THÈSE

DEVELOPMENT OF MULTIGRID SEQUENTIAL DATA ASSIMILATION STRATEGIES FOR COMPLEX UNSTEADY FLOWS

Presentée par

Gabriel Ionut MOLDOVAN

Pour l'obtention du grade de

DOCTEUR DE L'ÉCOLE NATIONALE SUPÉRIEURE DE MÉCANIQUE ET D'AÉROTECHNIQUE

(Diplôme National - Arrêté du 25 Mai 2016)

École Doctorale

Sciences et Ingenierie des Matériaux, Mécanique, Énergétique

Secteur de Recherche : *Mécanique des milieux fluides*

Sous la direction de **M. Marcello MELDI**

Co-encadrée par **M. Guillaume LEHNASCH**

- JURY -

RAPPORTEURS

Mme. Ivette Maria RODRÍGUEZ PÉREZ - *Professeur*,	UPC
M. Etienne MEMIN - *Directeur de Recherche*,	INRIA

MEMBRES

M. Marcello MELDI - *Professeur*,	LMFL-ENSAM
M. Guillaume LEHNASCH - *Maître de Conférence*,	Institut Pprime, ENSMA
M. Laurent CORDIER - *Directeur de recherche*,	Institut Pprime, CNRS
Mme. Maria Vittoria SALVETTI - *Professeur*,	Università di Pisa

INVITÉS

M. Frank HERVY - *Docteur*,	DGA

Abstract

The analysis and control of complex high-Reynolds-number flows of industrial and practical interest is one of the most distinctive open challenges that the scientific community must face for fluid mechanics applications in the coming decades. Modelling bias and uncertainty may strongly affect the predictive capabilities of both numerical simulations and experimental measurements. Under this perspective, data-driven tools from Data Assimilation, and, in particular, sequential tools such as the ensemble Kalman filter (EnKF), have been recently used to obtain a precise estimation of the physical flow state accounting for bias or uncertainty coming from real conditions in the performance of the investigative tool.

A newly developed sequential Data Assimilation algorithm, combining multi-grid aspects and the ensemble Kalman Filter, is presented in this PhD-study. The so-called MGEnKF algorithm (Multi-Grid Ensemble Kalman Filter) exploits physical states obtained on multiple grids of different resolution to perform state and parameter estimation procedures using the EnKF. More precisely, an ensemble of low-fidelity simulations of the flow is run on a coarse grid level together with a single higher-resolution simulation on the finest mesh level. The state estimation obtained at the coarse level and the associated ensemble statistics are used to filter the finest mesh solution and to optimize a set of parameters describing the model (boundary conditions, model parameters...). This procedure allows to i) reduce the computational costs of the EnKF and ii) ensure the conservativity and the smoothness of the final solution.

The assessment of the method is performed via the analysis of one-dimensional, two-dimensional and three-dimensional test cases, using different models of increasing complexity. The results show that the MGEnKF can successfully update the state of a system with the available observations to increase the global accuracy of the model. In addition, the parametric description of the numerical problem (in terms of prescribed boundary conditions, turbulence closures...) can be adequately optimized taking into account the different mesh resolutions employed in the algorithm. The MGEnKF opens interesting perspectives for potential application to in-streaming Data Assimilation techniques.

Résumé

La simulation fidèle d'écoulements à nombre de Reynolds élevé, en configuration complexe d'intérêt industriel, reste un défi majeur à relever par la communauté de la mécanique des fluides. Les biais et incertitudes de modélisation peuvent d'une part fortement altérer les capacités de prévision des simulations numériques. Les mesures expérimentales sont d'autre part toujours incomplètes et affectées par du bruit provenant des systèmes de mesure. Les outils d'assimilation de données, et en particulier les outils séquentiels tels que le filtre de Kalman d'ensemble (EnKF), ont ainsi récemment été introduits afin d'améliorer l'estimation de l'état physique d'écoulements. Ils permettent de combiner les données provenant à la fois des simulations et des mesures expérimentales en prenant en compte l'incertitude associée à chaque source, mais leur coût de mise en œuvre reste généralement très important.

Un nouvel algorithme séquentiel d'assimilation de données, combinant approches multigrilles et filtre de Kalman d'ensemble, est proposé dans ce travail de thèse. L'algorithme MGEnKF (Multi-Grid Ensemble Kalman Filter) exploite des simulations numériques effectuées sur plusieurs grilles de différentes résolutions sur lesquelles l'estimation d'état et l'optimisation paramétrique sont effectuées par des procédures EnKF. Plus précisément, un ensemble de simulations basse-fidélité est exécuté sur un niveau de grille déraffiné tandis qu'une unique simulation à plus haute résolution est considérée sur la grille la plus fine. L'estimation d'état obtenue au niveau grossier et les statistiques d'ensemble associées sont utilisées pour filtrer la solution sur maille fine et optimiser un ensemble de paramètres décrivant le modèle (conditions aux limites, paramètres du modèle...). Cette procédure permet à la fois de réduire les coûts de calcul de la méthode EnKF et d'assurer la conservativité et la régularité de la solution finale.

L'évaluation de la méthode est réalisée via l'analyse de cas tests de complexité croissante, allant de l'advection linéaire monodimensionnelle à l'application des équations de Navier Stokes en configuration tridimensionnelle. Les résultats montrent que l'algorithme MGEnKF permet d'effectuer des estimations d'état à coût très réduit avec des observations disponibles pour augmenter la précision globale du modèle. Par ailleurs, il est montré que la description paramétrique du problème numérique (en termes de conditions aux limites prescrites, de fermetures de turbulence...) peut être optimisée de manière adéquate en tenant compte des différentes résolutions de maillage employées dans l'algorithme. La méthode MGEnKF ouvre ainsi des perspectives intéressantes d'applications « in-streaming » de l'assimilation de données.

Acknowledgements

I cannot but reaffirm my deepest gratitude to my supervisors. Thank you Marcello, Guillaume and Laurent for having looked after me for all these four years. I look back on my first days as a PhD student and I am proud to say that I have come a long way thanks to you. I would also like to thank Maria-Vittoria and Alessandro for having welcomed me for a short stay in their research team in Pisa that greatly enriched my research activities.

I cannot forget to mention the many friends I made in the lab during this period: Arun, Muhitin, Armando, Yann, Florian, Clément, Ksenia, Di, Marcos, Lucas(IBM), Lucas(DA), Miguel, Yassine, Agostino, Antoine, Momo... Thank you guys for having made this period much more funny and more liveable. I will miss so much our totally non-serious coffee pauses and our nonsensical discussions.

Justine, thank you for always being there for me and for your unconditional support. I would also like to thank my parents and my sister for their continuous support.

This thesis would not have been possible without the financial support of the Direction générale de l'Armement (DGA) and the region Nouvelle-Aquitaine. I take this opportunity to express my deep gratitude for their funding.

Contents

List of Figures

List of Tables

Développement de stratégies d'assimilation de données séquentielles multigrilles pour les écoulements instationnaires complexes

Résumé : La simulation fidèle d'écoulements à nombre de Reynolds élevé en configuration complexe d'intérêt industriel, reste un défi majeur à relever par la communauté de la mécanique des fluides. Les biais et incertitudes de modélisation peuvent d'une part fortement altérer les capacités de prévision des simulations numériques. Les mesures expérimentales sont d'autre part toujours incomplètes et affectées par du bruit provenant des systèmes de mesure. Les outils d'assimilation de données, et en particulier les outils séquentiels tels que le filtre de Kalman d'ensemble (EnKF), ont ainsi récemment été introduits afin d'améliorer l'estimation de l'état physique d'écoulements. Ils permettent de combiner les données provenant à la fois des simulations et des mesures expérimentales en prenant en compte l'incertitude associée à chaque source, mais leur coût de mise en œuvre reste généralement très important.

Un nouvel algorithme séquentiel d'assimilation de données, combinant approches multigrilles et filtre de Kalman d'ensemble, est proposé dans ce travail de thèse. L'algorithme MGEnKF (MultiGrid Ensemble Kalman Filter) exploite des simulations numériques effectuées sur plusieurs grilles de différentes résolutions sur lesquelles l'estimation d'état et l'optimisation paramétrique sont effectuées par des procédures EnKF. Plus précisément, un ensemble de simulations basse-fidélité est exécuté sur un niveau de grille déraffiné tandis qu'une unique simulation à plus haute résolution est considérée sur la grille la plus fine. L'estimation d'état obtenue au niveau grossier et les statistiques d'ensemble associées sont utilisées pour filtrer la solution sur maille fine et optimiser un ensemble de paramètres décrivant le modèle (conditions aux limites, paramètres du modèle...). Cette procédure permet à la fois de réduire les coûts de calcul de la méthode EnKF et d'assurer la conservativité de la solution finale.

L'évaluation de la méthode est réalisée via l'analyse de cas tests de complexité croissante, allant de l'advection linéaire monodimensionnelle à l'application des équations de Navier Stokes en configuration 3D. Les résultats montrent que l'algorithme MGEnKF permet d'effectuer des estimations d'état à coût très réduit avec des observations disponibles pour augmenter la précision globale du modèle. Par ailleurs, il est montré que la description paramétrique du problème numérique (en termes de conditions aux limites prescrites, de fermetures de turbulence...) peut être optimisée de manière adéquate en tenant compte des différentes résolutions de maillage employées dans l'algorithme. La méthode MGEnKF ouvre ainsi des perspectives intéressantes d'applications « in-streaming » de l'assimilation de données.

Mots clés : Écoulement instationnaire (dynamique des fluides), Kalman, Filtrage de, Méthodes multigrilles (analyse numérique), Modélisation CFD, Navier-Stokes, Équations de–Solutions numériques, Assimilation de données, Filtre de Kalman d'ensemble, Algorithme MGEnKF (Multi-Grid Ensemble Kalman Filter)

Development of multigrid sequential Data Assimilation strategies for complex unsteady flows

Abstract: The analysis and control of complex high-Reynolds-number flows of industrial and practical interest is one of the most distinctive open challenges that the scientific community must face for fluid mechanics applications in the coming decades. Modelling bias and uncertainty may strongly affect the predictive capabilities of both numerical simulations and experimental measurements. Under this perspective, data-driven tools from Data Assimilation, and, in particular, sequential tools such as the ensemble Kalman filter (EnKF), have been recently used to obtain a precise estimation of the physical flow state accounting for bias or uncertainty coming from real conditions in the performance of the investigative tool.

A newly developed sequential Data Assimilation algorithm, combining multi-grid aspects and the ensemble Kalman Filter, is presented in this PhD-study. The so-called MGEnKF algorithm (Multi-Grid Ensemble Kalman Filter) exploits physical states obtained on multiple grids of different resolution to perform the state estimation and parametric optimization using EnKF procedures. More precisely, an ensemble of low-fidelity simulations of the flow is run on a coarse grid level together with a single higher-resolution simulation on the finest mesh level. The state estimation obtained at the coarse level and the associated ensemble statistics are used to filter the finest mesh solution and optimise a set of parameters describing the model (boundary conditions, model parameters...). This procedure allows to i) reduce the computational costs of the EnKF and ii) ensure the conservativity and the smoothness of the final solution.

The assessment of the method is performed via the analysis of one-dimensional, two-dimensional and three-dimensional test cases, using different models of increasing complexity. The results show that the MGEnKF can successfully update the state of a system with the available observations to increase the global accuracy of the model. In addition, the parametric description of the numerical problem (in terms of prescribed boundary conditions, turbulence closures...) can be adequately optimized taking into account the different mesh resolutions employed in the algorithm. The MGEnKF opens interesting perspectives for potential application to in-streaming Data Assimilation techniques.

Keywords: Unsteady flow (Fluid dynamics), Kalman filtering, Multigrid methods, Computational Fluid Dynamics, Navier-Stokes equations–Numerical solutions, Data Assimilation, Ensemble Kalman Filter, Multigrid Ensemble Kalman Filter (MGENKF) algorithm

Résume etendu

Introduction

Les simulations numériques en mécanique des fluides sont affectées par l'incertitude, et plus spécifiquement, par des erreurs systématiques provenant de simplifications de la géométrie, d'hypothèses de modélisation et d'erreurs de discrétisation numérique. De plus, les conditions aux limites peuvent difficilement reproduire les conditions réelles, car celles-ci ne peuvent pas être connues avec une précision parfaite. D'autre part, les mesures d'écoulement fournissent des informations de l'écoulement réel, mais elles sont généralement bruyantes et limitées dans le temps et l'espace. Les données numériques et les données expérimentales sont donc affectées par des biais et des incertitudes, ce qui entrave leur capacité à prédire/répliquer avec précision des écoulements complexes.

Dans ce contexte, l'assimilation de données peut être utilisée pour combiner les deux sources d'information et fournir une estimation de l'état qui minimise l'incertitude des outils conventionnels. Les simulations pourraient intégrer des données expérimentales afin de corriger les erreurs systématiques présentes dans les modèles numériques utilisés. Ce type de procédure a été employé et perfectionné pendant des décennies dans le milieu de la météorologie. Nous nous intéressons plus particulièrement aux outils d'assimilation séquentielle de données issus de la théorie d'estimation (ET) tels que le filtre de Kalman (KF), qui sont spécifiquement conçus pour mettre à jour progressivement un modèle en intégrant les données dès qu'elles sont disponibles.

La présente thèse s'attaque au problème de l'estimation séquentielle de l'état et des paramètres en mécanique des fluides. Dans ce domaine, il y a des défis particuliers auxquels on doit faire face lorsqu'on applique des algorithmes d'assimilation de données séquentielle pour combiner les données de modèle et d'observation :

1. Des problèmes de haute dimension qui aboutissent à des algorithmes impossibles à traiter.

2. Modèles hautement non linéaires (Navier-Stokes)

3. Grande sensibilité aux corrections brutes du modèle qui rompent la conservativité du modèle.

L'objectif de cette thèse est de développer un algorithme séquentiel d'assimilation de données qui peut être utilisé pour les écoulements instationnaires et avec un coût de calcul acceptable. Le filtre de Kalman d'ensemble (EnKF) est utilisé comme base pour l'algorithme d'estimation que nous voulons développer. Les deux principaux atouts du EnKF sont les suivants: il permet de traiter des problèmes non-linéaires et sa formulation conceptuelle simple assure une relative facilitée d'implémentation dans des codes déjà existants. L'EnKF est donc un candidat parfait pour l'assimilation des écoulements instationnaires. Il reste cependant deux obstacles importants : le coût de calcul et la conservativité. Nous tentons de surmonter ces deux problèmes à l'aide de la méthode multigrille.

Modèles numériques pour la mécanique des fluides

Cette thèse traite des problèmes de mécanique des fluides. Les équations qui régissent ce type de problèmes découlent des principes de conservation de la masse, de la quantité de mouvement et de l'énergie totale. On peut donc écrire un système fermé d'équations différentielles partielles pour modéliser le comportement des fluides :

1. Équation de continuité

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}.(\rho \mathbf{U}) = 0,$$

2. Équation de Navier-Stokes

$$\frac{\partial (\rho \mathbf{U})}{\partial t} + \boldsymbol{\nabla}.(\rho \mathbf{U}\mathbf{U}) = \rho \mathbf{g} - \boldsymbol{\nabla}\left(P + \frac{2}{3}\mu \boldsymbol{\nabla}.\mathbf{U}\right) + \boldsymbol{\nabla}.\left[\mu\left(\boldsymbol{\nabla}\mathbf{U} + (\boldsymbol{\nabla}\mathbf{U})^T\right)\right],$$

3. Équation de l'energie

$$\frac{\partial(\rho E)}{\partial t} + \boldsymbol{\nabla}.(\rho E \mathbf{U}) = \rho \mathbf{g}.\mathbf{U} - \boldsymbol{\nabla}.(P\mathbf{U}) - \boldsymbol{\nabla}.\left(\frac{2}{3}\mu(\boldsymbol{\nabla}.\mathbf{U})\mathbf{U}\right) + \boldsymbol{\nabla}.\left[\mu\left(\boldsymbol{\nabla}\mathbf{U} + (\boldsymbol{\nabla}\mathbf{U})^T\right).\mathbf{U}\right] + \boldsymbol{\nabla}.(\lambda \nabla T) + \rho S,$$

La simulation d'écoulements turbulents s'avère particulièrement complexe. Quand le nombre de Reynolds est élevé, la représentation précise des phénomènes physiques doit prendre en compte des interactions complexes à plusieurs échelles en raison du fort comportement non linéaire de l'écoulement. De plus, avec l'augmentation du nombre de Reynolds, des éléments de maillage plus fins et des pas de temps plus petits sont nécessaires pour capturer ces dynamiques, ce qui conduit rapidement à des problèmes difficilement calculables. Pour réduire le coût de calcul, la turbulence peut être modélisée. On peut distinguer trois techniques de simulation :

- Simulation numérique directe (DNS). Les équations de Navier-Stokes sont directement résolues pour déterminer le champ de vitesse instantané à un instant donné $\mathbf{U}(\mathbf{x}, t)$. Comme toutes les échelles de longueur et de temps de l'écoulement sont résolues, le coût de calcul augmente avec Re^3. Les DNS nécessitent des résolutions de maillage et des échelles de temps très fines, ce qui restreint cette méthode aux écoulements à faible nombre de Reynolds.

- Simulation des grandes échelles (LES). Les équations N-S sont résolues pour un champ de vitesse filtré $\mathbf{U}^*(\mathbf{x}, t)$ représentatif des plus grands tourbillons de l'écoulement turbulent. Les différentes échelles de l'écoulement sont séparées en considérant un filtre spectral. Les structures à grande échelle sont directement résolues, tandis que l'influence des petites échelles sur les plus grandes est modélisée (modélisation sous-maille) ; ce modèle, qui doit être intégré dans l'équation de la quantité de mouvement, est généralement dérivé en supposant que les petites échelles filtrées présentent des caractéristiques universelles. Le LES est moins exigeant en termes de calcul que le DNS, même s'il représente naturellement des écoulements tridimensionnels non stationnaires.

- Simulation Navier-Stokes à la moyenne de Reynolds (RANS). En utilisant une décomposition de Reynolds aux équations de Navier Stokes, on peut séparer la valeur moyenne de l'écoulement des fluctuations et dériver les équations de transport à partir de l'équation de quantité de mouvement pour les champs moyens $\overline{\mathbf{U}}$. Cette équation doit cependant être fermée en utilisant un modèle de turbulence.

Pour résoudre numériquement la forme complète ou fermée des équations de Navier-Stokes, une méthode de discrétisation est utilisée pour approximer les équations différentielles partielles par un système d'équations algébriques. Cette méthode décrit le processus par lequel les équations continues de Navier-Stokes sont transférées en contreparties discrètes. Au lieu de fournir une solution numérique continue, les différentes techniques de la dynamique des fluides numérique (CFD) fournissent des résultats à des endroits discrets dans le temps et l'espace.

La méthode de discrétisation utilisée est de la plus haute importance pour la qualité et la précision de la solution numérique obtenue. Il existe différentes approches, mais les plus utilisées sont : la méthode des différences finies (FD), des volumes finis (FV), des éléments finis (FE) et des éléments spectraux (SE). Chacune de ces méthodes donne la même solution avec des grilles très fines ; cependant, certaines méthodes sont plus appropriées que d'autres (en termes de précision par rapport aux exigences de calcul), selon le type de problème que l'on essaie d'approcher numériquement.

Pour la simulation d'écoulements instationnaires, la méthode de discrétisation en temps et en espace génère un système d'équations de la forme suivante :

$$\Phi \mathbf{u}^{\mathrm{k}} = \Psi \mathbf{u}^{\mathrm{k\text{-}1}} + \mathbf{B}\mathbf{b}^{\mathrm{k\text{-}1}},$$

où $\mathbf{u}^{\mathrm{k}}$ représente le vecteur d'état inconnu à un instant t_{k}, $\mathbf{u}^{\mathrm{k\text{-}1}}$ représente le vecteur d'état connu à un instant $t_{\mathrm{k\text{-}1}}$, Φ et Ψ sont des matrices obtenues via le processus de discrétisation, et les conditions aux limites sont représentées par le vecteur $\mathbf{b}^{\mathrm{k\text{-}1}}$ et par $\mathbf{B}$, qui est le modèle appliqué sur $\mathbf{b}^{\mathrm{k\text{-}1}}$. Cette équation peut être résolue en utilisant des méthodes itératives. La vitesse de convergence de ces méthodes

est faible. Pour accélérer les processus de convergence, la stratégie multigrille peut être utilisée. La base de cette classe de méthodes provient d'une observation sur le taux de convergence des méthodes itératives. En général, les composantes à basse fréquence du spectre d'erreur sont les plus lentes à être amorties dans la procédure itérative, tandis que les fréquences plus élevées sont rapidement lissées. La méthode multigrille exploite cette propriété : en quelques itérations dans un maillage initial donné, la partie haute fréquence du spectre d'erreur sera considérablement amortie. Les basses fréquences restantes peuvent être transférées vers des grilles successives plus grossières, où cette erreur devient automatiquement une partie de la gamme des hautes fréquences, en raison du changement de résolution spatiale. Cette stratégie permet d'amortir progressivement l'erreur à basse fréquence dans le spectre d'erreur en utilisant une hiérarchie de grilles.

Dans cette thèse, les idées derrière la méthode multigrille sont utilisés pour développer un algorithme de d'assimilation de données séquentielle à bas coût et qui permette de respecter la conservativité des équations.

Assimilation de données : Un aperçu général et un focus sur les algorithmes séquentiels

L'objectif de cette thèse est de développer un algorithme séquentiel d'assimilation de données. Les méthodes d'assimilation de données permettent de combiner des mesures/observations expérimentales avec des réalisations de modèles (CFD) dans le but d'améliorer la prédiction d'état fournie par ces derniers. Plus précisément, l'assimilation de données fournit un cadre mathématique solide et élégant pour cette entreprise. Les observations fournissent des informations sur l'écoulement "réel". Cependant, elles sont généralement bruyantes et limitées dans le temps et l'espace. On pourrait penser à simplement interpoler/extrapoler les données, mais cela ne fonctionne pas puisque les phénomènes analysés dans le domaine de la mécanique des fluides sont fortement non linéaires. En revanche, avec la CFD, on peut obtenir des représentations cartographiques complètes et sans bruit des phénomènes d'écoulement.

