. Oxford University Press. DOI: 10.1201/9781003059813.

Schmid, P. J. (2022). Dynamic mode decomposition and its variants. *Annual Review of Fluid Mechanics*, 54(1):225–254. DOI: 10.1146/annurev-fluid-030121-015835.

Schmidt, O. T. and Colonius, T. (2020). Guide to spectral proper orthogonal decomposition. *AIAA journal*, 58(3):1023–1033. DOI: 10.2514/1.J058809.

Schönfeld, T. and Rudgyard, M. (1999). Steady and unsteady flow simulations using the hybrid flow solver AVBP. *AIAA journal*, 37(11):1378–1385. DOI: 10.2514/2.636.

Seinfeld, J. and Pandis, S. (1998). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley & Sons, Inc, Hoboken, New Jersey. DOI: 10.1021/ja985605y.

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games (AM-28)*, volume 2. Princeton University Press Princeton. DOI: 10.1515/9781400881970-018.

Sirovich, L. (1987). Turbulence and the dynamics of coherent structures. I. Coherent structures. *Quarterly of applied mathematics*, 45(3):561–571. DOI: 10.1090/qam/910462.

Smagorinsky, J. (1963). General circulation experiments with the primitive equations: I. The basic experiment. *Mon. Wea. Rev.*, 91(3):99–164. DOI: 10.1175/1520-0493(1963)091<0099:GCEWTP>2.3.CO;2.

Smirnov, A., Shi, S., and Celik, I. (2001). Random flow generation technique for large eddy simulations and particle-dynamics modeling. *J. Fluids Eng.*, 123(2):359–371. ISSN 0098-2202. DOI: 10.1115/1.1369598.

Smith, P. J., Thornhill, G. D., Dance, S. L., Lawless, A. S., Mason, D. C., and Nichols, N. K. (2013). Data assimilation for state and parameter estimation: application to morphodynamic modelling. *Quarterly Journal of the Royal Meteorological Society*, 139 (671):314–327. DOI: 10.1002/qj.1944.

Sobol', I. M. (1990). On sensitivity estimation for nonlinear mathematical models (in Russian). *Matematicheskoe modelirovanie*, 2(1):112–118. URL http://mi.mathnet.ru/mm2320. Accessed: 2023-12-01.

Sood, I., Simon, E., Vitsas, A., Blockmans, B., Larsen, G. C., and Meyers, J. (2022). Comparison of large eddy simulations against measurements from the Lillgrund offshore wind farm. *Wind Energy Sci.*, 7(6):2469–2489. DOI: 10.5194/wes-7-2469-2022.

Soulhac, L., Salizzoni, P., Mejean, P., Didier, D., and Rios, I. (2012). The model SIRANE for atmospheric urban pollutant dispersion. Part II: Validation of the model on a real case study. *Atmos. Environ.*, 49:320–337. DOI: 10.1016/j.atmosenv.2011.11.031.

Soulhac, L., Salizzoni, P., Cierco, F.-X., and Perkins, R. (2011). The model SIRANE for atmospheric urban pollutant dispersion. Part I: Presentation of the model. *Atmos. Environ.*, 45(39):7379–7395. DOI: 10.1016/j.atmosenv.2011.07.008.

Sousa, J. and Gorlé, C. (2019). Computational urban flow predictions with Bayesian inference: Validation with field data. *Build. Environ.*, 154:13–22. ISSN 0360-1323. DOI: 10.1016/j.buildenv.2019.02.028.

Sousa, J., García-Sánchez, C., and Gorlé, C. (2018). Improving urban flow predictions through data assimilation. *Build. Environ.*, 132:282–290. ISSN 0360-1323. DOI: 10.1016/j.buildenv.2018.01.032.

Specht, D. F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks*, 2(6):568–576. DOI: 10.1109/72.97934.

Spicer, T. O. and Tickle, G. (2021). Simplified source description for atmospheric dispersion model comparison of the Jack Rabbit II chlorine field experiments. *Atmospheric Environment*, 244:117866. ISSN 1352-2310. DOI: 10.1016/j.atmosenv.2020.117866.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Series in Statistics. Springer Science & Business Media. DOI: 10.1007/978-1-4612-1494-6.

Stull, R. B. (1988). *An introduction to boundary layer meteorology*, volume 13 of *Atmospheric and Oceanographic Sciences Library*. Springer Science & Business Media. DOI: 10.1007/978-94-009-3027-8.

Sudret, B. (2008). Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 93(7):964–979. ISSN 0951-8320. DOI: 10.1016/j.ress.2007.04.002. Bayesian Networks in Dependability.