De plus, les modèles peuvent être utilisés pour effectuer des prédictions sur les états futurs. Cependant, il est bien connu qu'en raison de l'incertitude épistémique inhérente (erreurs systématiques), les prédictions des modèles sont sujettes à la dérive. Il est également important de noter que même si le modèle ne contenait aucune erreur systématique, les conditions initiales et les conditions aux limites ne peuvent être connues avec une précision parfaite. Le comportement à long terme des écoulements est imprévisible en raison de la grande sensibilité des équations de Navier Stokes aux conditions initiales et aux conditions aux limites, en particulier pour les régimes de Reynolds élevés. Pour résoudre ce problème, les prédictions du modèle doivent être corrigées par la réalité et l'assimilation de données fournit les outils mathématiques nécessaires à l'ajustement optimal de la sortie du modèle avec les observations.

Le problème d'assimilation de données est formulé comme une procédure d'optimisation dont l'objectif est de trouver la correspondance optimale entre la sortie du modèle et l'observation. Cette procédure d'optimisation peut être réalisée en utilisant au moins deux stratégies différentes :

- Assimilation variationnelle de données. Ces procédures d'optimisation classiques impliquent la minimisation d'une fonction de coût positive et quadratique qui exprime l'inadéquation entre les mesures et le modèle. Ceci est généralement défini comme un problème de minimisation de l'erreur par les moindres carrés. La formulation du problème variationnel d'assimilation de données repose sur deux éléments clés : le calcul des variations et la méthode d'adjoint. En bref, les équations d'Euler-Lagrange aux dérivées partielles issues du champ mathématique du calcul des variations fournissent un cadre général pour la minimisation des fonctionnelles, et la méthode adjointe rend ce problème de minimisation calculable. Les deux méthodes variationnelles les plus connues sont la 3D-VAR et la 4D-VAR.

- Assimilation séquentielle de données. Lorsque l'assimilation de données est abordée d'un point de vue statistique, on est naturellement attiré vers le cadre bayésien qui permet de combiner rigoureusement les deux sources d'information

disponibles et de fournir une loi de probabilité postérieure. En pratique, l'application directe du cadre bayésien s'avère très couteux. Par conséquent, il faut introduire une hypothèse forte qui permet de simplifier le problème d'estimation de la fonction de densité de probabilité (PDF) à l'estimation de ses premiers moments statistiques : toutes les PDF qui interviennent dans le problème d'estimation sont considérées comme étant distribuées de manière Gaussienne. Dans ce contexte, le problème d'estimation est réduit à la détermination de la moyenne et de la variance de l'état postérieur. Cela peut être fait en utilisant la méthode d'estimation de variance minimale, qui conduit naturellement au célèbre filtre de Kalman (KF) et ses variantes. Notamment, le filtre de Kalman d'ensemble (EnKF) est plus adapté pour les problèmes de très grande dimension et la méthode séquentielle d'assimilation la plus utilisée.

Multigrid Ensemble Kalman Filter (MGEnKF)

Malgré l'avantage que les EnKF offrent pour l'analyse des systèmes dynamiques à grande échelle, l'utilisation d'un ensemble suffisamment grand (60 à 100 membres sont généralement nécessaires) peut encore être prohibitive pour les applications avancées. Nous avons développé une stratégie EnKF qui repose sur la génération des membres de l'ensemble sur un sous-espace (maille plus grossière) du modèle original. Pour ce faire, nous exploitons les multiples niveaux de résolution naturellement utilisés par la procédure multigrille pour l'avancement temporel de l'écoulement. Dans une configuration à deux grilles, les membres de l'ensemble calculés sur le niveau de maille grossier accompagnent une seule simulation à haut raffinement, qui est mise à jour à l'aide des résultats de l'assimilation de la maille grossière. Pour cette raison, les coûts de calcul et le stockage en mémoire des variables physiques sont considérablement réduits.

L'algorithme d'estimation développé combine le EnKF classique avec les caractéristiques de la stratégie multigrille. Il est donc appelé Multigrid Ensemble Kalman Filter (MGEnKF). Deux procédures distinctes sont exécutées sur la grille grossière de l'algorithme MGEnKF : une boucle extérieure, où les observations utilisées sont

obtenues à partir d'une source d'information externe, et une boucle intérieure, où les solutions de la grille fine projetées sur la grille grossière sont utilisées comme observations de substitution. L'algorithme général complet de MGEnKF est structuré selon les opérations suivantes :

- **Prédiction**. La solution initiale sur la grille fine $\left(\mathbf{x}_{k-1}^{\mathrm{F}}\right)^{\mathrm{a}}$ est utilisée pour calculer un état de prévision $\left(\mathbf{x}_{k}^{\mathrm{F}}\right)^{\mathrm{f}}$

$$\left(\mathbf{x}_{k}^{\mathrm{F}}\right)^{\mathrm{f}} = \mathcal{M}_{k:k-1}^{\mathrm{F}}\left(\left(\mathbf{x}_{k-1}^{\mathrm{F}}\right)^{\mathrm{a}}, \overline{\theta_{k}^{\mathrm{f}}}\right) \tag{1}$$

  où $\mathcal{M}_{k:k-1}^{\mathrm{F}}$ est le modèle utilisé sur la grille fine et $\overline{\theta_{k}^{\mathrm{f}}}$ est un ensemble de paramètres libres décrivant la configuration du modèle sur la grille fine. Chaque membre i de l'ensemble calculé sur la grille grossière est également avancé dans le temps

$$\left(\mathbf{x}_{k}^{\mathrm{C}}\right)^{\mathrm{f},(\mathrm{i})} = \mathcal{M}_{k:k-1}^{\mathrm{C}}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a},(\mathrm{i})}, \theta_{k}^{\mathrm{f},(\mathrm{i})}\right) + \mathcal{C}_{k:k-1}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a},(\mathrm{i})}, \psi_{k}^{\mathrm{f},(\mathrm{i})}\right), \tag{2}$$

  où $\mathcal{M}_{k:k-1}^{\mathrm{C}}$ est le modèle à grille grossière paramétré par $\theta_{k}^{\mathrm{f},(\mathrm{i})}$, tandis que $\mathcal{C}_{k:k-1}$ est un terme de correction supplémentaire inclus pour compenser la perte de précision due aux calculs effectués sur la grille grossière (similaire à la stratégie utilisée par Brajard et al., 2021). Ce modèle supplémentaire, dont la structure doit être déterminée, est piloté par l'ensemble des paramètres libres $\psi_{k}^{\mathrm{f},(\mathrm{i})}$.

- **Projection sur la grille grossière et boucle interne**. On projette $\left(\mathbf{x}_{k}^{\mathrm{F}}\right)^{\mathrm{f}}$ sur l'espace de la grille grossière via un opérateur de projection Π_{C}, de manière à obtenir $\left(\mathbf{x}_{k}^{\mathrm{C}}\right)^{*}$:

$$\left(\mathbf{x}_{k}^{\mathrm{C}}\right)^{*} = \Pi_{\mathrm{C}}\left(\left(\mathbf{x}_{k}^{\mathrm{F}}\right)^{\mathrm{f}}\right). \tag{3}$$

  Dans cette étape, l'observation de substitution, $\left(\mathbf{y}_{k}^{\mathrm{C}}\right)^{\mathrm{SO}}$, est extraite de $\left(\mathbf{x}_{k}^{\mathrm{C}}\right)^{*}$ avec un opérateur d'observation $\left(\mathcal{H}_{k}^{\mathrm{C}}\right)^{\mathrm{SO}}$: $\left(\mathbf{y}_{k}^{\mathrm{C}}\right)^{\mathrm{SO}} = \left(\mathcal{H}_{k}^{\mathrm{C}}\right)^{\mathrm{SO}}\left(\mathbf{x}_{k}^{\mathrm{C}}\right)^{*}$. L'opérateur d'observation, $\left(\mathcal{H}_{k}^{\mathrm{C}}\right)^{\mathrm{SO}}$, détermine la région de la solution obtenue sur la grille fine qui est observée. Ensuite, l'observation de substitution est utilisée dans la boucle interne. Ici, l'EnKF est utilisé comme une méthode d'estimation des

paramètres seulement, c'est-à-dire, les états d'ensemble, $(\mathbf{x}_k^{\mathrm{C}})^{\mathrm{f,(i)}}$ ne sont pas modifiés, mais les paramètres libres $\psi_k^{\mathrm{f,(i)}}$ sont mis à jour $\psi_k^{\mathrm{a,(i)}}$. Cette optimisation vise une amélioration de la prédiction des membres de l'ensemble simulé sur la grille grossière via une mise à jour du terme $\mathcal{C}_{k:k-1}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a,(i)}}, \psi_k^{\mathrm{f,(i)}}\right)$.

- **Boucle extérieure**. Si une observation externe $(\mathbf{y}_k^{\mathrm{C}})^{\mathrm{O}}$ est disponible, la prévision d'ensemble $(\mathbf{x}_k^{\mathrm{C}})^{\mathrm{f},(i)}$ est corrigée avec la méthode EnKF pour obtenir $(\mathbf{x}_k^{\mathrm{C}})^{\mathrm{a},(i)}$ ainsi qu'une mise à jour des paramètres $\theta_k^{\mathrm{a},(i)}$.

- **Détermination des variables d'état sur la grille grossière**. Dans cette étape, l'état physique de la simulation principale (sur la grille fine) est mis à jour sur la grille grossière. Cette solution, que l'on appellera $\left(\mathbf{x}_k^{\mathrm{C}}\right)'$, est obtenue en utilisant la matrice de gain de Kalman $(\mathbf{K}_k^{\mathrm{C}})^{x,\mathrm{e}}$ obtenue dans la boucle extérieure (EnKF) :

$$\left(\mathbf{x}_k^{\mathrm{C}}\right)' = (\mathbf{x}_k^{\mathrm{C}})^* + (\mathbf{K}_k^{\mathrm{C}})^{x,\mathrm{e}}\left[(\mathbf{y}_k^{\mathrm{C}})^{\mathrm{O}} - (\mathcal{H}_k^{\mathrm{C}})^{\mathrm{O}}\left((\mathbf{x}_k^{\mathrm{C}})^*\right)\right].$$

- **Itération finale sur la grille fine**. La solution d'état de la grille fine $\left(\mathbf{x}_k^{\mathrm{F}}\right)'$ est déterminée en utilisant les résultats obtenus sur l'espace grossier : $\left(\mathbf{x}_k^{\mathrm{F}}\right)' = (\mathbf{x}_k^{\mathrm{F}})^{\mathrm{f}} + \Pi_{\mathrm{F}}\left(\left(\mathbf{x}_k^{\mathrm{C}}\right)' - (\mathbf{x}_k^{\mathrm{C}})^*\right)$. L'état $(\mathbf{x}_k^{\mathrm{F}})^{\mathrm{a}}$ est obtenu par une procédure itérative finale partant de $\left(\mathbf{x}_k^{\mathrm{F}}\right)'$, suivant l'approche multigrille.

Validation de l'algorithme MGEnKF

Dans ce chapitre, le filtre multigrille MGEnKF est systématiquement évalué et étudié sur un panel d'exemples de complexité croissante. Les cas étudiés concernent une équation d'advection en 1D linéaire, une équation de Burger avec un terme de viscosité, un modèle en 1D, des équations d'Euler, des équations de Navier-Stokes en 2D compressibles pour une configuration de type couche de mélange.

Tous les cas de test réalisés pour évaluer les performances de l'estimateur sont des expériences jumelles d'assimilation de données dont la structure est très similaire :

- Un modèle d'écoulement est utilisé avec une condition d'entrée paramétrée spécifique pour générer une solution de référence. Une observation synthétique est générée à partir de cette référence en ajoutant du bruit artificiel.

- Dans l'expérience d'assimilation de données, l'écoulement est initialisé à un état qui n'est pas l'état réel. Nous supposons en outre que certains paramètres de la condition d'entrée sont inconnus. L'objectif de l'estimateur est alors double : assimiler les observations synthétiques pour corriger l'état de l'écoulement instationnaire et d'estimer les paramètres inconnus.

Advection 1D linéaire

Dans cette expérience, l'équation d'advection 1D est considérée. Une condition de Dirichlet variable dans le temps est imposée à l'entrée :

$$u(x = 0, t) = c\ (1 + \theta \sin(2\pi t))\,,$$

où θ représente l'amplitude d'une perturbation sinusoïdale de période $\Delta T = 1$ et est fixée à $\theta = 0.015$. Dans l'expérience numérique effectuée, θ est considérée inconnue. L'algorithme MGEnKF est ici utilisé pour estimer θ et pour améliorer l'estimation d'état de la variable u avec l'observation synthétique. La précision de l'algorithme est analysé avec et sans la boucle interne du MGEnKF.

Les résultats obtenus (Fig. 0-1 et Fig. 0-2) montrent une nette amélioration de la précision de l'estimation de θ quand la boucle interne est utilisée.

Équation de Burgers non-linéaire en 1D

De manière similaire à ce qui a été fait dans l'expérience précédente, la performance du MGEnKF est ici étudiée pour ce cas test. Une condition de Dirichlet variable dans le temps est imposée à l'entrée :

$$u(x = 0, t) = u_0\,(1 + \theta \sin(2\pi t))\,,$$

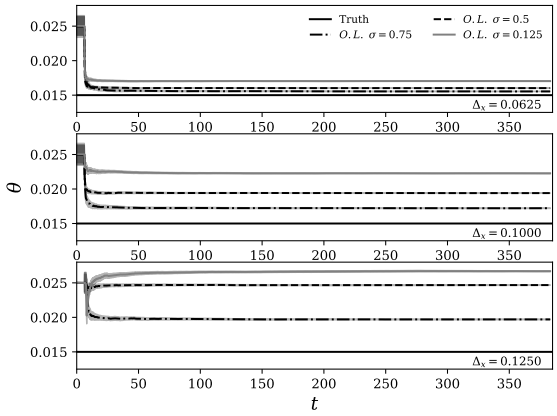


Figure 0-1: Estimation du paramètre θ sans la boucle interne.

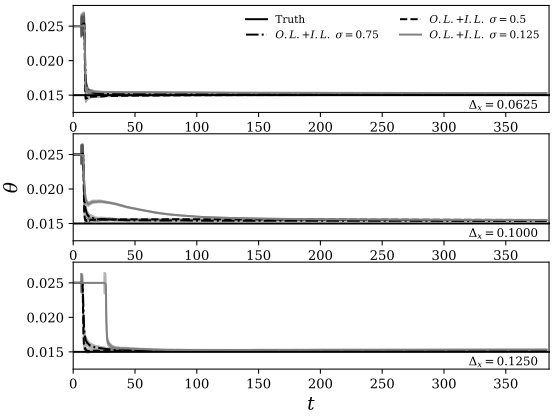


Figure 0-2: Estimation du paramètre θ avec la boucle interne.

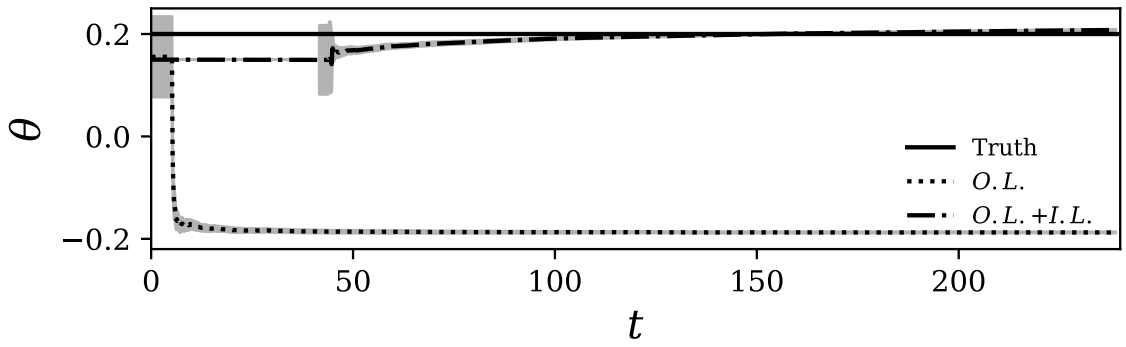


Figure 0-3: Estimation du paramètre θ avec et sans la boucle interne.

où u_0 est la vitesse caractéristique moyenne de l'écoulement et est fixée à $u_0 = 1$. Le paramètre θ représente l'amplitude d'un signal sinusoïdal dont la période est $\Delta T = 1$. Le paramètre d'amplitude a été fixé à $\theta = 0.2$. La précision de l'algorithme est analysé avec et sans la boucle interne du MGEnKF.

Les résultats présentés dans la Fig. 0-3 montrent que l'algorithme MGEnKF est beaucoup plus précis quand la boucle interne est activée. Comme dans le cas test précédent, un modèle de correction $\mathcal{C}$ a été conçu en analysant l'erreur de discrétisation liée à l'utilisation de maillages plus grossiers.

Équation d'Euler en 1D

Dans cette section, le modèle choisi pour l'expérience jumelle est celui des équations d'Euler en 1D. Contrairement aux équations de Burgers, dans les équations d'Euler, l'écoulement est considéré comme inviscide. Cependant, il existe des complexités supplémentaires liées aux effets du gradient de pression. Contrairement aux expériences jumelles réalisées dans les deux sections précédentes, dans cette section, nous essayons d'estimer un paramètre variant dans le temps lié à la condition limite d'entrée. Dans ce cas test, la boucle interne est désactivée. Le modèle d'Euler est beaucoup plus complexe que les deux modèles utilisés dans les expériences précédentes et la conception du modèle de correction, $\mathcal{C}$, est un vrai défi.

Une condition de vitesse de Dirichlet variant dans le temps est imposée à l'entrée :

$$u(x = 0, t) = u_0 \left(1 + \theta(t) \sin(2\pi f_c t)\right) \tag{4}$$

L'amplitude de l'onde sinusoïdale est déterminée par un paramètre variable dans le temps $\theta(t) = \theta_0 \left(1 + \sin\left(2\pi \frac{f_c}{10} t\right)\right)$, où θ_0 est une constante. Dans l'expérience jumelle d'assimilation de données, l'algorithme MGEnKF est utilisé pour estimer la valeur de $\theta(t)$ et pour corriger la variable d'état u avec l'observation synthétique.

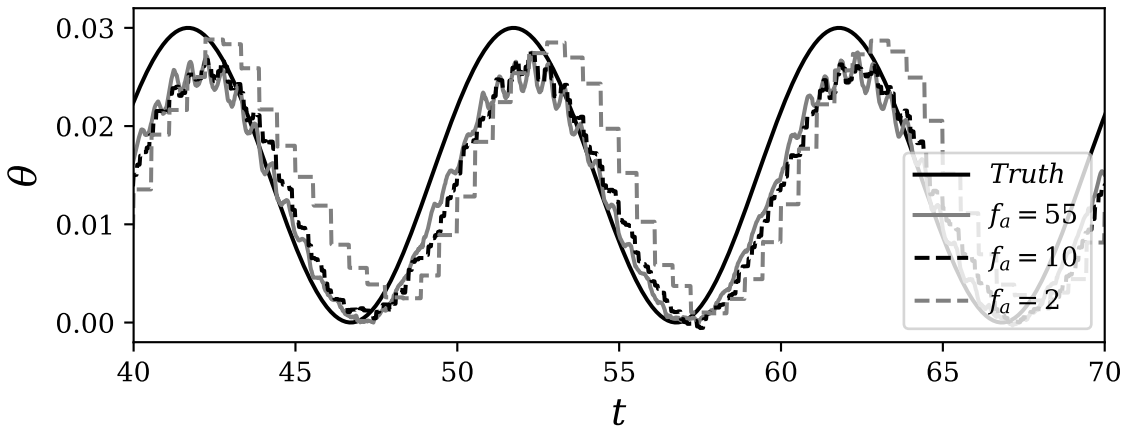


Figure 0-4: Estimation du parametre $\theta(t)$.

Concernant l'estimation du paramètre $\theta(t)$, les résultats obtenus (Fig. 0-4) mon-

trent que l'algorithme MGEnKF est capable d'estimer de manière précise un paramètre qui varie dans le temps.

Couche de mélange compressible en 2D

La couche de mélange bidimensionnelle évoluant dans l'espace à Re = 100 est étudiée ici en utilisant les équations compressibles de Navier Stokes. Comme dans le cas test précédent, la boucle internet est désactivée. Une condition de vitesse de Dirichlet variant dans le temps est imposée à l'entrée :

$$U_{\text{in}}(y,t) = \frac{U_1+U_2}{2} + \frac{U_1-U_2}{2}\tanh\left(\frac{2y}{\delta_0}\right) + U_{\text{Pert}}(y,t), \qquad -3 < y < 3$$

où

$$U_{\text{Pert}}(y,t) = \sum_{i=1}^{N_{\text{in}}} \epsilon_i(t)\frac{U_1+U_2}{2}[f_i(y)\sin(\omega_i t)].$$

Dans ce cas test, $N = 1$. Le paramètre ϵ_1 varie dans le temps selon une forme sinusoïdale : $\epsilon_1(t) = \epsilon(1+\sin(\omega_\epsilon t))$. Les valeurs des paramètres numériques caractérisant la perturbation sont $\epsilon = 0.15$ et $\omega_\epsilon = 0.62\omega_1$. Dans l'expérience jumelle d'assimilation de données, l'algorithme MGEnKF est utilisé pour estimer la valeur de $\epsilon_1(t)$ et pour corriger la prédiction d'état avec l'observation synthétique.

Les résultats obtenus pour l'estimation de $\epsilon_1(t)$ (Fig. 0-5) montrent que la tendance générale sinusoïdale est généralement respectée, bien qu'un décalage de phase relativement faible soit visible. Ces résultats pourraient être sensiblement améliorés avec l'implémentation de la boucle interne. Concernant l'estimation d'état effectuée, l'évolution de l'erreur quadratique moyennée (RMSE) de la variable ρv montre que l'algorithme réduit l'erreur totale du système en estimant $\epsilon_1(t)$ mais aussi en effectuant des corrections d'état sur la grille fine (Fig. 0-6).

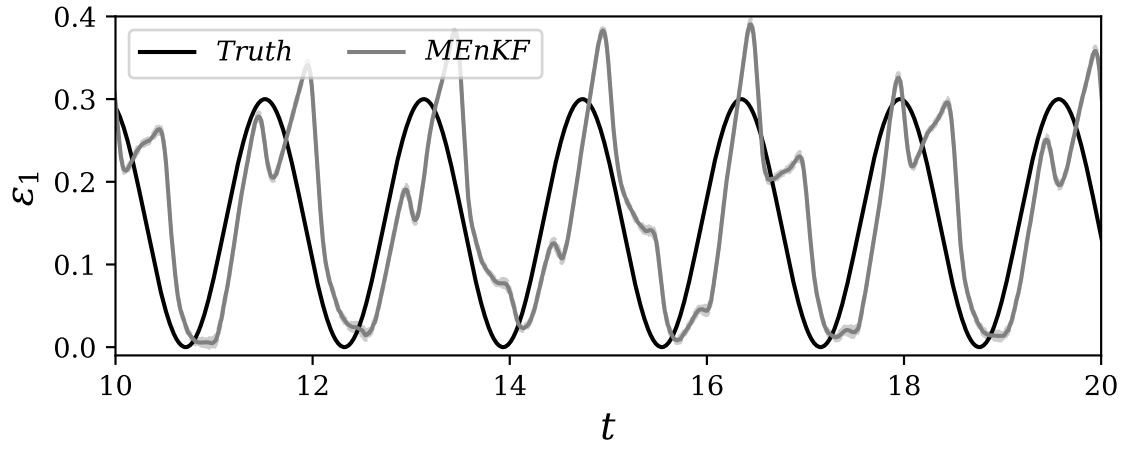


Figure 0-5: Estimation de $\epsilon_1(t)$.

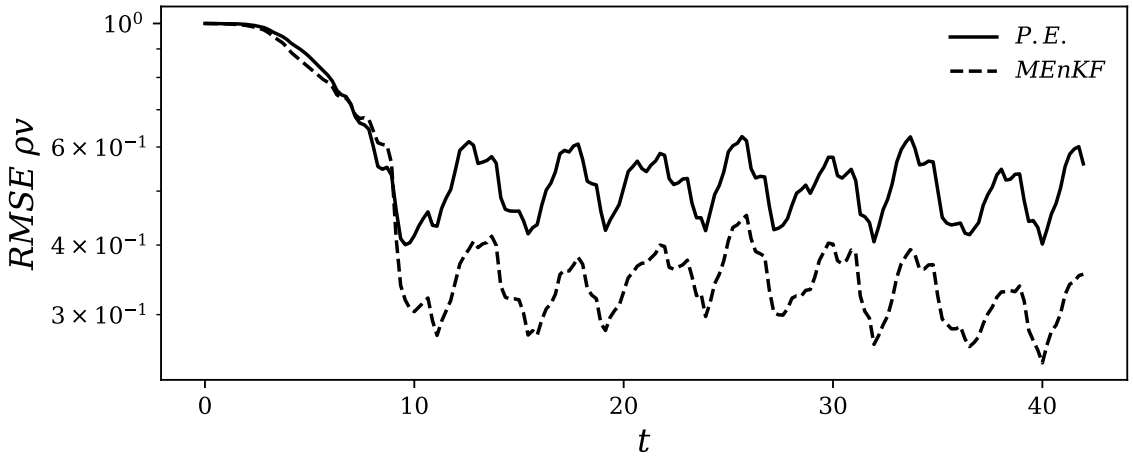


Figure 0-6: Erreur quadratique moyennée (RMSE) de la variable ρv. Comparaison être un test effectué avec estimation de paramètre seulement (P.E) et estimation de paramètre et d'état (MGEnKF).

Application de l'algorithme MGEnKF à un problème complexe : BARC

Dans ce chapitre, l'algorithme MGEnKF est testé sur une LES 3D d'un écoulement incompressible autour d'un cylindre rectangulaire à nombre Reynolds 40.000. Le domaine de calcul est présenté dans la Fig. 0-7.

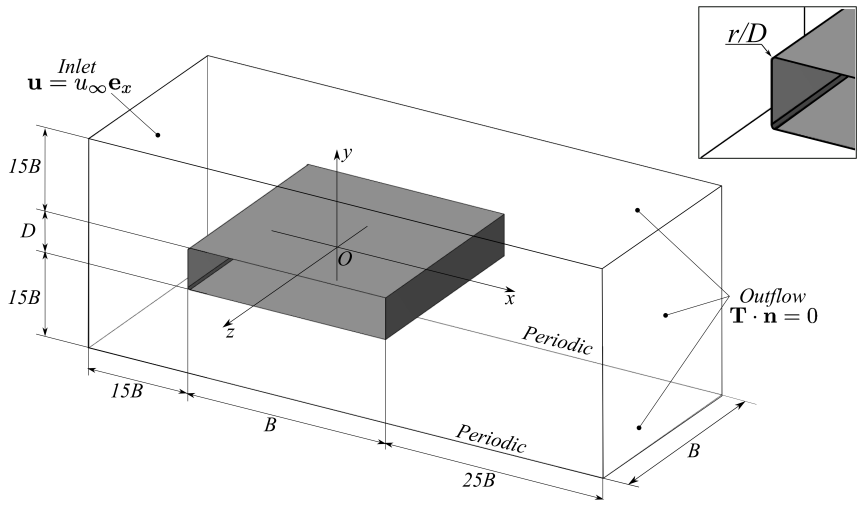


Figure 0-7: Domaine de calcul

Les données d'observation sont obtenues d'une simulation de référence effectuée sur un maillage très fin. Une zone d'observation 3D réduite au-dessus du cylindre a été sélectionnée avec des observations 3D de vitesse volumique. L'objectif est d'augmenter la capacité prédictive d'une LES à une résolution intermédiaire (réduction d'un facteur 2 en x, y, z) avec des observations synthétiques. Une seule simulation à résolution intermédiaire est lancée en parallèle avec une simulation d'ensemble sur une grille grossière (réduction d'un facteur 4 en x, y, z par rapport à la simulation de référence). En se basant sur la méthode MGEnKF, l'assimilation est effectuée vis-à-vis d'un paramètre de contrôle attaché au filtrage fréquentiel régissant la LES et en effectuant des procédures d'estimation d'état. Ce filtre est caractérisé par une fonction de transfert quadratique, qui agit du mode $N - k_c$ ($k_c = 3$ ici) jusqu'au mode le plus élevé N. La fonction de transfert s'écrit comme suit :

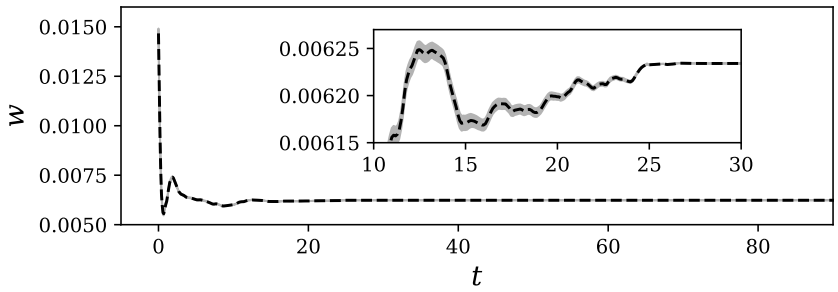


Figure 0-8: Estimation du parametre w.

$$\begin{cases} \sigma_k = 1 & k < k_c \\ \sigma_k = 1 - w\left(\frac{k-k_c}{N-k_c}\right)^2 & k_c \leq k \leq N, \end{cases}$$

et, une fois que le mode de coupure (k_c) est fixé, le filtrage ne peut être réglé que par un paramètre de pondération w. Ce paramètre est optimisé avec l'algorithme MGEnKF.

Les résultats numériques permettent de valider sur cet exemple difficile la procédure d'assimilation multigrille proposée dans cette thèse. Les résultats montrent que le paramètre w converge vers une valeur environ trois fois plus petite que la valeur de base (Fig. 0-8). Les résultats indiquent que la valeur antérieure de w, lorsqu'elle est utilisée en combinaison avec la grille intermédiaire, produit une dissipation excessive. On pourrait dire que le processus d'optimisation dans le MGEnKF réduit la dissipation interne de la LES, en réduisant la valeur de w, afin de compenser l'erreur numérique associée à l'utilisation d'un maillage moins raffiné. Une amélioration de la prédiction de la variance de la pression et de la pression moyenne à la surface supérieure du cylindre est observée (Fig. 0-9).

La taille de la région de recirculation est également améliorée et plus proche de la simulation de référence en comparaison avec des simulations de résolution équivalente et calibrées a priori (Fig. 0-10).

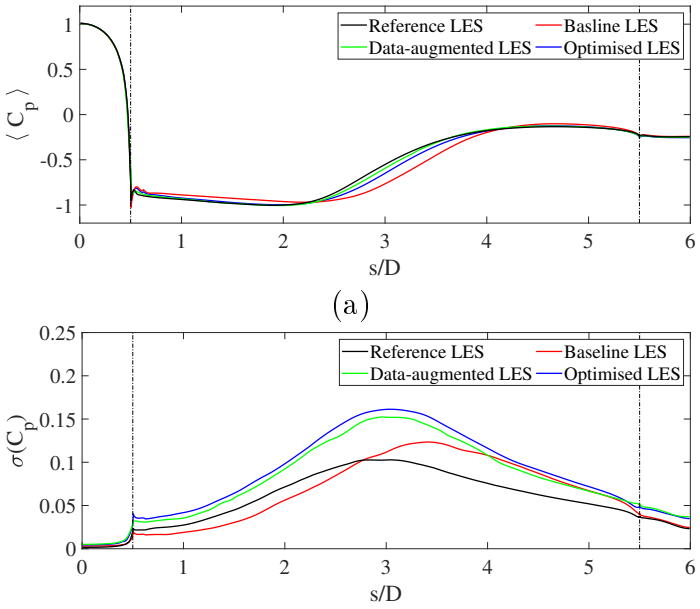


Figure 0-9: (a) Coefficient de pression moyenné dans le temps, sur surface du cylindre $\langle C_p \rangle$, et (b) son écart-type, $\sigma(C_p)$. Comparaison entre les résultats obtenus pour la LES de référence, la LES de base, et la LES améliorée avec l'algorithme MGEnKF.

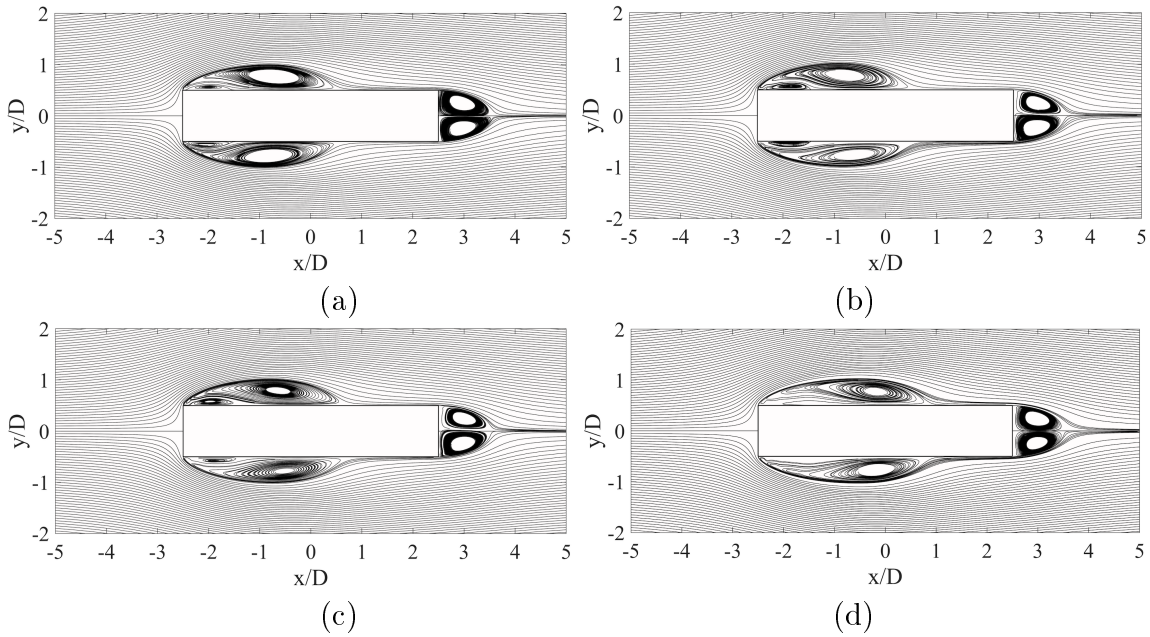


Figure 0-10: Lignes de courant de l'écoulement moyenné dans le temps pour (a) la LES de référence, (b) la LES améliorée avec MGEnKF, (c) la LES optimisée et (d) la LES de base.

Conclusions

La prédiction précise des écoulements turbulents reste un défi ouvert pour la recherche en mécanique des fluides. Les barrières technologiques concernant la puissance de calcul limitent l'utilisation d'outils de simulation numérique directe aux cas de faible Reynolds, et une révolution dans la modélisation de la turbulence n'est pas attendue dans un avenir proche.

Malgré ses limites, la CFD a été largement adoptée par les constructeurs d'avions et d'automobiles dans leurs procédures de conception ; cependant, les conditions hors conception, nécessitent des outils plus sophistiqués.

Les outils issus de la théorie de l'estimation commencent à être utilisés par les chercheurs en CFD dans le but d'augmenter la précision de prédiction des codes déjà existants. L'idée derrière cette classe de techniques est de combiner des données numériques avec des données expérimentales pour améliorer l'estimation d'état. Cela relève d'un domaine mathématique plus large, qui a été développé depuis des décennies et est largement utilisé dans les prévisions météorologiques : L'assimilation des données. Les techniques d'assimilation de données séquentielles, dont l'algorithme du filtre de Kalman est l'outil le plus connu, sont particulièrement adaptées aux écoulements instationnaires ; cependant, la résolution d'un problème d'assimilation de données avec une approche séquentielle s'avère très chère en termes de coût de calcul.

L'objectif de cette étude de doctorat était de développer un estimateur séquentiel qui pourrait être intégré dans des codes CFD largement utilisés. Nous avons développé un algorithme d'estimation combinant le Filtre de Kalman d'Ensemble et la stratégie multigrille pour assimiler les écoulements instationnaires.

L'algorithme dit MGEnKF (Multi-Grid Ensemble Kalman Filter) exploite les états physiques obtenus sur plusieurs grilles de résolutions différentes pour effectuer l'estimation d'état et l'optimisation paramétrique en utilisant des procédures EnKF. Plus précisément, lorsque seulement deux grilles sont utilisées, un ensemble de simulations basse-fidélité de l'écoulement est exécuté sur un niveau de grille grossier ainsi

qu'une seule simulation sur le niveau de maille le plus fin. L'estimation d'état obtenue au niveau grossier et les statistiques d'ensemble associées sont utilisées pour filtrer la solution à maille fine et optimiser un ensemble de paramètres décrivant le modèle (conditions aux limites, paramètres du modèle, etc). Cette procédure permet i) de réduire les coûts de calcul de l'EnKF et ii) d'assurer la conservativité de la solution finale.

D'une manière générale, cette méthode s'inscrit dans la classe des techniques *multiniveau* qui visent à améliorer l'estimation des statistiques de simulations numériques coûteuses en considérant différents niveaux de résolution.

L'évaluation de la méthode a été réalisée par l'analyse de cas de test unidimensionnels, bidimensionnels et tridimensionnels, en utilisant différents modèles de complexité croissante. Les résultats montrent que le MGEnKF peut mettre à jour avec succès l'état d'un système avec les observations disponibles pour augmenter la précision globale de la prédiction d'état fournie par le modèle. En outre, la description paramétrique du problème numérique (en termes de conditions aux limites prescrites, de fermetures de turbulence, etc) peut être optimisée de manière adéquate en tenant compte des différentes résolutions de maillage employées dans l'algorithme.

Le MGEnKF ouvre des perspectives intéressantes pour une application potentielle au contrôle actif des écoulements et au design industriel.

Chapter 1

Introduction

Context

Global economies and societies are preparing for an ecological transition that aims to revolutionize energy consumption. A change is needed in the energy paradigm by shifting from fossil-fuel-based economies to clean and renewable models. The pace at which this ongoing revolution will unfold will largely depend on the geopolitical context and the technological advancements that might help hasten this transition. Most experts agree that this process will take decades to unfold, and in this transitory period energy efficiency is more important than ever.

Road transport, civil aviation and navigation represent almost a third of the total energy consumption and greenhouse gas emissions in Europe (Eurostat, 2022). The regulations fixed by the European Union concerning CO2 emissions on newly manufactured vehicles are more stringent every year and the transport industry is compelled to continue the research and development efforts that have already been made in designing energy efficient solutions for mobility. The research efforts made in shape optimization over the last decades have been substantial, and aerodynamicists are reaching the limit to the lowest achievable drag.

The main challenge that aerodynamicists face when tackling the design optimization of vehicles is the high variability of flow regimes in operative conditions. However, technological advancement in flow control techniques and the advent of data-driven

methods could revolutionize the sector.

Inherent unsteady nature of flows

Transport vehicles encounter off-design conditions that are intrinsically unsteady. Dynamic phenomena such as flow separation and laminar-turbulent transition pose a real scientific challenge (see Rizzi and Luckring, 2021and Olson et al., 2013 for flow separation and Fedorov, 2011 for laminar-turbulent transition). They have a crucial influence on the loads, both steady and unsteady, acting on the vehicles. Taking ground vehicles as an example, massive separation happens at the trailing edges which forms the near wake. This gives birth to various instabilities evolving with very different length and time scales, such as the recently discovered bi-modal dynamics (Grandemange et al., 2013), which dominates the near wake. These instabilities, in a time-averaged context, set the pressure drag acting on the rear surface of the vehicles (Balachandarb et al., 1997).

To account for off-design conditions in the aerodynamic development, car manufacturers are showing increasing interest in active flow control techniques, which have the potential of reducing the aerodynamic drag by up to 25% (Brunton & Noack, 2015b; Choi et al., 2008; Choi et al., 2014). Successful open-loop or feed-back control systems tackling the near wake have been reported in the literature (Barros et al., 2016; Haffner et al., 2020; Li et al., 2016), with impressive accounts of drag decrease.

Limits of the traditional tools

In a flow control problem, having accurate information about the state of the system is essential for designing successful control strategies. Surface sensors can provide valuable information about the state of the flow at a given instant, which can be fed to a closed-loop control system. However, this information is usually insufficient for the design of robust control strategies. Complete maps of flow behaviour as the input to a control system is the ideal objective, but experimental techniques alone

cannot provide this type of data from local measurements; the noise associated to observations interacts in a spurious manner with the highly non-linear behaviour of the (normally) turbulent flows that have to be controlled, hindering the capabilities of existing flow estimation algorithms.

In the case of aerodynamic design, wind-tunnel experiments have been the classical aerodynamic development tool. However, since the 1980s, Computational Fluid Dynamics (CFD) has become essential in the design and analysis of transport vehicles. Numerical analysis tools tackling average properties of the flow have proven to be efficient for design optimization purposes. In nowadays engineering problems, Reynolds Averaged Navier Stokes (RANS) solvers remain the predominant tool used for the analysis and prediction of turbulent flow simulation, despite their well known limitations (Wilcox, 1994). The main issue with turbulence modelling is that there is no model that applies to all kinds of flows. Instead, there are several turbulence models, each one tailored for a specific application.

In a near future, CFD could be a valuable tool for active flow control, since it would allow obtaining more complete information about the state of the flow at a given instant. Furthermore, there is a vast potential for CFD to be more widely integrated in the aerodynamic design process of vehicles. For this to happen, most experts agree that flows presenting strong dynamic phenomena require resolution methods where the unsteady features are tackled explicitly. RANS or URANS (unsteady RANS) methods are not accurate in these cases and new resolution methods have to be envisaged. In the aerospace industry, for example, certain areas such as the high-lift wing design and wing control, where strong unsteady features are present, do not incorporate CFD tools in their aerodynamic development due to lack of accuracy in the traditional RANS approach (Spalart & Venkatakrishnan, 2016).

In the last decades, there has been an ongoing shift from RANS turbulence models to Large Eddy Simulation (LES). Over the years and with the experience acquired from RANS-type simulation, researchers have recognized that Direct Numerical Simulation (DNS)/LES methods are required to accurately reproduce complex flows. DNS is, however, computationally untreatable for flows with moderate to high Reynolds

numbers. LES is less expensive than DNS, but the cost associated with wall resolution (which may result in LES computations as expensive as DNS ones) does not allow for applications to very complex cases, unless wall modelling is used. Sub-grid scale modelling is also necessary for the scales that are smaller than the grid size. Wall and sub-grid scale modelling introduce, however, systematic errors in the simulations.

It appears that research in pure CFD methods has hit a ceiling. There has not been a revolutionary change in turbulence modelling over the past years and this trend does not seem to change in the near future (Spalart & Venkatakrishnan, 2016). Moreover, Moore's law shows signs of slowing down since beginning of the 2010s. This will delay the application of pure LES/DNS simulation strategies for high-Reynolds number turbulent flows of industrial interest. The rather recent introduction of hybrid RANS/LES approaches has enabled to consider more complex geometries at higher Reynolds numbers. It essentially consists in using RANS models in attached boundary layers while switching in LES-like mode resolution in separated regions. However, such approaches still require very fine-tuning of both numerical parameters and grid resolution, quite systematically case dependant, to avoid numerical artefacts such as transition delay, grid induced separation, etc.

In conclusion, to continue the increase in energy efficiency of transport vehicles, innovative active flow control techniques as well as improved aerodynamic design procedures are required. For this, new prediction tools that are able to accurately characterize complex flows have to be developed.

New tools coming from Estimation Theory

CFD is affected by uncertainty, and more specifically, by systematic errors coming from simplifications in geometry, modelling hypothesis and numerical discretization errors. Moreover, boundary conditions can hardly reproduce real conditions since these cannot be known with perfect precision. Flow measurements provide information of the *real* flow, but are usually noisy and limited in time and space. Both CFD and experimental data are thus affected by bias and uncertainty, which hinders their

capability to accurately predict/replicate complex flows.