Swischuk, R., Mainini, L., Peherstorfer, B., and Willcox, K. (2019). Projection-based model reduction: Formulations for physics-based machine learning. *Computers & Fluids*, 179:704–717. ISSN 0045-7930. DOI: 10.1016/j.compfluid.2018.07.021.

Taira, K., Brunton, S. L., Dawson, S. T. M., Rowley, C. W., Colonius, T., McKeon, B. J., Schmidt, O. T., Gordeyev, S., Theofilis, V., and Ukeiley, L. S. (2017). Modal analysis of fluid flows: An overview. *AIAA Journal*, 55(12):4013–4041. DOI: 10.2514/1.J056060.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9780898717921.

Thorpe, R. L., McCrink, M., and Gregory, J. W. (2018). Measurement of unsteady gusts in an urban wind field using a UAV-based anemometer. In *2018 Applied Aerodynamics Conference*. DOI: 10.2514/6.2018-4218.

Tian, X., Xie, Z., and Sun, Q. (2011). A POD-based ensemble four-dimensional variational assimilation method. *Tellus A: Dynamic Meteorology and Oceanography*. DOI: 10.1111/j.1600-0870.2011.00529.x.

Tipping, M. E. and Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482. DOI: 10.1162/089976699300016728.

Tissot, G., Cordier, L., Benard, N., and Noack, B. R. (2013). 4D-variational data assimilation for POD reduced-order models. In *Eighth International Symposium on Turbulence and Shear Flow Phenomena*. Begel House Inc. DOI: 10.1615/TSFP8.870.

Tominaga, Y. and Stathopoulos, T. (2007). Turbulent Schmidt numbers for CFD analysis with various types of flowfield. *Atmospheric Environment*, 41(37):8091–8099. ISSN 1352-2310. DOI: 10.1016/j.atmosenv.2007.06.054.

Tominaga, Y. and Stathopoulos, T. (2009). Numerical simulation of dispersion around an isolated cubic building: Comparison of various types of k–ϵ models. *Atmospheric Environment*, 43(20):3200–3210. ISSN 1352-2310. DOI: 10.1016/j.atmosenv.2009.03.038.

Tominaga, Y. and Stathopoulos, T. (2010). Numerical simulation of dispersion around an isolated cubic building: Model evaluation of RANS and LES. *Building and Environment*, 45(10):2231–2239. ISSN 0360-1323. DOI: 10.1016/j.buildenv.2010.04.004.

Tominaga, Y. and Stathopoulos, T. (2013). CFD simulation of near-field pollutant dispersion in the urban environment: A review of current modeling techniques. *Atmos. Environ.*, 79:716–730. DOI: 10.1016/j.atmosenv.2013.07.028.

Tominaga, Y. and Stathopoulos, T. (2017). Steady and unsteady RANS simulations of pollutant dispersion around isolated cubical buildings: Effect of large-scale fluctuations on the concentration field. *Journal of Wind Engineering and Industrial Aerodynamics*, 165:23–33. ISSN 0167-6105. DOI: 10.1016/j.jweia.2017.02.001.

Tominaga, Y., Mochida, A., Yoshie, R., Kataoka, H., Nozu, T., Yoshikawa, M., and Shirasawa, T. (2008). AIJ guidelines for practical applications of CFD to pedestrian wind environment around buildings. *J. Wind Eng. Ind. Aerodyn.*, 96(10):1749–1761. ISSN 0167-6105. DOI: 10.1016/j.jweia.2008.02.058. 4th International Symposium on Computational Wind Engineering (CWE2006).

Toparlar, Y., Blocken, B., Maiheu, B., and van Heijst, G. (2017). A review on the CFD analysis of urban microclimate. *Renew. Sust. Energ. Rev.*, 80:1613–1640. DOI: 10.1016/j.rser.2017.05.248.

Trystram, G. (2022). Calcul haute performance et ordinateurs superpuissants : la course à l'«exascale» (in French). *The Conversation*. URL https://theconversation.com/calcul-haute-performance-et-ordinateurs-superpuissants-la-course-a-l-exascale-194084. Accessed: 2023-09-20.

Turner, D. B. (1969). Workbook of atmospheric dispersion estimates: an introduction to dispersion modeling. Technical report, U.S. Environmental Protection Agency, Washington, DC. URL https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=9100JEIO.TXT. Accessed: 2023-12-01.

Vachon, G. (2001). *Transferts des polluants des sources fixes et mobiles dans la canopée urbaine: évaluation expérimentale*. PhD thesis, Nantes.

van Driest, E. R. (1956). On turbulent flow near a wall. *Journal of the aeronautical sciences*, 23(11):1007–1011. DOI: 10.2514/8.3713.