In this context, Data Assimilation can be used to *combine* both sources of information and provide a state estimate that minimizes the uncertainty in conventional tools. In an ideal flow control situation, CFD and measurement could be used in a synergistic way: the complete but systematically biased flow maps provided by CFD could be corrected *in streaming* by external noisy observation as it becomes available. Data Assimilation can also improve the analysis of flows around complex geometries. Simulations could integrate experimental data in order to correct the systematic errors present in the numerical models used. Meteorological researchers specializing in Data Assimilation techniques have perfected this type of procedures for decades. Notably, sequential Data Assimilation tools coming from Estimation Theory (ET) such as the Kalman Filter (KF) are specifically tailored for progressively updating a model by integrating data as soon as it becomes available. As we will later see in this manuscript, the numerical methods used to simulate unsteady flow are highly compatible with this type of procedure.

This PhD study tackles the sequential *state & parameter estimation* problem in fluid mechanics. In this field, there are particular challenges that one has to face when applying sequential DA algorithms to combine model and observation data:

- High dimensional problems ($\mathcal{O}(10^9)$) which results in untreatable algorithms
- Highly non-linear models (Navier-Stokes)
- High sensitivity to brute model corrections that break *conservativity* in the model.

The objective of this PhD is to develop a sequential Data Assimilation algorithm that can be used for unsteady flows and with an acceptable computational cost. The Ensemble Kalman Filter (EnKF) is used as foundation for the estimation algorithm we want to develop. The two main assets of the EnKF is that it allows to treat fully non-linear problems, and its simple conceptual formulation ensures a relative ease of implementation to already existing codes. The EnKF is thus a perfect candidate for

assimilation of unsteady flows. There remain however two significant obstacles to the implementation of EnKF techniques to unsteady problems: computational cost and conservativity. We attempt to overcome these two issues with the aid of the multigrid method.

Thesis Outline

In this research work, we present a newly developed sequential Data Assimilation algorithm, combining multi-grid aspects and the ensemble Kalman Filter. The so-called Multi-Grid Ensemble Kalman Filter (MGEnKF) algorithm exploits physical states obtained on multiple grids of different resolution to perform state estimation and parametric optimization using EnKF procedures. More precisely, an ensemble of low-fidelity simulations of the flow is run on a coarse grid level together with a single, high-fidelity simulation run on the finest mesh level. The state estimation obtained at the coarse level and the associated ensemble statistics are used to filter the finest mesh solution and to optimize a set of parameters describing the model (boundary conditions, model parameters, ...). This procedure allows to i) reduce the computational costs of the EnKF and ii) ensure the conservativity and the smoothness of the final solution.

We start by introducing the classical numerical models used for the simulation of unsteady flows, and we treat Data Assimilation techniques. The method is thereafter presented together with some validation test cases. The manuscript is thus organized as follows:

- In Chapter 2, we present the numerical models used in fluid mechanics, and we provide a short introduction to the numerical techniques used to numerically approximate unsteady flow problems, with a special emphasis on multigrid methods.

- In Chapter 3, Data Assimilation methodologies are treated. The two families of Data Assimilation techniques, i.e., sequential and variational methods are

presented, with a literature review of the relatively recent usage of these class of methods in fluid mechanics research.

- In Chapter 4 the Multigrid Ensemble Kalman Filter is presented in detail.
- In Chapter 5, the MGEnKF is applied to the analysis of different test cases. Several dynamical systems of increasing complexity have been chosen in order to highlight different properties of the algorithm.
- In Chapter 6, the MGEnKF is used to improve the prediction accuracy of a intermediate resolution Large Eddy Simulation with synthetic observation data. The experiment is carried out on the flow around a 5:1 rectangular cylinder at high Reynolds number ($Re = 40000$).
- In Chapter 7, the main findings of this PhD study are summarized with recommendations for future research.

Chapter 2

Numerical Models in Fluid Mechanics

2.1 Continuum Mechanics: Governing Equations

This PhD study deals with problems of fluid mechanics. The equations that govern these type of problems derive from the conservation principles of mass, momentum and total energy. Throughout this chapter, the notation introduced by Jasak, 1996 for describing the governing equations is used. Let us recall the rate of change in time of an intensive property of the system, θ, through the material derivative

$$\frac{d}{dt}\int_{\mathcal{V}(t)} \rho\theta\left(\mathbf{x},t\right) d\mathcal{V} = \frac{\partial}{\partial t}\int_{\mathcal{V}(t)} \rho\theta d\mathcal{V} + \oint_{\partial\mathcal{V}(t)} d\mathbf{S}.\left(\rho\theta\mathbf{U}\right), \tag{2.1}$$

where ρ is the density, $\mathcal{V}\left(t\right)$ is the time-dependent control volume, $\partial\mathcal{V}\left(t\right)$ is the surface of the control volume, $\mathbf{U}$ represents the velocity vector, and $d\mathbf{S}$ is the unit of surface times the vector pointing in the outward direction of $\partial\mathcal{V}\left(t\right)$. The rate of change of θ within the control volume, $\mathcal{V}$, is thus equal to the sum of the constraints applied on the surface, $\mathbf{F}_S$, and the volume sources $S_{\mathcal{V}}$.

$$\frac{\partial}{\partial t}\int_{\mathcal{V}(t)} \rho\theta d\mathcal{V} + \oint_{\partial\mathcal{V}(t)} d\mathbf{S}.\left(\rho\theta\mathbf{U}\right) = \int_{\mathcal{V}(t)} S_{\mathcal{V}} d\mathcal{V} + \oint_{\partial\mathcal{V}(t)} d\mathbf{S}.\mathbf{F}_S\left(\theta\right) \tag{2.2}$$

Equation (2.2) can also be written in the differential form

$$\frac{\partial(\rho\theta)}{\partial t} + \boldsymbol{\nabla}.(\rho\theta\mathbf{U}) = S_{\mathcal{V}}(\theta) + \boldsymbol{\nabla}.\mathbf{F}_S(\theta). \tag{2.3}$$

Starting from the general form provided in (2.3), the conservation principles of continuum mechanics can be written in the differential form (Aris, 1990):

1. Mass conservation

$$\frac{\partial\rho}{\partial t} + \boldsymbol{\nabla}.(\rho\mathbf{U}) = 0, \tag{2.4}$$

2. Momentum conservation

$$\frac{\partial(\rho\mathbf{U})}{\partial t} + \boldsymbol{\nabla}.(\rho\mathbf{U}\mathbf{U}) = \rho\mathbf{g} + \boldsymbol{\nabla}.\sigma, \tag{2.5}$$

3. Energy conservation

$$\frac{\partial(\rho E)}{\partial t} + \boldsymbol{\nabla}.(\rho E\mathbf{U}) = \rho\mathbf{g}.\mathbf{U} + \boldsymbol{\nabla}.(\sigma.\mathbf{U}) - \boldsymbol{\nabla}.\mathbf{q} + \rho S, \tag{2.6}$$

where σ is the stress tensor, E is the total energy per unit volume, S is the volume energy source, $\mathbf{g}$ represents a general volume source term and is usually considered to model gravity effects, and $\mathbf{q}$ is the heat flux. The mass conservation equation (2.4) does not include source terms, since mass cannot be created nor destroyed. The total energy of the system, E, is calculated as the sum of the kinetic energy, E_K, and the internal energy e. In fluid mechanics studies, chemical, nuclear and other forms of energy are considered only on configurations that explicitly need them. Therefore, E can be written as

$$E = E_K + e(P,T) = \frac{1}{2}\mathbf{U}.\mathbf{U} + e(P,T), \tag{2.7}$$

where the internal energy, e, is defined as a function of temperature, T, and pressure P

$$e = e(P,T). \tag{2.8}$$

These equations are valid for any system modelled as a continuous medium. However, the number of unknown variables is larger than the number of equations in the system. In order to "close" the system, it is necessary to define additional constitutive relations between the unknown physical quantities. These relations are specific to the considered continuum system. For Newtonian fluids, the constitutive relations used are:

1. Equation of state. For ideal gases:

$$\rho = PrT, \tag{2.9}$$

   where r is the ideal gas constant.

2. Fourier's law of heat conduction

$$\mathbf{q} = -\lambda \boldsymbol{\nabla} T, \tag{2.10}$$

   where λ is the heat conductivity.

3. Newton's law for describing stress tensor components

$$\sigma = -\left(P + \frac{2}{3}\mu \boldsymbol{\nabla}.\mathbf{U}\right)\mathbf{I} + \mu\left[\boldsymbol{\nabla}\mathbf{U} + (\boldsymbol{\nabla}\mathbf{U})^T\right], \tag{2.11}$$

   where μ is the dynamic viscosity.

4. Sutherland's Law for expressing the dynamic viscosity, μ as a function of T

$$\mu = \mu_{\text{ref}} + \left(\frac{T}{T_{\text{ref}}}\right)^{\frac{3}{2}} \frac{T_{\text{ref}} + \mathcal{S}}{T + \mathcal{S}} \tag{2.12}$$

   where the variables with the subscript ref indicate reference values and $\mathcal{S}$ is the Sutherland temperature.

5. Constitutive relation for heat conductivity λ. The dynamic viscosity of a fluid, μ is correlated with its heat conductivity, λ through the Prandtl number, $Pr =$

$\frac{c_p \mu}{\lambda}$, where c_p is the specific heat of the fluid. A constitutive relation for λ can be obtained by considering a constant Prandtl number.

The conservation laws in continuum mediums together with these constitutive laws produce a closed system of partial differential equations:

1. Continuity equation

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}.\left(\rho \mathbf{U}\right) = 0, \tag{2.13}$$

2. Navier-Stokes equation (also Momentum equation)

$$\begin{aligned}\frac{\partial \left(\rho \mathbf{U}\right)}{\partial t} + \boldsymbol{\nabla}.\left(\rho \mathbf{U}\mathbf{U}\right) = \rho \mathbf{g} - \boldsymbol{\nabla}\left(P + \frac{2}{3}\mu \boldsymbol{\nabla}.\mathbf{U}\right) \\ + \boldsymbol{\nabla}.\left[\mu\left(\boldsymbol{\nabla}\mathbf{U} + \left(\boldsymbol{\nabla}\mathbf{U}\right)^T\right)\right],\end{aligned} \tag{2.14}$$

3. Energy equation

$$\begin{aligned}\frac{\partial \left(\rho E\right)}{\partial t} + \boldsymbol{\nabla}.\left(\rho E \mathbf{U}\right) = \rho \mathbf{g}.\mathbf{U} - \boldsymbol{\nabla}.\left(P\mathbf{U}\right) - \boldsymbol{\nabla}.\left(\frac{2}{3}\mu\left(\boldsymbol{\nabla}.\mathbf{U}\right)\mathbf{U}\right) \\ + \boldsymbol{\nabla}.\left[\mu\left(\boldsymbol{\nabla}\mathbf{U} + \left(\boldsymbol{\nabla}\mathbf{U}\right)^T\right).\mathbf{U}\right] + \boldsymbol{\nabla}.\left(\lambda \nabla T\right) + \rho S,\end{aligned} \tag{2.15}$$

Equations (2.13), (2.14) and (2.15) constitute the so-called *compressible* Navier-Stokes equations. For many flows of engineering interest, this set of equations can be further simplified assuming incompressible and isothermal flow, i.e ρ = const., $\lambda = \infty$. The so-called *incompressible* Navier-Stokes system of equations can thus be written in the following form:

$$\boldsymbol{\nabla}.\mathbf{U} = 0 \tag{2.16}$$

$$\frac{\partial \mathbf{U}}{\partial t} + \boldsymbol{\nabla}.\left(\mathbf{U}\mathbf{U}\right) = \mathbf{g} - \boldsymbol{\nabla} p + \boldsymbol{\nabla}.\left(\nu \boldsymbol{\nabla}\mathbf{U}\right), \tag{2.17}$$

where $p = P/\rho$ is the kinematic pressure, $\mathbf{g}$ the acceleration of gravity and $\nu = \mu/\rho$ the kinematic viscosity.

2.2 Turbulence Modelling

The discussion in the previous section indicates that every flow configuration can be studied numerically once the discretization method of the partial differential equations, the constitutive relations and the boundary/initial conditions are set. However, for high Reynolds numbers regimes, the accurate representation of the physical phenomena must take into account complex multiscale interactions due to the strong non-linear behaviour of the flow. In addition, with increasing Reynolds numbers, finer mesh elements and smaller time steps are needed to capture these dynamics, which rapidly leads to computationally intractable problems.

In terms of energy, the largest eddies present in the turbulent flow contain most of the kinetic energy. This energy is transferred through a non-linear process to the lower scale eddies, down to the Kolmogorov scale, where viscous dissipation is predominant. The energy spectrum of turbulence, $E(k)$, represents the contribution of turbulent kinetic energy by wavenumber k . It follows that $E(k)$ is related to the mean turbulent kinetic energy in the flow as:

$$\int_0^\infty E(k)\,\mathrm{d}k = \frac{1}{2}\sum_{i=1}^{3}\overline{U_i'U_i'}, \tag{2.18}$$

where U_i' represents the i^{th} fluctuating velocity component (3 different components in 3D). The overbar on the variable represents the ensemble average.

The simulation of turbulent flows can be approached in three different forms, depending on the wavenumber range that is either calculated or modelled with the set of equations (see Fig. 2-1):

- Direct numerical simulation (DNS) (Eswaran & Pope, 1988, 1988; Rogallo & Moin, 1984; Rogers & Moin, 1987). The N-S equations ((2.16) and (2.17) or (2.13), (2.14) and (2.15)) are directly solved to determine the instantaneous velocity field at a given instant $\mathbf{U}(\mathbf{x}, t)$. Since all the lengthscales and timescales of the flow are resolved, the computational cost increases as Re^3 (Pope, 2000). DNS require very fine mesh resolutions and time-step sizes, restricting this

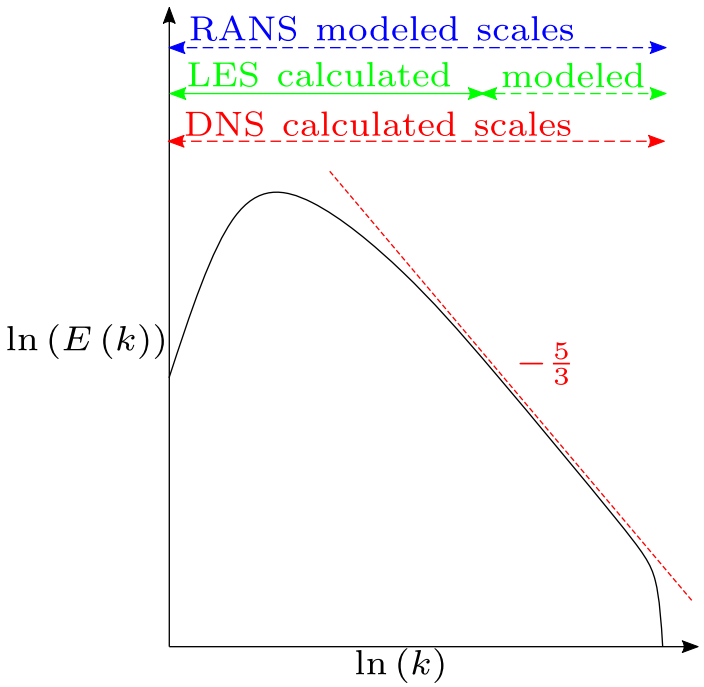


Figure 2-1: Schematic illustration of the distribution of turbulent kinetic energy by wavenumber and the different strategies available for simulation of turbulent flows.

method to flows with low Reynolds number.

- Large eddy simulation (LES) (Deardorff, 1970; Moin & Kim, 1982; Sagaut, 2006). The N-S equations are solved for a filtered velocity field $\mathbf{U}^*(\mathbf{x}, t)$ representative of the larger eddies of the turbulent flow. The different scales of the flow are separated by considering a spectral filter. Large scale structures are directly resolved, while the influence of the smaller scales on the larger ones is modelled (sub-grid modelling); this model, which must be integrated in the momentum equation, is usually derived on the assumption that the small filtered scales exhibit universal features. LES is less computationally demanding than DNS, even though it naturally represents unstationary, three-dimensional flows.

- Reynolds-Averaged Navier-Stokes simulation (RANS) (Pope, 2000; Wilcox, 1994). Using a Reynolds decomposition to the N-S equations, one can separate the mean value of the flow from the fluctuations and derive transport equations from the momentum equation for the average fields $\overline{\mathbf{U}}$. This equation must be however closed using a turbulence model.

A general formulation for the momentum equation which is valid for most of the numerous RANS and LES models presented in the literature is the following:

$$\frac{\partial \overline{\mathbf{U}}}{\partial t} + \boldsymbol{\nabla}.\left(\overline{\mathbf{U}}\overline{\mathbf{U}}\right) = \mathbf{g} - \boldsymbol{\nabla}\overline{p} + \boldsymbol{\nabla}.\left(\nu\boldsymbol{\nabla}\overline{\mathbf{U}}\right) + \tau_T \tag{2.19}$$

where $\overline{\mathbf{U}}$ and $\overline{p}$ are reduced-order representations (filtered for LES, averaged for RANS) of the flow field and τ_T is a tensor representing the effects of turbulence. In this equation, the last term represents either the turbulent stress (RANS approach) or the subgrid stress (LES approach) contributions. The closure of this last term has been at the core of intense research in the turbulence community for decades, without leading to a satisfactory universal model. Data assimilation strategies are likely to overcome partially the drawback of such turbulence closures by correcting in a more optimal way and on a case-specific basis the parameters that affect turbulence models.

2.3 Discretisation method

To computationally solve the complete or closed form of the Navier-Stokes equations, a discretization method is used to approximate the partial differential equations by a system of algebraic equations. This method describes the process by which the continuous Navier-Stokes equations are transferred into discrete counterparts. Instead of providing a continuous approximate numerical solution, Computational Fluid Dynamics (CFD) provides results at discrete locations in time and space.

The discretization method used is of utmost importance to the quality and accuracy of the numerical solution obtained. There are different approaches but the most used are (Ferziger & Peric, 1996): finite difference (FD), finite volume (FV), finite element (FE) and spectral element (SE) methods. Each one of these methods yields the same solution with very fine grids; however, some methods are more suitable than others (in terms of accuracy vs computational requirements), depending on the type of problem one is trying to numerically approximate. Interested readers are referred to Ferziger and Peric, 1996 and (Hirsch, 2007) for a comprehensive review on

these discretization techniques. Below, we briefly describe the finite difference, finite volume and spectral element methods:

1. **Finite Difference**. The starting point is the generic conservation law written in the differential form

$$\frac{\partial(\rho\theta)}{\partial t} + \boldsymbol{\nabla}.(\rho\theta\mathbf{U}) = S_{\mathcal{V}}(\theta) + \boldsymbol{\nabla}.\mathbf{F}_S(\theta). \tag{2.20}$$

   The objective is to obtain a numerical solution of this equation given initial and boundary conditions. In FD methods, the numerical grid must be structured. Each grid node is identified by a set of indices that represent the lines that intersect at this point, (i, j, k) in 3D.

   At each node, there is an unknown variable and an algebraic equation expressing the relation between the information at the node and its neighbours. This relation is obtained by approximating each term of the PDE with a finite difference approximation. These approximations can be obtained, at different order of accuracy, by combining Taylor's expansion of values at each point of the stencil. By comparison with other methods, FD offers an ideal framework to characterize the numerical error (of diffusive or dispersive nature) and possibly reduce it through coefficient optimization. In Sec. 5.2.1, the approximation error on coarse grids is more particularly characterized through the introduction of the numerical gain of schemes. The use of an inner loop of optimization is introduced in our MGEnKF algorithm to calibrate coefficients of diffusive and dispersive correction terms. The FD method has been used in almost all the numerical tests performed in this PhD study.

2. **Finite Volume** The starting point now is the generic conservation law in the integral form

$$\frac{\partial}{\partial t}\int_{\mathcal{V}\,t} \rho\theta d\mathcal{V} + \oint_{\partial\mathcal{V}\,t} d\mathbf{S}.(\rho\theta\mathbf{U}) = \int_{\mathcal{V}\,t} S_{\mathcal{V}} d\mathcal{V} + \oint_{\partial\mathcal{V}\,t} d\mathbf{F}.\mathbf{S}_S(\theta). \tag{2.21}$$

   The computational domain is subdivided into smaller control volumes (CVs) by

a grid which defines the CVs' boundaries or central points, depending on the strategy used. The generic conservation law in the integral form (2.21) applies to each CV, and to the whole control volume defined by the computational domain. Volume integrals containing divergence terms are then converted to surface integrals using the divergence theorem. Thereafter, these terms are evaluated as fluxes at the surface of each CV.

If the conservation equations for all CVs are summed, the global conservation equation over the whole domain is obtained, since surface integrals over the boundaries and over the inner CV faces cancel out. Global conservation is thus intrinsic to the method and constitutes one of the main advantages over FD, where conservation is more difficult to enforce. Algebraic equations are obtained for each CV by approximating volume and surface integrals using quadrature formulae.

3. **Spectral element method**. The spectral element method (Patera, 1984) is a high-order weighted residual technique that can be viewed as an extension of the finite element method. In this method, the computational domain is split into N_e deformable hexahedral elements. The solution on each element is represented in terms of N_{th} order tensor-product polynomials. Spectral element methods exhibit reduced numerical dispersion and dissipation. A well-known spectral-element-based CFD code which has been used in the context of this PhD study (see Chapter 6) is Nek5000 (Fischer et al., 2008).

2.4 Methods for Unsteady flows

This PhD study tackles the state and parameter estimation problem for unsteady flows. It is therefore necessary to introduce the key numerical ingredients of the discretization methods used in this type of problems. More specifically, in this section we focus on the available time discretization methods for partial differential equations. Considering the nature of time, all solution methods to unsteady flows follow a time-

marching structure, where the unsteady solution of the system is obtained in a step-by-step manner.

Let us consider the 1D linear advection equation of a scalar quantity u advected with the constant velocity c:

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0. \tag{2.22}$$

Advection is usually the main driving force in flows of engineering interest, and the linear advection equation represents the simplest flow model prototype. It describes the linear transport of a quantity by motion of a fluid.

We choose to describe time advancing in the frame of FD method for the discretization of this equation. The simplest solution method one could use to solve (2.22) is the first-order Euler method, which is actually derived via a first-order Taylor series development:

$$u_{\mathrm{j}}^{\mathrm{k}} = u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + \Delta t\, f, \tag{2.23}$$

where the spatial discretization is performed on a Cartesian mesh with a constant size Δx, Δt is the time step, $u_{\mathrm{j}}^{\mathrm{k}}$ represents the discrete numerical solution at the spatial location $x_{\mathrm{j}} = (j-1)\Delta x$ at time $t_{\mathrm{k}} = k\Delta t$, and f represents the discretization of the spatial derivative term $c\frac{\partial u}{\partial x}$. At this point, one has two options for the evaluation of f:

- The function f is evaluated at an instant $t_{\mathrm{k\text{-}1}}$. In this case, one retrieves the forward first-order Euler method

$$u_{\mathrm{j}}^{\mathrm{k}} = u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + \Delta t\, f\left(u_{\mathrm{k\text{-}1}}\right). \tag{2.24}$$

- The function f is evaluated at an instant t_{k}. In this case, one retrieves the backward, first-order Euler method

$$u_{\mathrm{j}}^{\mathrm{k}} = u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + \Delta t\, f\left(u_{\mathrm{k}}\right). \tag{2.25}$$

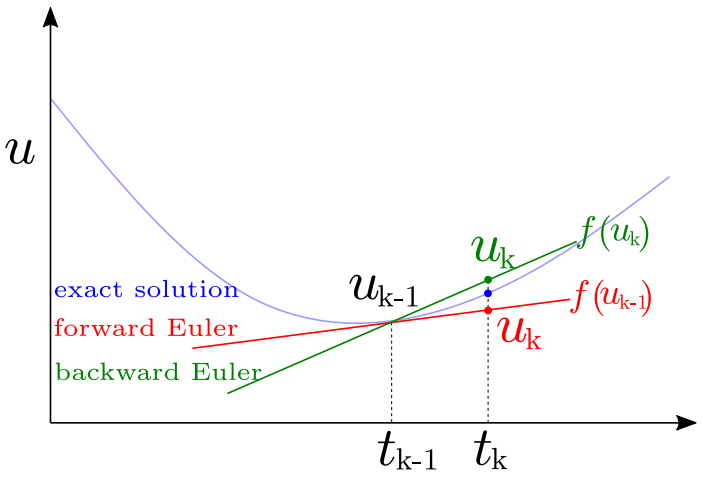


Figure 2-2: Schematic illustration of forward and backward Euler method for time integration.

The backward and forward Euler method are also illustrated in Fig. 2-2. This example allows introducing two different class of methods for resolving unsteady problems: explicit and implicit methods. In explicit methods, the system of equations obtained when applying (2.24) can be written as

$$\Phi \mathbf{u}^{\mathrm{k}} = \Psi \mathbf{u}^{\mathrm{k\text{-}1}} + \mathbf{B}\mathbf{b}^{\mathrm{k\text{-}1}}, \tag{2.26}$$

where $\mathbf{u}^{\mathrm{k}}$ represents the unknown state vector at an instant t_{k}, $\mathbf{u}^{\mathrm{k\text{-}1}}$ represents the known state vector at an instant $t_{\mathrm{k\text{-}1}}$, Φ and Ψ are matrices obtained via the discretization process, and boundary conditions are represented by the *control vector* $\mathbf{b}^{\mathrm{k\text{-}1}}$ and by $\mathbf{B}$, which is the *control input model* applied on $\mathbf{b}^{\mathrm{k\text{-}1}}$. In the framework of Data Assimilation, it is more common to use the linear or non-linear model operator $\mathcal{M}$, which includes the terms Φ, Ψ and $\mathbf{B}$ and allows writing the time advancement in a simplified manner: $\mathbf{u}^{\mathrm{k}} = \mathcal{M}\left(\mathbf{u}^{\mathrm{k\text{-}1}}\right)$. In the explicit case, the matrix of the unknown variables, Φ, is diagonal. The right-hand side of (2.26) can be straightforwardly evaluated, since the state of the system at the time $t_{\mathrm{k\text{-}1}}$ is available. The inversion of Φ is straightforward. In practical applications of explicit time-marching solution methods, the matrix form is not even used, since a node-by-node resolution of the system is much more efficient.

In implicit methods, the resulting system of equations obtained from the appli-

cation of (2.25) to each node can be written in the exact same form. However, the matrix Φ is no longer diagonal, since the spatial discretization terms present in f are evaluated at an instant where the flow variables are unknown. The inversion of Φ is much more complex. When Φ has a simple structure, such as block bidiagonal or block tridiagonal, one can use state-of-the art algorithms to perform the inversion efficiently (Ferziger & Peric, 1996).

The number of operations per time-step required for implicit methods is significantly higher when compared to explicit methods. This is compensated when numerical stability is taken into consideration. In general, implicit methods are much more stable than explicit ones, allowing for significantly larger time-steps Δt.

There is no clear "better" method for a given application (Ferziger & Peric, 1996; Hirsch, 2007). With very small Δt, both methods should provide the same solution. Depending on the particularities of the flow that is being simulated, there might be significant computational time gains if one chooses to implement implicit time-marching schemes. Implicit schemes are also more widely used in steady flow simulation. In this type of problems, the flow is started from an initial state and is advanced in time in a pseudo time-marching fashion until a steady or converged state is found. The objective is to evacuate the initial transient as fast as possible. The higher CFL numbers (the Courant–Friedrichs–Lewy number is defined as $CFL = \frac{u\Delta t}{\Delta x}$) that can be used with implicit methods allow dissipating the initial transient more efficiently than when equivalent explicit methods are used.

In setting up high-fidelity unsteady LES or DNS, the prime requirement is accuracy, since the simulation has to be able to precisely capture the different lengthscales and timescales present in the flow. In this case, techniques designed for steady problems (high CFL numbers) are not suitable. Even if numerical stability with high CFL numbers can be reached via implicit methods, the accuracy of the simulation can be severely penalized. Since accuracy requirements limit the size of the time-step, explicit methods can be efficient and the extra expense associated with implicit methods is not justified; most DNS and LES reported in the literature have used explicit time advancement methods.

There is, however, a notable exception: in the case of wall-bounded flow simulation, the regions of the flow where boundary layers are attached contain very small characteristic near-wall structures. The viscous terms involving normal derivatives to the wall may thus be sources of numerical instability, and therefore an implicit treatment of these terms might be necessary for stabilizing the simulation (Ferziger & Peric, 1996). Moreover, in simulation of flows around complex geometries, more terms might have to be treated implicitly. In general, some specific physical aspects are better represented in an implicit way, among which several problems related with immersed bodies – confined spaces.

2.5 On the sensitivity of compressible/incompressible solvers to *state-estimation procedures*

The implementation of data assimilation strategies to numerical flow solvers is particularly challenging. The numerical strategies used for solving compressible or incompressible flows are fundamentally different, and the sensitivity of numerical solvers to state-estimation procedures can vary depending on the type of flow.

The particular feature of the incompressible form of Navier Stokes equation is that pressure and velocity fields are implicitly coupled. If temperature gradients are sufficiently low, the energy equation is decoupled from the other conservation laws and the temperature can be considered just as a passive scalar. Segregated solvers are the most widely used approaches to solve such systems. Amongst the most used algorithms in LES are the Pressure Implicit with Splitting of Operators (PISO) method (Ferziger & Peric, 1996; Jasak, 1996) and the Pressure Implicit Method for Pressure Linked Equations (PIMPLE) (Holzmann, 2017) methods. The PIMPLE algorithm is actually a combination of the PISO and the Semi Implicit Method for Pressure Linked Equations (SIMPLE) (Ferziger & Peric, 1996; Jasak, 1996) algorithms.

Let us consider the basic structure of the segregated PISO solver used for incompressible flow. The starting point is the incompressible Navier-Stokes equations

(2.14). Using an implicit time discretization and a linearization of the non-linear term about the previous time-step, one obtains the following discretized system:

$$a_p \mathbf{u}_p = -\sum_N a_N \mathbf{u}_N + \Phi_0 (\mathbf{u}_0) - \nabla P = \Phi (\mathbf{u}) - \nabla P \tag{2.27}$$

where a_p and a_N are coefficients coming from the discretization of the momentum equation, $\mathbf{u}$ is the velocity vector and P is the pressure. The subscripts p and N are coefficients coming from the discretization process at a given mesh element and its neighbours, respectively. Φ_0 represents an operator applied on a known velocity from the previous time step.

The PISO algorithm operates in the following steps:

1. A predicted velocity field $\mathbf{u}^*$ is obtained from (2.27) by evaluating ∇P with the pressure solution obtained on the previous time step.

2. $\mathbf{u}^*$ does not usually comply with the zero-divergent condition:

$$\nabla \cdot \mathbf{u} = \sum_f S \times \mathbf{u}_f = 0 \tag{2.28}$$

$$\mathbf{u}_f = \left(\frac{\Phi (\mathbf{u})}{a_p} \right)_f - \left(\frac{\nabla P}{a_p} \right)_f \tag{2.29}$$

   where the subscript f indicates that the terms to which it applies are interpolated on the face centres of the mesh elements and S is the corresponding surface area. Combining (2.28) and (2.29), one can obtain a Poisson equation for P:

$$\nabla \cdot \left(\frac{\nabla P}{a_p} \right) = \nabla \cdot \left(\frac{\Phi (\mathbf{u})}{a_p} \right) = \sum_f S \times \left(\frac{\Phi (\mathbf{u})}{a_p} \right)_f . \tag{2.30}$$

   The first pressure guess used to obtain $\mathbf{u}^*$ in step 1 is updated using (2.30).

3. The pressure field obtained in step 2 does not satisfy (2.27). The velocity field is then updated with the new pressure field to obtain $\mathbf{u}^{**}$ and the procedure is repeated until convergence.

The PISO algorithm has a predictor/corrector structure, where a first prediction is then corrected to satisfy the divergence-free condition. This can be exploited in *state estimation* procedures. Meldi and Poux, 2017 have adapted the PISO algorithm to perform Kalman filtering and demonstrated the feasibility of the approach for Data Assimilation purposes. In the assimilation strategy they devised, the predicted velocity obtained in step 1 of the PISO algorithm, u^*, is updated with observation. Thereafter, the divergence-free condition is imposed on the estimated field via the PISO loop. The intrinsic structure of such algorithms thus appear to be better suited to the application of state-estimation procedures, in so far as the velocity field can be assimilated without any pressure field knowledge. Pressure and velocity will just adapt to each-other via the PISO loop.

This has an important physical consequence: the flow information carried by the estimate, which is a consequence of combining model and observation, is propagated in the entire domain in a single time-step. This constitutes a very powerful *regularisation* technique natural to the PISO algorithm. From another perspective, one could also conceive the state-estimate obtained in step 1 as a disturbance to the original predicted flow. This disturbance tends to propagate to the entire domain as the flow adapts to it. Its characteristic propagation speed is necessarily linked to the speed of the sound in the flow. However, in the incompressible flow context, the speed of sound is infinity, thus the flow is regularized in a single time-step.

Data assimilation techniques are more complex to apply to the compressible form of Navier Stokes equations, since all variables become strongly coupled and the time integration used needs to update the whole vector of conservative variables. DA algorithms introduce optimized state estimation based on probabilistic considerations, but do not naturally include any physical constraint enforcing the conservativity of state corrections. The direct application of state correction on variables (the velocity for example) considered individually, is thus likely to lead to the continuous generation of artificial waves. Iterative solvers with implicit time-marching schemes might be beneficial in this kind of situation, since they can provide a natural *regularization* of the flow. This issue is partially addressed through the algorithm proposed in this

work.

2.6 Iterative methods for flow calculation

Let us focus on the resolution methods available for implicit time integration of partial differential equations. This type of time discretization strategies can be convenient for state-estimation problems due to the iterative nature of the resolution methods used to solve the system.

We have seen that, for both implicit and explicit methods for time integration, the discretization process produces an algebraic system of equations for the unknown flow variables at each node. This system of equations might be huge in realistic CFD applications. For explicit methods, the resolution of the system is straightforward; however, in the case of implicit methods, a full algebraic system has to be solved, i.e., a matrix has to be inverted at each time step. In this section, we focus on the existing strategies to efficiently solve the type of algebraic system of equations that one encounters when using implicit time-discretisation schemes. In particular, we focus on the multigrid method, which is a key ingredient of the algorithm proposed in this work.

There are two large families of methods available for resolving linear algebraic systems: direct methods and iterative methods. In direct methods, the objective is to develop algorithms capable of efficiently inverting Φ with the least number of operations and thus to obtain the exact solution of the system. In iterative methods, one starts from an approximated solution, which is progressively corrected towards the exact solution. In practice, direct methods are seldom used in CFD, since they cannot be used directly with non-linear systems and computational costs do not scale well with increasing mesh size. The resolution methods used for implicit time integration are thus almost exclusively iterative.

Iterative and direct methods for resolving algebraic systems of equations have been treated extensively in the literature. The interested reader is referred to Barrett et al., 1994 and Saad, 2003 (among others) for a comprehensive review. In this PhD

study, we focus on the Multigrid Method, which is an efficient iterative technique, with the highest convergence rate observed amongst all the other iterative methods. The iterative nature of multigrid strategies together with their multilevel structure could be exploited in applications of DA algorithms to unsteady flow to i) ensure numerical stability and ii) reduce the computational costs.

2.6.1 Multigrid Methods

Let us consider a linear system of equations of the form

$$A\mathbf{u} = \mathbf{f} \tag{2.31}$$

Note that (2.31) can be linked with (2.26), since $A = \Phi$ and $\mathbf{f} = \Psi\mathbf{u}^{\text{k-1}} + \mathbf{B}\mathbf{b}^{\text{k-1}}$. A classical iterative scheme will be used for the resolution of this system. To this end, a fixed point iteration strategy can be employed. Let us start by decomposing the matrix

$$A = M - N, \tag{2.32}$$

where M is a non-singular matrix. The fixed point system can thus be written as

$$\mathbf{u} = F\mathbf{u} + M^{-1}\mathbf{f}, \tag{2.33}$$

where $F = M^{-1}N$. Given an initial guess $\mathbf{u}^0$, a fixed point iteration can be applied to (2.33)

$$\mathbf{u}^{\text{l}+1} = F\mathbf{u}^{\text{l}} + M^{-1}\mathbf{f}, \;\; l = 0, 1, 2, 3, ..., \tag{2.34}$$

where l is the iteration step. The error at an iteration step is given by

$$\mathbf{e}^{\text{l}} = \mathbf{u} - \mathbf{u}^{\text{l}}, \tag{2.35}$$

and the residual by

$$\mathbf{r}^{\text{l}} = \mathbf{f} - A\mathbf{u}^{\text{l}}. \tag{2.36}$$

Since (2.31) can be written as

$$A\left(\mathbf{u}^{\mathrm{l}}+\mathbf{e}^{\mathrm{l}}\right)=\mathbf{f}, \tag{2.37}$$

it follows that

$$A\mathbf{e}^{\mathrm{l}}=\mathbf{f}-A\mathbf{u}^{\mathrm{l}}=\mathbf{r}^{\mathrm{l}}, \tag{2.38}$$

and the residual correction equation can be written as $\mathbf{u}=\mathbf{u}^{\mathrm{l}}+\mathbf{e}^{\mathrm{l}}$.

In multigrid methods, one can use (2.38) for updating $\mathbf{u}^{\mathrm{l}}$. An approximation of the error $\tilde{\mathbf{e}}^{\mathrm{l}}$ is computed from (2.38), and an update for $\mathbf{u}^{\mathrm{l}}$ is calculated as $\mathbf{u}^{\mathrm{l}+1}=\mathbf{u}^{\mathrm{l}}+\tilde{\mathbf{e}}^{\mathrm{l}}$. This iterative procedure is performed on a hierarchy of grids.

The basis for this class of methods comes from an observation on the rate of convergence of iterative methods, which is controlled by the eigenvalues of the method-dependant iteration matrix. On a general basis, low-frequency components in the error spectrum are the slowest to be damped in the iterative procedure, while higher frequencies are rapidly smoothed. The multigrid method exploits this property: with a few iteration sweeps through a given initial mesh, the high frequency part of the error spectrum will be significantly damped. The remaining low frequencies could be transferred to successive coarser grids, where this error automatically becomes part of the high frequency range, due to the change in spatial resolution. This strategy allows to progressively damp the low-frequency error in the error spectrum using a hierarchy of coarser grids.

The last necessary ingredient to formalize this method is the definition of a transfer function (or projection operator Π) enabling the communication between different grids. For this, the simplest option is linear interpolation. A schematic representation of a three-level multigrid strategy in a V-cycle arrangement is presented in Fig. 2-3 as an example.

It should be noted that the interpolation of a flow solution from a coarse grid to a fine grid or vice-versa may not guarantee conservativity, i.e., the total energy of the flow might not necessarily be conserved thus introducing a systematic error in the transformation. This is not necessarily an issue in the multigrid strategy,

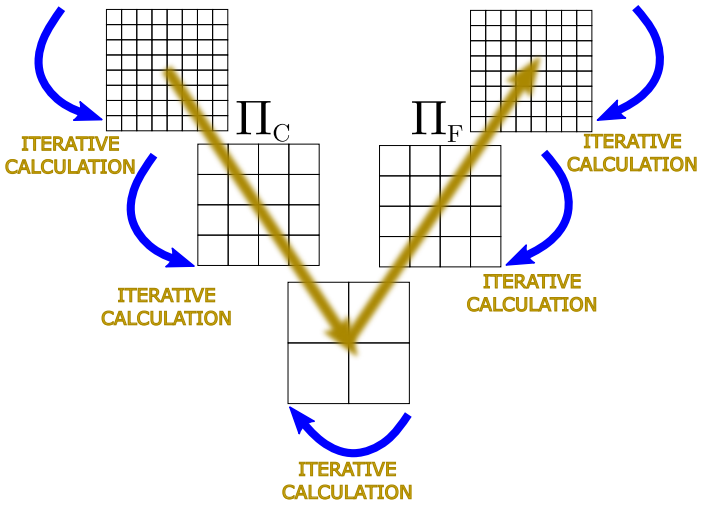


Figure 2-3: Schematic illustration of a three-level multigrid V-cycle.

since the final iterative sweeps are always performed on the fine grid, thus ensuring the highest possible level of accuracy. Therefore, the multigrid strategy acts as an "accelerator" of the classical iterative procedure, where a single grid is used. In the context of the sequential algorithm that we will present in the next chapters, state updates on the finest level of the multigrid approach will be obtained from statistics calculated on a coarser, lower-fidelity model. In this situation, projection operators have to be used and "conservativity" is broken twice: the statistical state update does not respect the original model and the interpolators used disrupt even more the total energy of the flow. However, the projection term is used to extend a correction term calculated on the coarse grid to the fine grid. This correction term is normally orders of magnitude lower than the actual magnitude of the variables of the flow, thus interpolation errors can be neglected. One could also consider interpolation errors as an intrinsic modelling uncertainty, degrading the overall confidence in the forecast.

Multigrid approaches are extensively used in CFD applications (Ferziger & Peric, 1996; Hackbusch, 1985). The solution is found on the computational grid by updating an initial guess via multiple estimations obtained on a hierarchy of discretizations. Two well-known families of multigrid approaches exist, namely the *algebraic* multigrid method and the *geometric* multigrid method. With algebraic multigrid methods (see Falgout, 2006 and references therein), a hierarchy of operators is directly constructed

from the state transition model (represented by Ψ and Φ). The geometric multigrid methods provide the solution via a set of operations performed in two (or more) meshes. In this PhD study, we focus on the latter technique, and we consider the simplified case of two grids. Thereafter, the variables defined on the fine grid will be denoted with the superscript F ($\mathbf{u}^{\mathrm{F}}$ for instance), and those defined on the coarse grid will be denoted with the superscript C ($\mathbf{u}^{\mathrm{C}}$ for instance).

The coarse-level representation $\mathbf{u}^{\mathrm{C}}$ is usually obtained suppressing multiple mesh elements from the initial fine-level one $\mathbf{u}^{\mathrm{F}}$ or by linear interpolation. This operation may be defined by a coarsening ratio parameter r_{C}, which indicates the total number of elements on the fine grid over the number of elements conserved in the coarse grid. Among the numerous algorithms proposed for geometric multigrid, we use the Full Approximation Scheme (FAS), which is a well-documented strategy (Brandt, 1977; Wesseling & Oosterlee, 2001).

Let us recall the general two-time levels implicit discretization of a non-linear set of partial differential equations:

$$\Phi\left(\mathbf{u}^{\mathrm{k}}\right) = \Psi\left(\mathbf{u}^{\mathrm{k}}\right) + \mathbf{B}\left(\mathbf{b}^{\mathrm{k\text{-}1}}\right) := \mathbf{c}. \tag{2.39}$$

The general formulation of two-grid algorithm that is followed in the present work is described hereafter. The time subscript k is excluded for clarity. The superscript l represents the iteration step of the procedure.

1. Starting from an initial solution on the fine grid $\left(\mathbf{u}^{0}\right)^{\mathrm{F}}$, an iterative loop (or smoothing operation which can be linear or non-linear) is used on (2.39) to obtain a first solution $\left(\mathbf{u}^{1}\right)^{\mathrm{F}}$. A residual $\left(\delta^{1}\right)^{\mathrm{F}} = \mathbf{c}^{\mathrm{F}} - \Psi^{\mathrm{F}}\left(\mathbf{u}^{1}\right)^{\mathrm{F}}$ is calculated, where the high-frequency error components in the error spectrum are smoothed.

2. $\left(\mathbf{u}^{1}\right)^{\mathrm{F}}$ and $\left(\delta^{1}\right)^{\mathrm{F}}$ are projected from the fine grid to the coarse grid space via a projection operator Π_{C}, so that $\left(\mathbf{u}^{1}\right)^{\mathrm{C}}$ and $\left(\delta^{1}\right)^{\mathrm{C}}$ are obtained. Similarly, the state transition model Ψ^{F} is projected on the coarse grid (i.e., re-estimated based on the projection of the solution of the fine grid onto the coarse grid) to obtain Ψ^{C}. Finally, we evaluate $\mathbf{c}^{\mathrm{C}} = \Psi^{\mathrm{C}}\left(\mathbf{u}^{1}\right)^{\mathrm{C}} + \left(\delta^{1}\right)^{\mathrm{C}}$.