Vardoulakis, S., Fisher, B. E., Pericleous, K., and Gonzalez-Flesca, N. (2003). Modelling air quality in street canyons: a review. *Atmospheric Environment*, 37(2):155–182. ISSN 1352-2310. DOI: 10.1016/S1352-2310(02)00857-9.

Vasaturo, R., Kalkman, I., Blocken, B., and van Wesemael, P. (2018). Large eddy simulation of the neutral atmospheric boundary layer: Performance evaluation of three inflow methods for terrains with different roughness. *J. Wind Eng. Ind. Aerodyn.*, 173: 241–261. DOI: 10.1016/j.jweia.2017.11.025.

Vendel, F. (2011). *Modélisation de la dispersion atmosphérique en présence d'obstacles complexes : application à l'étude de sites industriels*. PhD thesis, Ecole Centrale de Lyon. URL https://theses.hal.science/tel-00601470. Accessed: 2023-12-01.

Vermorel, O., Quillatre, P., and Poinsot, T. (2017). LES of explosions in venting chamber: A test case for premixed turbulent combustion models. *Combustion and Flame*, 183: 207–223. ISSN 0010-2180. DOI: 10.1016/j.combustflame.2017.05.014.

Vervecken, L., Camps, J., and Meyers, J. (2015). Stable reduced-order models for pollutant dispersion in the built environment. *Build. Environ.*, 92:360–367. DOI: 10.1016/j.buildenv.2015.05.008.

Vieira Da Rocha, T. and André, J.-M. (2021). État de l'art de la recherche scientifique sur l'impact climatique des traînées de condensation des avions (in French). Technical report, ADEME, CITEPA. URL https://librairie.ademe.fr/mobilite-et-transport/4617-etat-de-l-art-de-la-recherche-scientifique-sur-l-impact-climatique-des-trainees-de-condensation-des-avions.html. Accessed: 2023-12-01.

Villa, T. F., Gonzalez, F., Miljievic, B., Ristovski, Z. D., and Morawska, L. (2016)a. An overview of small unmanned aerial vehicles for air quality measurements: Present applications and future prospectives. *Sensors*, 16(7). ISSN 1424-8220. DOI: 10.3390/s16071072.

Villa, T. F., Salimi, F., Morton, K., Morawska, L., and Gonzalez, F. (2016)b. Development and validation of a UAV based system for air pollution measurements. *Sensors*, 16(12). ISSN 1424-8220. DOI: 10.3390/s16122202.

Vinuesa, R. and Brunton, S. L. (2022). Enhancing computational fluid dynamics with machine learning. *Nature Computational Science*, 2(6):358–366. ISSN 2662-8457. DOI: 10.1038/s43588-022-00264-7.

Väkevä, M., Hämeri, K., Kulmala, M., Lahdes, R., Ruuskanen, J., and Laitinen, T. (1999). Street level versus rooftop concentrations of submicron aerosol particles and gaseous pollutants in an urban street canyon. *Atmospheric Environment*, 33(9):1385–1397. ISSN 1352-2310. DOI: 10.1016/S1352-2310(98)00349-5.

Wang, G., Duchaine, F., Papadogiannis, D., Duran, I., Moreau, S., and Gicquel, L. Y. (2014). An overset grid method for large eddy simulation of turbomachinery stages. *Journal of Computational Physics*, 274:333–355. DOI: 10.1016/j.jcp.2014.06.006.

Wang, R., Kashinath, K., Mustafa, M., Albert, A., and Yu, R. (2020)a. Towards physics-informed deep learning for turbulent flow prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1457–1466, New York, NY, USA. Association for Computing Machinery. ISBN 9781450379984. DOI: 10.1145/3394486.3403198.

Wang, Y., Walters, S., Overton-Katz, N., Guzik, S. M., and Gao, X. (2020)b. CFD modeling of bluff-body stabilized premixed flames with data assimilation. In *AIAA Scitech 2020 Forum*. DOI: 10.2514/6.2020-0352.

Wiener, N. (1938). The homogeneous chaos. *American Journal of Mathematics*, 60(4): 897–936. DOI: 10.2307/2371268.

Wiersema, D. J., Lundquist, K. A., and Chow, F. K. (2020). Mesoscale to microscale simulations over complex terrain with the immersed boundary method in the weather research and forecasting model. *Mon. Weather Rev.*, 148(2):577–595. DOI: 10.1175/MWR-D-19-0071.1.