3. A few iterative sweeps are performed on (2.39) to obtain $(\mathbf{u}^2)^{\mathrm{C}}$ on the coarse grid using as initial solution $(\mathbf{u}^1)^{\mathrm{C}}$.

4. The updated variable on the fine grid is obtained as $(\mathbf{u}^2)^{\mathrm{F}} = (\mathbf{u}^1)^{\mathrm{F}} + \Pi_{\mathrm{F}} \left((\mathbf{u}^2)^{\mathrm{C}} - (\mathbf{u}^1)^{\mathrm{C}} \right)$ where Π_{F} is a projection operator from the coarse grid to the fine grid.

5. At last, the final solution $(\mathbf{u}^3)^{\mathrm{F}}$ is obtained with an iterative loop applied to (2.39) on the fine grid, starting from the intermediate solution $(\mathbf{u}^2)^{\mathrm{F}}$.

This procedure can be repeated multiple times imposing $(\mathbf{x}^0)^{\mathrm{F}} = (\mathbf{x}^3)^{\mathrm{F}}$ at the beginning of each cycle. When the convergence is reached, the fine grid solution at time instant k is equal to $(\mathbf{x}^3)^{\mathrm{F}}$. The strategy used when more than two discretization levels are employed is similar. In this work, the two projection operators (Π_{F} and Π_{C}) are chosen to be 4-th order Lagrange interpolators.

The inherent algorithmic structure of the multigrid strategy could be exploited in state-estimation problems where flows are directly updated using observation and error statistics. The idea is to let the flow "adapt" to the imposed correction through the iterative loops of the multigrid algorithm. But a compromise is needed between regularization and state correction: the updated state is naturally closer to the *true* state of the flow, but the stability of the numerical simulation needs to be assured and thus the need of "regularization".

Chapter 3

Data assimilation: A general overview and a focus on sequential algorithms

3.1 Introduction

The key ingredients of the numerical resolution methods of unsteady flows have been presented in the previous chapter. In a flow simulation, the discretization strategy employed together with the imposed boundary conditions constitute the *state transition model*, or *model* of the system. State predictions obtained from models are necessarily affected by uncertainty Nonetheless, these predictions can be corrected by observation. Data assimilation provides the mathematical tools necessary for this venture.

What does Data Assimilation (DA) mean? The simplest explanation one could give is that it is a method for combining experimental measurements/observations with model realizations (CFD) with the objective of improving the state prediction provided by the latter. More specifically, DA provides a sound and elegant mathematical framework for this endeavour. Observations supply information of the "real" flow. However, they are usually noisy and limited in time and space. One could think of simply interpolating/extrapolating the data, but this does not work since the phenomena analysed in the domain of fluid mechanics are strongly non-linear. On the other hand, with CFD one can obtain complete noise-free map representations of

flow phenomena. Moreover, models can be used to perform predictions about future states. However, it is well known that due to inherent epistemic uncertainty (systematic errors), model predictions are prone to drifting. It is also important to note that even if the model contained no systematic errors, initial and boundary conditions cannot be known with perfect precision. Long term behaviour of flows is unpredictable due to the high sensitivity of the N-S equations to initial and boundary conditions, in particular for high Reynolds regimes. To address this issue, model predictions have to be corrected by reality and DA provides the mathematical tools necessary for the optimal fitting of the model output with observations.

The DA problem is formulated as an optimization procedure where the objective is to find the optimal match between model output and measurement. This optimization procedure can be performed using at least two different strategies:

1. Classical optimization procedures involving the minimization of a positive, quadratic cost function that expresses the mismatch between measurements and model. This is usually defined as a least-squares error minimization problem.

2. Statistical optimization involving the minimization of the uncertainty of the model output with observation data.

3.2 Uncertainty in model and observation

Model and observations are the two sources of information that allow to analyse physical systems, but the data gathered from them is imperfect and limited. Before establishing the optimal mathematical framework for combining them, a method to rigorously quantify the uncertainty in models and observations is required. The field of uncertainty quantification (UQ) provides the tools necessary to replace the subjective concept of *confidence* in the data by a rigorous and mathematical quantification of the uncertainty in a system.

There are two main types of uncertainty: epistemic and aleatory. On the one hand, epistemic uncertainty is related to the imperfect knowledge of the system and

it can be reduced; however, its sources are not usually known *a priori*. On the other hand, some systems are inherently noisy with a built-in randomness (or stochasticity). This aleatory uncertainty is irreducible.

Let us take a closer look at the uncertainty present in models (CFD in our case) and the errors related to experimental measurements. Model assumptions, discretization and imperfect knowledge of boundary/initial conditions distort the output that one can obtain from numerical simulation. While most of the uncertainty present in CFD is of epistemic nature, i.e., it can be reduced by using more accurate models and better discretization strategies, boundary and initial conditions in realistic simulations exhibit aleatory uncertainty since they cannot be known perfectly. In the case of experimental measurements, there is epistemic uncertainty related to the imperfect characterization of the set-up and miss-calibration, as well as random errors caused by inherently unpredictable fluctuations in the readings of the measurement apparatus.

The uncertainty in both model and observation can be represented via probability density functions (PDF), as shown in Fig. 3-1. To make predictions about future states, one can forecast a prior distribution using UQ tools, to obtain a probabilistic description (PDF) of future states. If observation is available, the obtained model prediction can be corrected with DA tools. An accurate quantification of the uncertainty is crucial in data assimilation methods, since it allows to meaningfully weight the different ingredients of the assimilation process.

UQ is a wide domain, covering many problems in the natural sciences. A detailed description of UQ methods, although important to DA, is out of the scope of this present research work. Interested readers are referred to Owhadi et al., 2013 and Le Maître and Knio, 2010 for a complete, mathematical overview of UQ.

3.3 Notation in DA problems

The necessary ingredients to formalize the DA problem have been presented. Two sources of information are available, and their uncertainty is quantified via PDFs. In

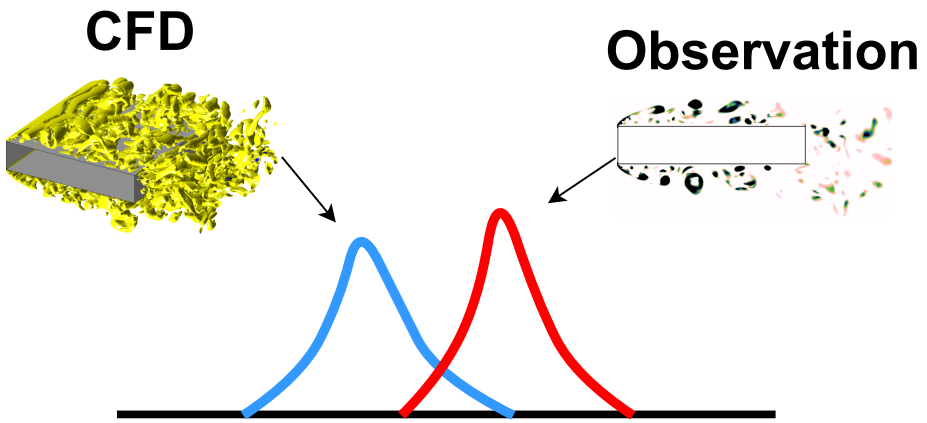


Figure 3-1: Schematic representation of the uncertainty in model (CFD) and observation, both depicted via probability density functions (PDF).

this section, we introduce the notation for DA problems used throughout this research work. Our notation is identical (when possible) to the one formalized by Ide et al., 1999 and used in other DA books (Asch et al., 2016). We consider the evolution of a system from time $t_{k\text{-}1}$ to time t_k obtained with a discrete physical model:

$$\mathbf{x}_k = \mathcal{M}_{k:k\text{-}1}\left(\mathbf{x}_{k\text{-}1}, \boldsymbol{\theta}_k\right) + \boldsymbol{\eta}_k, \tag{3.1}$$

where $\mathbf{x}_k$ represents the state vector of dimension N_x at an instant t_k, $\mathcal{M}_{k:k\text{-}1}$ is the non-linear model operator, and $\boldsymbol{\theta}_k$ contains the parameters that affect the model. The term $\boldsymbol{\eta}_k$ is associated with uncertainties in the model prediction which, as discussed before, could emerge, for example, from incomplete knowledge of initial/boundary conditions. These uncertainties are modelled by a zero-mean Gaussian distribution characterized by a covariance $\mathbf{Q}_k$, *i.e.* $\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$. An error covariance matrix, $\mathbf{P}_k = \mathbb{E}\left[\left(\mathbf{x}_k - \mathbb{E}(\mathbf{x}_k)\right)\left(\mathbf{x}_k - \mathbb{E}(\mathbf{x}_k^{f/a})\right)^\top\right]$, associated to $\mathbf{x}_k$ characterizes the total uncertainty in the system at a time t_k. As we will see later in the manuscript, even though the real distributions might not be Gaussian, this assumption is very advantageous from a mathematical point of view.

Throughout this manuscript, we used the standard notations employed in data assimilation studies. We will therefore make the difference between $\mathcal{M}_{k:k\text{-}1}$, which is the non-linear dynamical model, and $\mathbf{M}_{k:k\text{-}1}$, which is its linearized version. In this

last case, (3.1) can be written as

$$\mathbf{x}_{\mathrm{k}} = \mathbf{M}_{\mathrm{k:k\text{-}1}}\mathbf{x}_{\mathrm{k\text{-}1}} + \boldsymbol{\eta}_{\mathrm{k}}, \tag{3.2}$$

Indirect observations of $\mathbf{x}_{\mathrm{k}}$ are available in the components of the observation vector $\mathbf{y}^{\mathrm{o}}_{\mathrm{k}}$. These two variables are related by:

$$\mathbf{y}^{\mathrm{o}}_{\mathrm{k}} = \mathcal{H}_{\mathrm{k}}(\mathbf{x}_{\mathrm{k}}) + \boldsymbol{\epsilon}^{\mathrm{o}}_{\mathrm{k}}. \tag{3.3}$$

where $\mathcal{H}_{\mathrm{k}}$ is the non-linear observation operator which maps the model state space to the observed space. The available measurements are also affected by uncertainties which are assumed to follow a zero-mean Gaussian distribution characterized by a covariance $\mathbf{R}_{\mathrm{k}}$, *i.e.* $\boldsymbol{\epsilon}^{\mathrm{o}}_{\mathrm{k}} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{\mathrm{k}})$. The so-called observation error covariance matrix, $\mathbf{R}$, describes instrument (aleatory uncertainty) and calibration (epistemic uncertainty) errors in the available measurement data. The dimension of the observation vector $\mathbf{y}^{\mathrm{o}}_{\mathrm{k}}$, N_y, is usually order of magnitudes smaller than the dimension of the state $N_y \ll N_x$. The linearized version of (3.3) can be written as

$$\mathbf{y}^{\mathrm{o}}_{\mathrm{k}} = \mathbf{H}_{\mathrm{k}}\mathbf{x}_{\mathrm{k}} + \boldsymbol{\epsilon}^{\mathrm{o}}_{\mathrm{k}}, \tag{3.4}$$

where $\mathbf{H}_{\mathrm{k}}$ is the linearized observation operator.

In the standard DA notation, subscripts are used to denote time/space indexes; superscripts refer to the nature of the DA variables: "a" for analysis (the estimate obtained with the DA procedure), "b" for initial guess or background, "f" for forecast, "o" for observation and "t" for the real or *true* state of the system (usually unknown).

3.4 Variational DA

Two sources of information for describing fluid flow have been identified. The notation used to describe them has been defined in the previous section. We are now ready to introduce the first class of DA methods that aim at combining experimental and

numerical data: variational DA.

How do we combine observation and model in this framework? In variational DA, one can introduce a cost function describing the mismatch between observation and model with the objective of minimizing it. These methods are based on optimal control theory and require numerical optimization techniques. In the minimization procedure, the gradient of the cost function is sought with the aid of the adjoint method.

In this section, the basic ingredients required for the formulation of the variational DA problem are presented. Thereafter, 3D-VAR and 4D-VAR methods are briefly introduced. A literature review of the applications of variational methods to fluid mechanics problems is also provided, with a focused view on the strong points and the difficulties that one faces when applying these class of methods to the analysis of fluid flows.

Variational DA methods have been extensively studied in the context of weather forecast for decades. Solution methods for variational DA problems involve the combination of skills from different mathematical domains: functional analysis, numerical optimization, partial differential equations, variational calculus, ... The interested reader is referred to the following bibliographic references for an exhaustive guide to these different fields: Courant and Hilbert, 1989 for variational calculus; Kreyszig, 1978 and Reed and Simon, 1981 for functional analysis; and Nocedal and Wright, 2006 for numerical optimization techniques. Furthermore, we recommend the book written by Asch et al., 2016 and the thesis by M. Rochoux, 2014, which cover comprehensively both statistical and variational approaches to DA.

3.4.1 Formulation of the 3D-VAR and 4D-VAR methods

The formulation of the variational DA problem is based on two key ingredients: *calculus of variations* and the *adjoint method*. Considering that variational methods were not used in this PhD study, these two key elements are not treated in detail in this manuscript. In short, the partial differential Euler-Lagrange equations coming from the *calculus of variations* mathematical field provide a general framework for

minimizing functionals, and the adjoint method makes this minimization problem computationally tractable. Let us review the key points in the formulation of the two well-known variational DA methods: 3D-VAR and 4D-VAR (Asch et al., 2016; Talagrand, 2014).

The 3D-VAR and 4D-VAR methods are presented in the discrete context. Consider the following cost function where the time subscript is dropped for simplicity:

$$\mathcal{J}(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right)^{\mathrm{T}}\mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right) + \frac{1}{2}\left(\mathbf{H}(\mathbf{x})-\mathbf{y}^{\mathrm{o}}\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{H}(\mathbf{x})-\mathbf{y}^{\mathrm{o}}\right) \tag{3.5}$$

where $\mathbf{x}, \mathbf{x}^{\mathrm{b}}$ and $\mathbf{y}^{\mathrm{o}}$ represent the target, background and observed state, respectively; $\mathbf{H}$ represents the observation matrix (linearized version of $\mathcal{H}$) and $\mathbf{B}$ and $\mathbf{R}$ are the background and observation error covariance matrices, respectively. This cost function attempts to combine an *a priori* knowledge about the state of the system (background) with an actual observation of the state. The target state, $\mathbf{x}^{\mathrm{a}}$, should minimize the deviation to the background state, $\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right)$, and the deviation to the observed state, $(\mathbf{H}(\mathbf{x})-\mathbf{y}^{\mathrm{o}})$, taking into account observation and background errors ($\mathbf{R}$ and $\mathbf{B}$, respectively).

3.4.1.1 3D-VAR

Consider the state of the system at a time t_{k}. If a background state, $\mathbf{x}^{\mathrm{b}} = \mathbf{x}^{\mathrm{t}} + \boldsymbol{\epsilon}^{\mathrm{b}}$, is available as well as an observation of the form $\mathbf{y}^{\mathrm{o}} = \mathbf{H}\mathbf{x}^{\mathrm{t}} + \boldsymbol{\epsilon}^{\mathrm{o}}$, the minimization of the cost function (3.5) produces an estimate at the discrete time t_{k}. This procedure is known as three-dimensional variational analysis (3D-VAR).

It is interesting to note that there is a very well known analytical solution for the minimization of the cost function in (3.5). One can take the gradient of $\mathcal{J}$ and apply the optimality condition

$$\nabla\mathcal{J}(\mathbf{x}^{\mathrm{a}}) = \mathbf{B}^{-1}\left(\mathbf{x}^{\mathrm{a}}-\mathbf{x}^{\mathrm{b}}\right) + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{y}^{\mathrm{o}}-\mathbf{H}\mathbf{x}^{\mathrm{a}}\right) = 0, \tag{3.6}$$

where $\mathbf{x}^{\mathrm{a}}$ is the target analysis state

$$\mathbf{x}^{\mathrm{a}} = \arg\min \mathcal{J}\left(\mathbf{x}\right) \tag{3.7}$$

The well-known solution of this equation is

$$\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{b}} + \mathbf{K}\left(\mathbf{y}^{\mathrm{o}} - \mathbf{H}\mathbf{x}^{\mathrm{b}}\right), \tag{3.8}$$

where $\mathbf{K}$ represents the gain matrix,

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}\left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)^{-1}. \tag{3.9}$$

The step-by-step resolution of (3.6) is not reported here for simplicity. A detailed resolution of this equation can be found in any DA books (see Asch et al., 2016).

The solution obtained for (3.7) shows that the cost function $\mathcal{J}$ presented in (3.5) is minimized when $\mathbf{x}^{\mathrm{a}}$ is constructed as a linear combination between a background state and a vector expressing the difference between measurement and background, usually denominated as innovation vector $\mathbf{d} = \mathbf{y}^{\mathrm{o}} - \mathbf{H}\mathbf{x}^{\mathrm{b}}$. The innovation vector is multiplied by a gain matrix that depends strongly on the error statistics of background and measured state. A closer look at the calculation of $\mathbf{K}$ shows that, in the scalar case, the denominator term, $\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$, represents the sum of background and observation error covariances in the observation space, while the numerator, $\mathbf{B}\mathbf{H}^{\mathrm{T}}$, expresses only the background error. This is analogous to a variance ratio

$$\frac{\sigma^{\mathrm{b}}}{\sigma^{\mathrm{b}} + \sigma^{\mathrm{o}}}, \tag{3.10}$$

where σ^{b} and $\sigma^{\mathbf{o}}$ represent the background and the observation error variances, respectively. This shows that the optimization of the cost function in (3.6) produces a result with a strong statistical tie.

In real situations, the dimensions of the error covariance matrices can be so large that their storage/manipulation can be problematic. The direct calculation of the

gain $\mathbf{K}$ with (3.9) is thus usually infeasible.

To avoid this issue, the cost function defined in (3.5) can be minimized in an iterative fashion by combining an adjoint approach for the computation of the gradient with a descent algorithm for obtaining the direction where minima can be found. State-of-the-art numerical optimization techniques allow performing this task in a computationally efficient manner (Nocedal and Wright, 2006).

3.4.1.2 4D-VAR

Consider a situation where multiple observations over a time interval are available. These observations are available at a succession of instants, t_{k}, $k = 0, 1, ..., K$, and can be expressed via (3.4).

Model dynamics during the time interval analysed can be represented using (3.2). However, the model is considered to be deterministic, *i.e.* $\boldsymbol{\eta}_{\mathrm{k}} = 0$. With known background initial state, $\mathbf{x}_0^{\mathrm{b}}$, and associated covariance matrix, $\mathbf{P}_0^{\mathrm{b}}$, one can slightly modify the cost function presented in (3.5) and express it as

$$\begin{aligned} \mathcal{J}(\mathbf{x}_0) &= \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}^{\mathrm{b}}\right)^{\mathrm{T}} \left(\mathbf{P}_0^{\mathrm{b}}\right)^{-1} \left(\mathbf{x}_0 - \mathbf{x}^{\mathrm{b}}\right) \\ &\quad + \sum_{\mathrm{k}=0}^{\mathrm{K}} \frac{1}{2}\left(\mathbf{H}_{\mathrm{k}}\mathbf{x}_{\mathrm{k}} - \mathbf{y}_{\mathrm{k}}^{\mathrm{o}}\right)^{\mathrm{T}} \left(\mathbf{R}_{\mathrm{k}}\right)^{-1} \left(\mathbf{H}_{\mathrm{k}}\mathbf{x}_{\mathrm{k}} - \mathbf{y}_{\mathrm{k}}^{\mathrm{o}}\right). \end{aligned} \tag{3.11}$$

The state sought with the minimization of $\mathcal{J}$ is the initial state $\mathbf{x}_0$ of the system that best fits the available observation. When the model is considered to be error-free (deterministic approach), the analysis performed to find $\mathbf{x}_0$ is denominated strong constraint 4D-VAR. This means that the sequence of states obtained through the optimization approach, starting from $\mathbf{x}_1$ to $\mathbf{x}_{\mathrm{K}}$, have to strictly respect the model equation (3.2). In Fig. 3-2, the differences between 3D and 4DVar are depicted. While 3DVar performs an analysis at a specific time instant, 4DVar performs an optimization procedure over a defined assimilation time window.

If model uncertainty is considered in (3.2), and if it is assumed that model errors

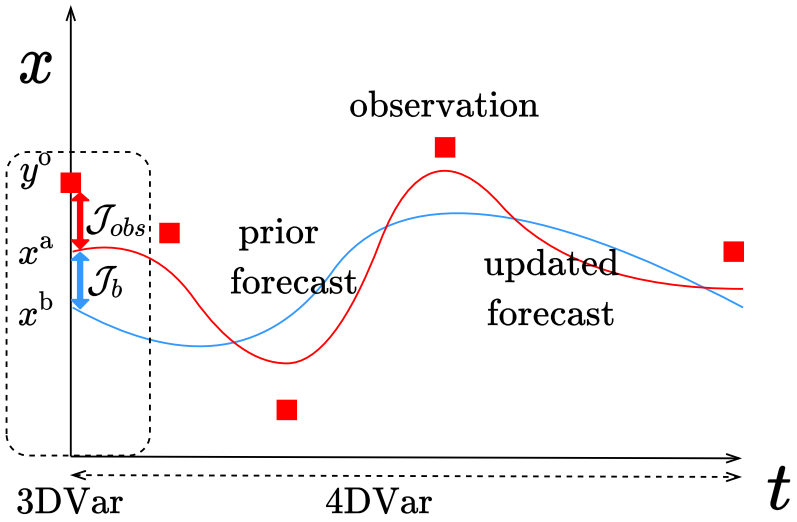


Figure 3-2: 3DVar vs 4DVar. While 3DVar is performed at a specific time instant with a single observation, a 4DVar analysis is performed over a time window including multiple observations.

are uncorrelated in time and uncorrelated with the background and observation errors, the objective function becomes

$$\begin{aligned}\mathcal{J}\left(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{\mathrm{K}}\right) &= \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}^{\mathrm{b}}\right)^{\mathrm{T}}\left(\mathbf{P}_0^{\mathrm{b}}\right)^{-1}\left(\mathbf{x}_0 - \mathbf{x}^{\mathrm{b}}\right) \\ &+ \frac{1}{2}\sum_{\mathrm{k}=0}^{\mathrm{K}}\left(\mathbf{H}_{\mathrm{k}}\mathbf{x}_{\mathrm{k}} - \mathbf{y}_{\mathrm{k}}^{\mathrm{o}}\right)^{\mathrm{T}}\left(\mathbf{R}_{\mathrm{k}}\right)^{-1}\left(\mathbf{H}_{\mathrm{k}}\mathbf{x}_{\mathrm{k}} - \mathbf{y}_{\mathrm{k}}^{\mathrm{o}}\right) \\ &+ \frac{1}{2}\sum_{\mathrm{k}=1}^{\mathrm{K}}\left(\mathbf{x}_{\mathrm{k}} - \mathbf{M}_{\mathrm{k:k\text{-}1}}\mathbf{x}_{\mathrm{k\text{-}1}}\right)^{\mathrm{T}}\left(\mathbf{Q}_{\mathrm{k}}\right)^{-1}\left(\mathbf{x}_{\mathrm{k}} - \mathbf{M}_{\mathrm{k:k\text{-}1}}\mathbf{x}_{\mathrm{k\text{-}1}}\right). \quad (3.12)\end{aligned}$$

Note that the cost function is now dependent on all the successive states and not only on the initial state. This analysis is denominated weak constraint 4D-VAR.

As discussed before, for the optimization procedure to be computationally feasible, one has to resort to the adjoint method. The details of this procedure are not reported here. The interested reader is referred to Talagrand, 2014 and Asch et al., 2016 for a detailed description of the usage of the adjoint methods in variational DA.

3.4.2 Applications of Variational DA algorithms for fluid mechanics problems: a critical review

We proceed now with a review of the usage of variational methods in DA problems for fluid mechanics applications. We will see that starting from the late 2000s, researchers in fluid mechanics have shown increasing interest in variational DA algorithms. Today, this is a hot topic of research in the community. Variational DA methods can be applied to a wide range of problems such as reconstruction of instantaneous/mean flows, estimation of realistic boundary conditions and data-driven models.

$\hookrightarrow$***Flow reconstruction***

Variational DA tools have been extensively used for flow reconstruction problems, where the target function is defined in terms of the deviation of a model (e.g., steady state solution of the Navier Stokes equations) to available observation. Examples of the usage of this approach in fluid mechanics are vast. At first, these methods were aimed at improving particle image velocimetry (PIV) techniques. Classical PIV algorithms had relied on correlation methods to obtain the velocity fields. In Ruhnau et al., 2005, the authors extracted the velocity fields from PIV measurements with a variational approach, where the velocity field obtained through minimization of a cost function also satisfies additional physical constraints (model embedded in this cost function). Different physical regularization techniques were investigated, such as imposing the Stokes equations (Ruhnau & Schnörr, 2006) or the unstationary vorticity transport equations (Ruhnau et al., 2007) to the reconstruction of the velocity fields.

Later, other authors have used the full Reynolds-Averaged Navier-Stokes operator to constrain the minimization of the cost function in the flow reconstruction procedure. Foures et al., 2014 attempted to reconstruct the time-averaged velocity fields from a succession of PIV snapshots of the flow around a cylinder at low Reynolds number using a variational approach. Similarly, Symon et al., 2017, reconstructed the average flow around an idealized airfoil from 2D PIV measurements under high-Reynolds conditions, and more recently, Symon et al., 2020 combined a variational

DA approach with resolvent analysis for reconstructing mean and unsteady flows.

↪***Improvement of reduced-order models***

The improvement of reduced-order models for fluid mechanics with variational DA techniques has also been explored. D'adamo et al., 2007 used a variational approach together with PIV measurements to estimate the dynamic coefficients of a POD Galerkin model. Artana et al., 2012 investigated strong and weak constraint variational methods to estimate the initial state of the POD modes as well as the reduced-order model coefficients in an attempt to reduce the stability issues inherent to standard POD Galerkin methods. Tissot, 2014 used the 4D variational method to derive POD and DMD reduced-order models for the turbulent wake behind a cylinder from PIV measurements.

↪***Estimation of realistic boundary conditions and sensitivity of flow to changes in initial/boundary conditions***

The determination of realistic boundary conditions in fluid mechanics situations is a problem that can also be tackled with variational DA algorithms. Gronskis et al., 2013 generated realistic inflow boundary conditions for DNS simulations from noisy data using a variational approach. Theoretical analysis in turbulence, tackling the sensitivity of the flow with respect to initial conditions, have been performed using a variational approach in Mons et al., 2014. Using a similar approach, Mons et al., 2017 and Mons and Marquet, 2021 investigated the use of variational methods to develop an optimal sensor placement technique which consisted in identifying flow regions that have the greatest sensitivity to changes in initial/boundary conditions or model parameters.

↪***Hybrid methods***

One of the major drawbacks of variational DA methods is the need of adjoint and tangent linear operators, which can be very cumbersome to derive and to maintain. To avoid this issue, fluid mechanics researchers are starting to resort to hybrid ensemble

variational methods (usually called EnVAR methods) where the sensitivity of the cost function is no longer evaluated with the adjoint method but via an ensemble of realizations. Yang et al., 2015 compared the performance of the traditional 4D-VAR method to the performance of the hybrid ensemble-based 4D-VAR for reconstructing flow fields from incomplete noisy data. Similarly, Mons et al., 2016 compared the estimation capabilities of variational (4D-VAR), sequential (EnKS) and ensemble-variational (4DEnVAR) data assimilation techniques in the context of unsteady CFD. It was demonstrated that the three techniques can be used to accurately reconstruct the initial and boundary conditions of an unsteady 2D flow from noisy observations.

Other authors have also reported successful application of the 4DEnVAR method to fluid mechanics problems. In Yang and Mémin, 2017, the authors used this data assimilation technique to infer both the initial conditions and the inhomogeneous time-varying parameters of a stochastic subgrid model. Mons et al., 2019 reported successful reconstruction of the spatial distribution of a scalar source in a turbulent channel flow from remote measurements of concentration using ensemble variational techniques. In a very recent work of the same author (Mons et al., 2021), corrections to LES-subgrid models were developed through variational assimilation of measured statistical quantities.

$\hookrightarrow$ ***Weak-constraint 4D-VAR applications***

Finally, an important aspect, which is yet to be fully explored, is the consideration of modelling errors in the optimization of the cost function. 4D-VAR applications in fluid mechanics are usually strongly constrained, i.e., the model is considered to be deterministic. Weak constraint 4D-VAR applications are less common due to their increased computational cost. Moreover, when modelling uncertainty is taken into account, both modelling and observation errors are usually considered in terms of additive Gaussian noise (diagonal covariance matrices $\mathbf{Q}$ and $\mathbf{R}$). It is also common to assume that model and observation errors are uncorrelated, but in real applications this might not be realistic (Tandeo et al., 2020).

Chandramouli et al., 2020 have partially tackled this issue by considering a stochas-

tic subgrid model to the 3D Navier-Stokes equations that accounts for unresolved turbulent scales of motion. This stochastic subgrid model was parametrized, and the associated coefficients were optimized with a 4D variational approach. The assimilation strategy proposed is a compromise between strong and weak variational methods, where the model error is not considered from the pure non-informative Gaussian point of view but rather through a parametrized stochastic model.

$\hookrightarrow$***Conclusions***

In conclusion, applications of variational DA methods in fluid mechanics problems have been extensively used in the last years. These methods provide a robust strategy for combining experimental and numerical data, an endeavour still at an early stage in fluid mechanics research. Here are some of the strong points of variational DA strategies applied to fluid mechanics:

1. The 4DVar method has proven to be an excellent tool for flow reconstruction from limited observation data.

2. Full non-linear analysis can be performed with variational DA. Non-linear model and observation operators do not pose an intrinsic challenge to the existing state-of-the-art numerical optimization techniques.

3. Variational methods do not require the explicit propagation of the error covariance matrices, as it is the case in statistical estimation techniques (KF). This procedure is implicit within the optimization of the cost function, which significantly reduces the computational costs of the assimilation and makes the 4D-VAR an attractive operational solution (as it is the case in numerical weather forecast).

However, there are significant challenges that one faces when applying variational DA methods to fluid mechanics problems:

1. Variational DA methods require the implementation of an adjoint code to compute the gradient of $\mathcal{J}$. This can be highly code-intrusive and extremely cum-

bersome to develop. Ensemble-variational methods allow avoiding this issue, providing an interesting alternative to classical variational approaches.

2. In variational DA methods, the analysed state is obtained through the minimization of a cost function. The propagation of the error covariances is implicit within this optimization procedure. The advantage of this is a significant computational cost reduction. At the same time, this represents an important drawback since, depending on the application, accurate knowledge of the error statistics at a given instant can be important.

3. In 4DVar methods, the gradient of $\mathcal{J}$ with respect to the control vector is efficiently evaluated with the adjoint method. However, depending on the size of the state vector, computational costs can increase significantly. This is especially the case in 3D unsteady flow, since the forward solution needs to be stored in memory for the back propagation via the adjoint code.

3.5 Sequential data assimilation in fluid dynamics

We have introduced variational DA strategies and reviewed multiple examples of flow analysis performed with the tools that this class of methods provides. In this section, we will introduce a second approach to DA directly based on the Bayes' theorem: sequential Data Assimilation. It should be noted that even though variational DA algorithms are usually presented in the optimal control theory framework, they can also be derived with the Bayesian approach.

3.5.1 Introduction

From the beginning of times, scientists have tried to predict future events, whether it be in the field of astronomy, agriculture, or meteorology. With the discovery of the natural laws, there have been passionate debates on whether the laws of physics can or cannot provide perfect predictions (determinism vs indeterminism). In numerous applications, deterministic approaches can provide highly accurate forecasts. However,

sometimes our numerical models for forecasting events just don't work. The causes for this can be related to various reasons, among which are erroneous model assumptions, discretization techniques used, or imperfect knowledge of initial/boundary conditions (Asch et al., 2016).

The type of models used in fluid mechanics tend to be strongly non-linear. Direct numerical simulations of turbulence attempt to simulate *turbulence* itself, which has an inherent *unpredictable* nature. It is obvious that the forecasts that one can perform with CFD can be affected by the lack of accuracy in the initial/boundary conditions of the model. Moreover, CFD researchers and engineers are usually forced to resort to reduced-order modelling (e.g., turbulence modelling), artificially increasing thus the uncertainty in the forecast. For all these reasons, purely deterministic predictions are doomed to drift from reality. Accurate forecasts require formal consideration of the uncertainty present in the system. This can be done with a stochastic/probabilistic approach.

3.5.2 Bayes' Theorem

How to condition numerical forecasts? The best way to approach this problem is by taking a probabilistic approach. Luckily, an elegant mathematical framework can be used for that. Bayes' theorem is a mathematical formula for evaluating conditional probability and it is stated as following:

$$P\left(A|B\right) = \frac{P\left(B|A\right)P\left(A\right)}{P\left(B\right)}. \tag{3.13}$$

This formula provides an updated predicted probability or posterior (an updated knowledge of the system), $P\left(A|B\right)$, that combines the prior knowledge of the system, $P\left(A\right)$, with new observation, $P\left(B\right)$, given the conditional probability $P\left(B|A\right)$, which can be interpreted as the *likelihood*. This is at the core of what we are trying to do with DA. Our final objective is to improve the output of our numerical models by assimilating more and more evidence (observation) of the system as it becomes available.

3.5.3 Formulation of the estimation problem in the Bayesian framework

Now that we have a very promising candidate in the form of the Bayes' theorem for combining data and model, how do we formulate Bayes' rule in the context of fluid mechanics analysis? For that, we need a dynamical model (3.1) to describe the evolution in time of the system $\mathbf{x}_\mathrm{k}$. Available observation data, $\mathbf{y}_\mathrm{k}^\mathrm{o}$, can be defined using (3.3).

Model and observation data are affected by uncertainty. To reflect this, both sources of information are represented as random variables described in terms of probability density functions (PDFs). The objective is to determine the *real* state of a system by "fusing" model and observation. This combination procedure is also known as *state estimation*. From a Bayesian point of view, the estimator should provide the posterior distribution $P(\mathbf{x}|\mathbf{y}^\mathrm{o})$ of the random variable $\mathbf{x}$ conditioned on the observation $\mathbf{y}^\mathrm{o}$. This can be obtained by directly applying Bayes' theorem

$$P(\mathbf{x}|\mathbf{y}^\mathrm{o}\ ) = \frac{P(\mathbf{y}^\mathrm{o}|\mathbf{x})\,P(\mathbf{x})}{P(\mathbf{y}^\mathrm{o})}. \tag{3.14}$$

In (3.14), three key components are found: $P(\mathbf{x})$ is the PDF describing the most accurate knowledge of the system prior to assimilation of observation, $P(\mathbf{y}^\mathrm{o}|\mathbf{x})$ is the likelihood or joint probability of the observation conditioned on the true state of the system (the observation one would obtain if the true state of the system was known), and $P(\mathbf{y}^\mathrm{o})$ represents the marginal distribution or total probability of $\mathbf{y}^\mathrm{o}$

$$P(\mathbf{y}^\mathrm{o}) = \int_{-\infty}^{\infty} P(\mathbf{y}^\mathrm{o}|\mathbf{x})\,P(\mathbf{x})\,\mathrm{d}\mathbf{x}, \tag{3.15}$$

and acts as a normalization factor.

Let us now focus on the sequence in time of model and measurement errors during the time interval where DA is performed. Consider that the random error variables, $\boldsymbol{\eta}_\mathrm{k}$ (model (3.1)) and $\boldsymbol{\epsilon}_\mathrm{k}^\mathrm{o}$ (observation (3.3)), where $k = 1, ..., K$, are not correlated and their corresponding PDFs, P_η and P_ϵ , characterize their distribution. The different

terms in Bayes' equation (3.14) can be rewritten in function of these PDFs as follows:

$$P\left(\mathbf{x}_{\mathrm{k}}|\mathbf{x}_{\mathrm{k\text{-}1}}\right) = P_{\eta}\left[\mathbf{x}_{\mathrm{k}} - \mathcal{M}_{\mathrm{k:k\text{-}1}}\left(\mathbf{x}_{\mathrm{k\text{-}1}}\right)\right] \tag{3.16}$$

$$P\left(\mathbf{y}^{\mathrm{o}}_{\mathrm{k}}|\mathbf{x}_{\mathrm{k}}\right) = P_{\epsilon}\left[\mathbf{y}^{\mathrm{o}}{}_{\mathrm{k}} - \mathcal{H}_{\mathrm{k}}\left(\mathbf{x}_{\mathrm{k}}\right)\right] \tag{3.17}$$

Let us now evaluate the likelihood of the observation data within the observation window $[t_0, t_{\mathrm{K}}]$. This is calculated as a function of the sequences of states and observations, which can be defined as $\mathbf{x}_{\mathrm{K:0}} = \{\mathbf{x}_{\mathrm{K}}, \mathbf{x}_{\mathrm{K\text{-}1}}, ..., \mathbf{x}_0\}$ and $\mathbf{y}^{\mathrm{o}}_{\mathrm{K:1}} = \{\mathbf{y}^{\mathrm{o}}_{\mathrm{K}}, \mathbf{y}^{\mathrm{o}}_{\mathrm{K\text{-}1}}, ..., \mathbf{y}^{\mathrm{o}}_{1}\}$, respectively. Assuming uncorrelated observational errors in time, the mutual likelihood of all the measurements can be written as the product of the joint probabilities at each instant t_{k}

$$P\left(\mathbf{y}^{\mathrm{o}}{}_{\mathrm{K:1}}|\mathbf{x}_{\mathrm{K:0}}\right) = \prod_{\mathrm{k}=1}^{\mathrm{K}} P\left(\mathbf{y}^{\mathrm{o}}_{\mathrm{k}}|\mathbf{x}_{\mathrm{k}}\right) = \prod_{\mathrm{k}=1}^{\mathrm{K}} P_{\epsilon}\left[\mathbf{y}^{\mathrm{o}}_{\mathrm{k}} - \mathcal{H}_{\mathrm{k}}\left(\mathbf{x}_{\mathrm{k}}\right)\right]. \tag{3.18}$$

Furthermore, assuming a Markovian process, i.e., the probability of $\mathbf{x}_{\mathrm{k}}$ conditioned on the entire history of $\mathbf{x}$ only depends on the most recent known state $\mathbf{x}_{\mathrm{k\text{-}1}}$, the prior PDF can also be written as the product of the individual probabilities at each instant

$$P\left(\mathbf{x}_{\mathrm{K:0}}\right) = P\left(\mathbf{x}_0\right) \prod_{\mathrm{k}=1}^{\mathrm{K}} P\left(\mathbf{x}_{\mathrm{k}}|\mathbf{x}_{\mathrm{k\text{-}1}}\right) = P\left(\mathbf{x}_0\right) \prod_{\mathrm{k}=1}^{\mathrm{K}} P_{\eta}\left[\mathbf{x}_{\mathrm{k}} - \mathcal{M}_{\mathrm{k:k\text{-}1}}\left(\mathbf{x}_{\mathrm{k\text{-}1}}\right)\right]. \tag{3.19}$$

Finally, with Bayes's theorem (3.14) and combining (3.18) and (3.19) one can obtain the posterior distribution of the state as follows

$$P\left(\mathbf{x}_{\mathrm{K:0}}|\mathbf{y}^{\mathrm{o}}_{\mathrm{K:1}}\right) \propto P\left(\mathbf{x}_0\right) \prod_{\mathrm{k}=1}^{\mathrm{K}} P\left(\mathbf{y}^{\mathrm{o}}_{\mathrm{k}}|\mathbf{x}_{\mathrm{k}}\right) P\left(\mathbf{x}_{\mathrm{k}}|\mathbf{x}_{\mathrm{k\text{-}1}}\right) =$$

$$P\left(\mathbf{x}_0\right) \prod_{\mathrm{k}=1}^{\mathrm{K}} P_{\epsilon}\left[\mathbf{y}^{\mathrm{o}}_{\mathrm{k}} - \mathcal{H}_{\mathrm{k}}\left(\mathbf{x}_{\mathrm{k}}\right)\right] P_{\eta}\left[\mathbf{x}_{\mathrm{k}} - \mathcal{M}_{\mathrm{k:k\text{-}1}}\left(\mathbf{x}_{\mathrm{k\text{-}1}}\right)\right]. \tag{3.20}$$

This equation represents a hidden Markov chain. Essentially, this means that the current probabilistic state of the system only depends on the most recently known state. Therefore, new data can be integrated through Bayes' rule as soon as it becomes

available. This very powerful result comes from the strong assumptions that were considered concerning model and observation errors. These errors were assumed to be uncorrelated in time, which is advantageous from a mathematical point of view but, seldom valid in realistic applications. There is, however, a notable exception; high-Reynolds turbulent flows show eminent Markovian characteristics (Pedrizzetti & Novikov, 1994).

One can differentiate three different state-estimation problems depending on the time interval where this estimation is actually performed (see Wiener, 1949 and Carrassi et al., 2018 for a comprehensive review):

1. Prediction problem: In this type of problems, one estimates $P\left(\mathbf{x}_{\mathrm{w}}|\mathbf{y}^{\mathrm{o}}_{\mathrm{k:1}}\right)$ with $\mathrm{w} > k$. Basically, one can rely on the model to predict the probability of future phenomena from the last available posterior distribution $P\left(\mathbf{x}_{\mathrm{k}}|\mathbf{y}^{\mathrm{o}}_{\mathrm{k:1}}\right)$. This problem is formally addressed by integrating (3.16) (Chapman-Kolmogorov equation)

$$P\left(\mathbf{x}_{\mathrm{w}}|\mathbf{y}^{\mathrm{o}}_{\mathrm{k:1}}\right) = \int_{-\infty}^{\infty} P_{\eta}\left[\mathbf{x}_{\mathrm{w}} - \mathcal{M}_{\mathrm{w:k}}\left(\mathbf{x}_{\mathrm{k}}\right)\right] P\left(\mathbf{x}_{\mathrm{k}}|\mathbf{y}^{\mathrm{o}}_{\mathrm{k:1}}\right) \mathrm{d}\mathbf{x}_{\mathrm{k}} \tag{3.21}$$

2. Filtering problem: The objective is now the estimation of the conditional distribution of the state $P\left(\mathbf{x}_{\mathrm{k}}|\mathbf{y}^{\mathrm{o}}_{\mathrm{k:1}}\right)$ at a given time t_{k}. In this type of problems, new data is integrated sequentially as soon as it becomes available (Jazwinski, 1970). Solution to filtering problems involves a two-step procedure where at first, a given conditional PDF is forwarded in time until a new observation is provided. At this specific point, an analysis step based on Bayes' theorem is performed to update the state

$$P\left(\mathbf{x}_{\mathrm{k}}|\mathbf{y}^{\mathrm{o}}_{\mathrm{k:1}}\right) = \frac{P_{\epsilon}\left[\mathbf{y}^{\mathrm{o}}_{\mathrm{k}} - \mathcal{H}_{\mathrm{k}}\left(\mathbf{x}_{\mathrm{k}}\right)\right] P\left(\mathbf{x}_{\mathrm{k}}|\mathbf{y}^{\mathrm{o}}_{\mathrm{k\text{-}1:1}}\right)}{\int_{-\infty}^{\infty} P_{\epsilon}\left[\mathbf{y}^{\mathrm{o}}_{\mathrm{k}} - \mathcal{H}_{\mathrm{k}}\left(\mathbf{x}_{\mathrm{k}}\right)\right] P\left(\mathbf{x}_{\mathrm{k}}|\mathbf{y}^{\mathrm{o}}_{\mathrm{k\text{-}1:1}}\right) \mathrm{d}\mathbf{x}_{\mathrm{k}}}. \tag{3.22}$$

   This procedure is repeated in time alternating forward and analysis phases, sequentially updating the propagated PDFs with available data.

3. Smoothing problem: In this type of state-estimation problems, the target condi-

tional PDF at a given time t_k is based on past, present and future observations. Clearly, this type of analysis provides more accurate estimates when compared to the estimates one would obtain through filtering. Smoothed estimates can be obtained via a recursive forward/backward application of the Bayes's rule together with the propagation of the PDFs with the Chapman-Kolmogorov equation.