Winiarek, V., Bocquet, M., Saunier, O., and Mathieu, A. (2012). Estimation of errors in the inverse modeling of accidental release of atmospheric pollutant: Application to the reconstruction of the cesium-137 and iodine-131 source terms from the Fukushima Daiichi power plant. *Journal of Geophysical Research: Atmospheres*, 117(D5). DOI: 10.1029/2011JD016932.

Wise, D., Boppana, V., Li, K., and Poh, H. (2018). Effects of minor changes in the mean inlet wind direction on urban flow simulations. *Sustain. Cities Soc.*, 37:492–500. ISSN 2210-6707. DOI: 10.1016/j.scs.2017.11.041.

Wood, C. R., Arnold, S. J., Balogun, A. A., Barlow, J. F., Belcher, S. E., Britter, R. E., Cheng, H., Dobre, A., Lingard, J. J. N., Martin, D., Neophytou, M. K., Petersson,

F. K., Robins, A. G., Shallcross, D. E., Smalley, R. J., Tate, J. E., Tomlin, A. S., and White, I. R. (2009). Dispersion experiments in Central London: The 2007 DAPPLE project. *Bulletin of the American Meteorological Society*, 90(7):955 – 970. DOI: 10.1175/2009BAMS2638.1.

Wu, J., Cai, J., Yuan, S., Zhang, X., and Reniers, G. (2021). CFD and EnKF coupling estimation of LNG leakage and dispersion. *Safety Science*, 139:105263. ISSN 0925-7535. DOI: 10.1016/j.ssci.2021.105263.

Xiao, D., Heaney, C., Fang, F., Mottet, L., Hu, R., Bistrian, D., Aristodemou, E., Navon, I., and Pain, C. (2019). A domain decomposition non-intrusive reduced order model for turbulent flows. *Computers & Fluids*, 182:15–27. ISSN 0045-7930. DOI: 10.1016/j.compfluid.2019.02.012.

Xiao, H., Wu, J.-L., Wang, J.-X., Sun, R., and Roy, C. (2016). Quantifying and reducing model-form uncertainties in Reynolds-averaged Navier–Stokes simulations: A data-driven, physics-informed Bayesian approach. *Journal of Computational Physics*, 324: 115–136. ISSN 0021-9991. DOI: 10.1016/j.jcp.2016.07.038.

Xie, Y., Franz, E., Chu, M., and Thuerey, N. (2018). TempoGAN: A temporally coherent, volumetric GAN for super-resolution fluid flow. *ACM Trans. Graph.*, 37(4). ISSN 0730-0301. DOI: 10.1145/3197517.3201304.

Xie, Z. and Castro, I. P. (2006). LES and RANS for turbulent flow over arrays of wall-mounted obstacles. *Flow, Turbulence and Combustion*, 76(3):291–312. ISSN 1573-1987. DOI: 10.1007/s10494-006-9018-6.

Xin-Yi, G., Hu, S., De, X., Zheng-Tao, Z., Fei, S., and Hua-Bin, Y. (2018). An overview of contour detection approaches. *International Journal of Automation and Computing*, 15(IJAC-2017-10-257):656. ISSN 2731-538X. DOI: 10.1007/s11633-018-1117-z.

Yang, Y., Robinson, C., Heitz, D., and Mémin, E. (2015). Enhanced ensemble-based 4DVar scheme for data assimilation. *Computers & Fluids*, 115:201–210. ISSN 0045-7930. DOI: 10.1016/j.compfluid.2015.03.025.

Yee, E. and Biltoft, C. A. (2004). Concentration fluctuation measurements in a plume dispersing through a regular array of obstacles. *Boundary-Layer Meteorology*, 111(3): 363–415. DOI: 10.1023/B:BOUN.0000016496.83909.ee.

Yee, E., Gailis, R. M., Hill, A., Hilderman, T., and Kiel, D. (2006). Comparison of wind-tunnel and water-channel simulations of plume dispersion through a large array of obstacles with a scaled field experiment. *Boundary-Layer Meteorology*, 121(3):389–432. ISSN 1573-1472. DOI: 10.1007/s10546-006-9084-2.

Yue Yang, G.-W. H. and Wang, L.-P. (2008). Effects of subgrid-scale modeling on lagrangian statistics in large-eddy simulation. *Journal of Turbulence*, 9:N8. DOI: 10.1080/14685240801905360.

Zhang, C., Collin, A., Moireau, P., Trouvé, A., and Rochoux, M. C. (2019). State-parameter estimation approach for data-driven wildland fire spread modeling: Application to the 2012 rxcadre s5 field-scale experiment. *Fire Safety Journal*, 105:286–299. ISSN 0379-7112. DOI: 10.1016/j.firesaf.2019.03.009.