3.5.4 Gaussian distributions

We have formulated the estimation problem in the Bayesian framework and defined three different types of estimation problems. The expressions obtained for the conditional PDFs allow a complete probabilistic description of all the possible states of a system. However, in practice it is hard to evaluate the complete probability densities. Therefore, it is more judicious to target particular characteristics of these PDFs. The statistical DA problem is thus formulated as an optimization involving the minimization of the uncertainty in the posterior with observation data:

1. Conditional mode estimation: These type of methods target the mode of the distribution, i.e., the most probable state, and are also referred to as maximum a posteriori (MAP) estimators.

2. Conditional mean estimation: Contrary to conditional mode estimation methods, in these type of estimators, the target is the mean of the distribution, and since the mean always provides the minimum squared error estimate (Jazwinski, 1970), they are usually called minimum squared error estimators.

Bayes' rule provides an elegant and logical framework for combining model and observation. However, the extremely large dimensions of the state vectors we usually deal with in fluid mechanics impedes the usage of the Bayesian approach in its entirety. Notably, it is extremely difficult to advance the PDFs in time. One way to reduce the computational burden is to assume that model and observation uncertainties follow Gaussian distributions. This strong hypothesis leads to a very convenient

mathematical simplification: Gaussian PDFs are fully described by their first and second statistical moments, i.e., by the mean of the distribution and a covariance matrix.

Let us formalize the definition of the state in the Gaussian framework. Consider the state of the system as a random vector $\mathbf{x} \in \mathbb{R}^n$ with a Gaussian distribution, whose joint PDF can be written as

$$P_{\mathbf{x}}(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right). \tag{3.23}$$

The mean $\mu \in \mathbb{R}^n$, which is defined as

$$\mu = \mathbb{E}(\mathbf{x}), \tag{3.24}$$

and the covariance matrix $\mathbf{\Sigma}$, an $n \times n$ symmetric and positive definite matrix

$$\mathbf{\Sigma} = \mathrm{cov}(\mathbf{x}), \tag{3.25}$$

completely define the Gaussian distribution of the random state variable

$$\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Sigma}). \tag{3.26}$$

3.5.5 Sequential DA Filters

We have seen that, when DA is addressed from a statistical point of view, we are naturally drawn towards the Bayesian framework which allows to rigorously combine the two sources of information available and provide a posterior probability law. One could use particle filters (Doucet & Johansen, 2011); however, it is computationally infeasible for most realistic applications to estimate the full PDF distribution. Therefore, one has to introduce a strong hypothesis that allows to simplify the estimation of the PDF problem to the estimation of its first statistical moments: all the PDFs that intervene in the estimation problem are considered to be Gaussian distributed.

In this context, the estimation problem is reduced to the determination of the mean and the variance of the posterior state. There exist proven and mature statistical methods for determining the optimal estimate (Anderson and Moore, 1979, Garthwaite et al., 2002, Ross, 2014). Conditional mean and conditional mode estimation methods have been introduced above. Conditional mean or minimum variance (MV) estimation naturally leads to the well-known Kalman Filter, while conditional mode or maximum a posteriori (MAP) estimators are intrinsically related to 3D-VAR methods. This draws an elegant connection between optimal control based variational methods and statistical sequential methods.

In the previous section, it was shown that variational methods are based on the minimization of a cost function that expresses the difference between model and observations within a given assimilation time window. In other words, multiple observations are integrated at the same time and the optimization procedure accounts for all the data at once in a numerical process that is strongly stationary. This is interesting from a reanalysis point of view (similar to the previously introduced smoothing problem) but, certain phenomena encountered in fluid mechanics present strong unstationary and non-linear features. In this type of scenario, it would be highly beneficial to be able to just update the previous estimate as new observations become available, without having to consider all the previous history of the state. This is in the core of the sequential Kalman filter (KF), which constitutes the main foundation of this PhD thesis.

3.5.5.1 Kalman Filter

In 1960, Rudolph E. Kalman published a paper introducing a new approach to linear filtering and prediction problems (Kalman, 1960). In this work, a new solution method to the filtering problem is described, which will later be known as the Kalman Filter (KF). The derivation of the method is not given here for simplicity. It is based on the minimization of the variance in the posterior distribution. The interested reader can find numerous works treating this issue (Jazwinski, 1970, Talagrand, 1997, Bain and Crisan, 2009, Asch et al., 2016).

The classical KF provides an estimate of the state of a physical system at time k ($\mathbf{x}_k$), given the initial estimate $\mathbf{x}_0$, a set of observations, $\mathbf{y}_k^o$, that can be described using (3.4), and the information of a linear dynamical model (3.2). Following the notation generally used in DA literature, the *forecast/analysis* states and error covariances are indicated as $\mathbf{x}_k^{f/a}$ and $\mathbf{P}_k^{f/a}$, respectively.

We recall that the notation used is introduced in Sec. 3.3. The estimated state is obtained via a recursive procedure:

1. A predictor (forecast) phase, where the analysed state of the system at a previous time-step is used to obtain an *a priori* estimation of the state at the current instant. This prediction, which is obtained relying on the model only, is not conditioned by observation at time k:

$$\mathbf{x}_k^f = \mathbf{M}_{k:k-1}\mathbf{x}_{k-1}^a \tag{3.27}$$

$$\mathbf{P}_k^f = \mathbf{M}_{k:k-1}\mathbf{P}_{k-1}^a\mathbf{M}_{k:k-1}^\top + \mathbf{Q}_k \tag{3.28}$$

2. An update (analysis) step, where the state-estimation is updated accounting for observation at the time k:

$$\mathbf{K}_k = \mathbf{P}_k^f\mathbf{H}_k^\top\left(\mathbf{H}_k\mathbf{P}_k^f\mathbf{H}_k^\top + \mathbf{R}_k\right)^{-1} \tag{3.29}$$

$$\mathbf{x}_k^a = \mathbf{x}_k^f + \mathbf{K}_k\left(\mathbf{y}_k^o - \mathbf{H}_k\mathbf{x}_k^f\right) \tag{3.30}$$

$$\mathbf{P}_k^a = \left(I - \mathbf{K}_k\mathbf{H}_k\right)\mathbf{P}_k^f \tag{3.31}$$

A complete filtering phase from an instant t_{k-1} to an instant t_k is illustrated in Fig. 3-3. The optimal prediction of the state ($\mathbf{x}_k^a$) is obtained via the addition to the predictor estimation ($\mathbf{x}_k^f$) of a correction term determined via the so-called *Kalman gain* $\mathbf{K}_k$. The classical KF algorithm is not suited for direct application to the analysis of complex flows, since the classical KF formulation is developed for linear systems. Moreover, the knowledge of the error covariances $\mathbf{Q}_k$ and $\mathbf{R}_k$ is usually limited. One simple, classical simplification is to consider that errors for each component are completely uncorrelated in space and from other components, i.e. $\mathbf{Q}_k$ and $\mathbf{R}_k$ are

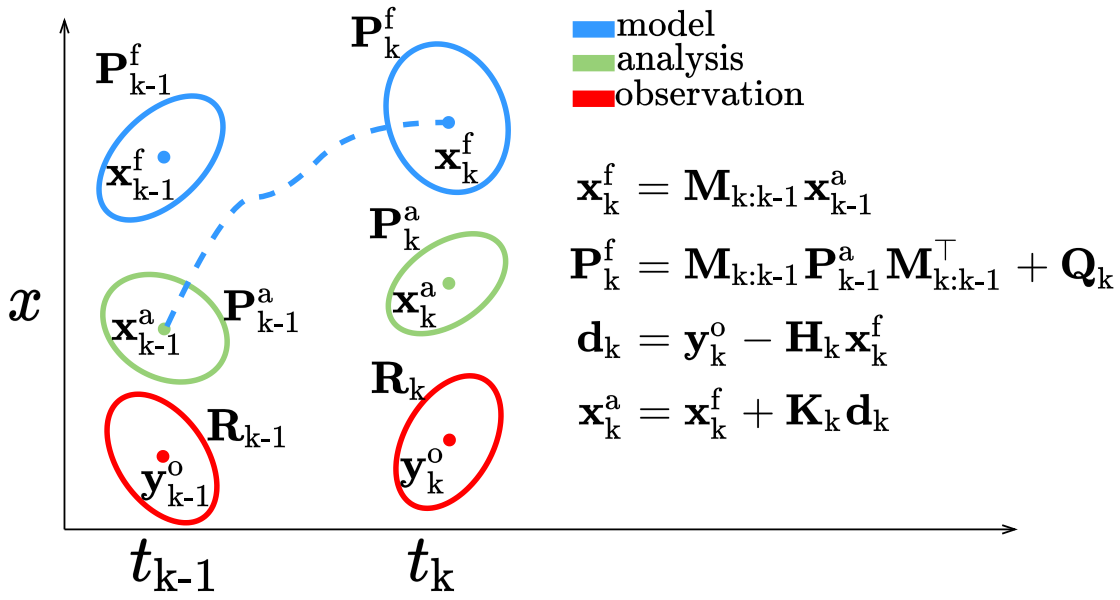


Figure 3-3: Schematic representation of the classical Kalman Filter.

considered to be diagonal (Brunton and Noack, 2015a; Stonebridge, 2017).

Furthermore, KF relies on the transport of a very large error covariance matrix $\mathbf{P}_k$. Matrix inversion operations (see (3.29)) are also required. Depending on the size of the observation vector, this might be computationally intractable.

3.5.5.2 Ensemble Kalman Filter

In practical fluid mechanics applications, we are usually confronted with strongly non-linear models and depending on the type of experimental observation available, we might also have to deal with non-linear measurement operators. In this context, we have to consider the full non-linear form of the equations (3.1) and (3.3), with the non-linear operators $\mathcal{H}$ and $\mathcal{M}$, respectively. The classical KF is not tailored for this type of situations. Extensions to the KF capable of dealing with these non-linearities have been developed, such as the Extended Kalman Filter (EKF)(Jazwinski, 1970), where the mean and the covariance are linearized about the current state. However, this method is hardly affordable in high dimensional systems and is accurate only when the degree of non-linearity remains low (Asch et al., 2016).

An alternative approach to deal with non-linearities and even non-Gaussian statis-

tics is the Ensemble Kalman Filter (EnKF), where the PDF of the state is described by an ensemble of possible model realizations. This method is extremely easy to implement and very intuitive. Consider a forecast with uncertainty on the initial conditions. If one considers multiple statistically probable initial states, each one of them will produce a slightly different trajectory as they are evolved in time with the model. Thereafter, the variance in the ensemble can be used to approximate error statistics.

The Ensemble Kalman Filter (EnKF) (Evensen, 1994, 2009) relies on the estimation of $\mathbf{P}_\mathrm{k}$ by means of an ensemble. More precisely, the error covariance matrix is approximated using a finite ensemble of model states of size N_e. If the ensemble members are generated using stochastic Monte-Carlo sampling, the error in the approximation decreases with a rate of $\frac{1}{\sqrt{N_\mathrm{e}}}$. We provide here a general description of the method. The interested reader should know that there are a number of countermeasures that one must put in place (localization, inflation) to make this method work in realistic cases, and should find detailed description of such measures in the works of Carrassi et al., 2018 and Asch et al., 2016.

Given an ensemble of forecast/analysed states at a certain instant k, the ensemble matrix is defined as

$$\boldsymbol{\mathscr{E}}_\mathrm{k}^{\mathrm{f/a}} = \left[\mathbf{x}_\mathrm{k}^{\mathrm{f/a},(1)}, \cdots, \mathbf{x}_\mathrm{k}^{\mathrm{f/a},(N_\mathrm{e})}\right] \in \mathbb{R}^{N_x \times N_\mathrm{e}}. \tag{3.32}$$

To reduce the numerical cost of implementation, the normalised ensemble anomaly matrix is then specified as

$$\mathbf{X}_\mathrm{k}^{\mathrm{f/a}} = \frac{\left[\mathbf{x}_\mathrm{k}^{\mathrm{f/a},(1)} - \overline{\mathbf{x}_\mathrm{k}^{\mathrm{f/a}}}, \cdots, \mathbf{x}_\mathrm{k}^{\mathrm{f/a},(N_\mathrm{e})} - \overline{\mathbf{x}_\mathrm{k}^{\mathrm{f/a}}}\right]}{\sqrt{N_\mathrm{e}-1}} \in \mathbb{R}^{N_x \times N_\mathrm{e}}, \tag{3.33}$$

where the ensemble mean $\overline{\mathbf{x}_\mathrm{k}^{\mathrm{f/a}}}$ is obtained as

$$\overline{\mathbf{x}_\mathrm{k}^{\mathrm{f/a}}} = \frac{1}{N_\mathrm{e}} \sum_{i=1}^{N_\mathrm{e}} \mathbf{x}_\mathrm{k}^{\mathrm{f/a},(\mathrm{i})}. \tag{3.34}$$

The error covariance matrix $\mathbf{P}_{\mathrm{k}}^{\mathrm{f/a}}$ can thus be estimated via the information derived from the ensemble. This estimation, hereafter denoted with the superscript e, can be factorized into:

$$\mathbf{P}_{\mathrm{k}}^{\mathrm{f/a,e}} = \mathbf{X}_{\mathrm{k}}^{\mathrm{f/a}} \left(\mathbf{X}_{\mathrm{k}}^{\mathrm{f/a}}\right)^{\top} \in \mathbb{R}^{N_x \times N_x}. \tag{3.35}$$

The goal of the EnKF is to mimic the BLUE (Best Linear Unbiased Estimator) analysis of the Kalman filter. For this, Burgers et al., 1998 showed that the observation must be considered as a random variable with an average corresponding to the observed value and a covariance $\mathbf{R}_{\mathrm{k}}$ (the so-called *data randomization* trick). Therefore, given the discrete observation vector $\mathbf{y}_{\mathrm{k}}^{\mathrm{o}} \in \mathbb{R}^{N_y}$ at an instant k, the ensemble of perturbed observations is defined as

$$\mathbf{y}_{\mathrm{k}}^{\mathrm{o,(i)}} = \mathbf{y}_{\mathrm{k}}^{\mathrm{o}} + \boldsymbol{\epsilon}_{\mathrm{k}}^{\mathrm{o,(i)}}, \quad \text{with} \quad i = 1, \cdots, N_{\mathrm{e}} \quad \text{and} \quad \boldsymbol{\epsilon}_{\mathrm{k}}^{\mathrm{o,(i)}} \sim \mathcal{N}(0, \mathbf{R}_{\mathrm{k}}). \tag{3.36}$$

A normalized anomaly matrix of the observations errors is defined as

$$\mathbf{E}_{\mathrm{k}}^{\mathrm{o}} = \frac{1}{\sqrt{N_{\mathrm{e}} - 1}} \left[\boldsymbol{\epsilon}_{\mathrm{k}}^{\mathrm{o,(1)}} - \overline{\boldsymbol{\epsilon}^{\mathrm{o}}}, \boldsymbol{\epsilon}_{\mathrm{k}}^{\mathrm{o,(2)}} - \overline{\boldsymbol{\epsilon}^{\mathrm{o}}}, \cdots, \boldsymbol{\epsilon}_{\mathrm{k}}^{\mathrm{o,(}N_{\mathrm{e}}\mathrm{)}} - \overline{\boldsymbol{\epsilon}^{\mathrm{o}}}, \right] \in \mathbb{R}^{N_y \times N_{\mathrm{e}}}, \tag{3.37}$$

where $\overline{\boldsymbol{\epsilon}^{\mathrm{o}}} = \dfrac{1}{N_{\mathrm{e}}} \displaystyle\sum_{i=1}^{N_{\mathrm{e}}} \boldsymbol{\epsilon}_{\mathrm{k}}^{\mathrm{o,(i)}}$.

The covariance matrix of the measurement error can then be estimated as

$$\mathbf{R}_{\mathrm{k}}^{\mathrm{e}} = \mathbf{E}_{\mathrm{k}}^{\mathrm{o}} \left(\mathbf{E}_{\mathrm{k}}^{\mathrm{o}}\right)^{\top} \in \mathbb{R}^{N_y \times N_y}. \tag{3.38}$$

By combining the previous results, we obtain (see Asch et al., 2016) the standard stochastic EnKF algorithm. The corresponding analysis step consists of updates performed on each of the ensemble members, as given by

$$\mathbf{x}_{\mathrm{k}}^{\mathrm{a,(i)}} = \mathbf{x}_{\mathrm{k}}^{\mathrm{f,(i)}} + \mathbf{K}_{\mathrm{k}}^{\mathrm{e}} \left(\mathbf{y}_{\mathrm{k}}^{\mathrm{o,(i)}} - \mathcal{H}_{\mathrm{k}} \left(\mathbf{x}_{\mathrm{k}}^{\mathrm{f,(i)}}\right)\right), \tag{3.39}$$

where the Kalman Gain can be calculated as

$$\mathbf{K}_k^e = \mathbf{X}_k^f \left(\mathbf{Y}_k^f\right)^\top \left(\mathbf{Y}_k^f \left(\mathbf{Y}_k^f\right)^\top + \mathbf{E}_k^o \left(\mathbf{E}_k^o\right)^\top\right)^{-1}, \tag{3.40}$$

and where $\mathbf{Y}_k^f = \mathbf{H}_k \mathbf{X}_k^f$.

A schematic representation of the functioning of the EnKF is given in Fig. 3-4. A version of the Ensemble Kalman filter algorithm using the previously defined anomaly matrices is given in Sec. A.2. This is the version we use in our applications.

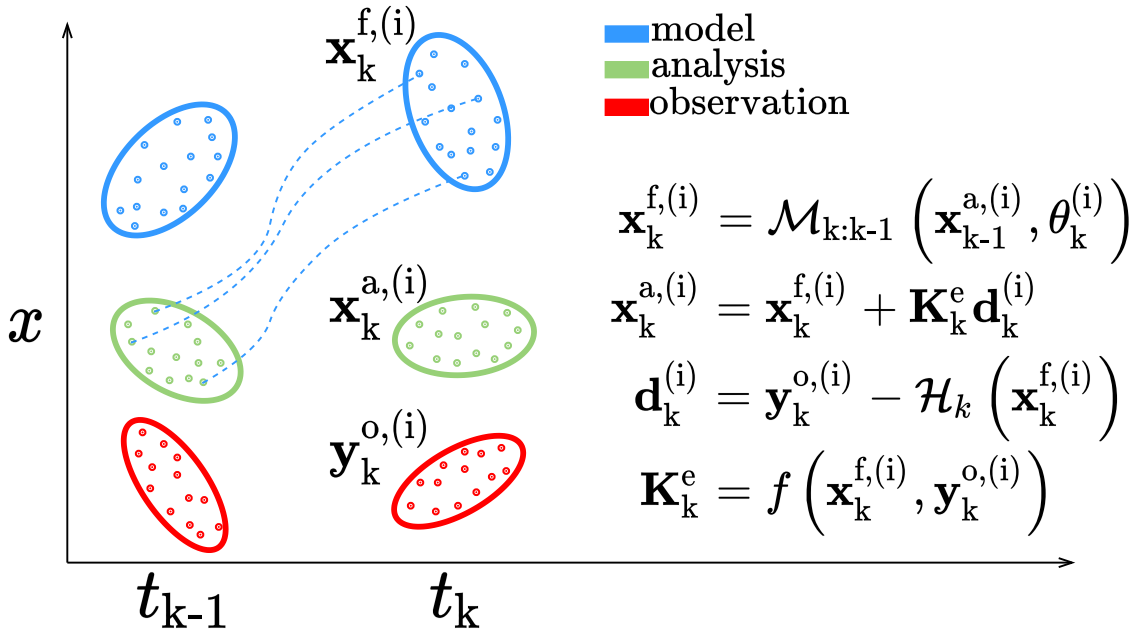


Figure 3-4: Schematic representation of the Ensemble Kalman Filter (EnKF).

State-of-the-art approaches based on the EnKF are arguably the most advanced forms of state-estimation available in the field of DA methods. These techniques have been extensively applied in the last decade in meteorology and geoscience (see Asch et al., 2016 and references therein).

3.5.5.3 Dual Ensemble Kalman filter

In this section, we extend the classical EnKF framework presented in Sec. 3.5.5.2 by considering the case of a parametrized model such as (3.1). The objective is to enable the model to generate accurate forecasts. For this, we need to determine good

estimates of both model state variables, $\mathbf{x}_{\mathrm{k}}$, and parameters, $\boldsymbol{\theta}_{\mathrm{k}}$, given noisy observations $\mathbf{y}_{\mathrm{k}}^{\mathrm{o}}$. One approach is provided by *joint estimation*, where state and parameter vectors are concatenated into a single joint state vector (state augmentation). After Moradkhani et al., 2005, the drawback of such strategy is that, by increasing the number of unknown model states and parameters, the degree of freedom in the system increases and makes the estimation unstable and intractable, especially in the non-linear dynamical model. An alternative approach is provided by Moradkhani et al., 2005, called *dual estimation*. Two interactive filters are used, the former for the estimation of the parameters from a guessed state solution, the latter for the update of the state variables from the estimated previous parameters.

In the first step of the algorithm, the ensemble of the analysed parameters is updated following the classical KF equation:

$$\boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{a,(i)}} = \boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{f,(i)}} + \mathbf{K}_{\mathrm{k}}^{\theta,\mathrm{e}} \left(\mathbf{y}_{\mathrm{k}}^{\mathrm{o,(i)}} - \mathbf{y}_{\mathrm{k}}^{\mathrm{f,(i)}} \right) \quad \text{with} \quad i = 1, \cdots, N_e \tag{3.41}$$

where $\mathbf{y}_{\mathrm{k}}^{\mathrm{f,(i)}} = \mathcal{H}_{\mathrm{k}} \left(\mathbf{x}_{\mathrm{k}}^{\mathrm{f,(i)}} \right)$.

The Kalman gain responsible for correcting the parameter trajectories in the ensemble is obtained as follows:

$$\mathbf{K}_{\mathrm{k}}^{\theta,\mathrm{e}} = \boldsymbol{\Theta}_{\mathrm{k}}^{\mathrm{f}} \left(\mathbf{Y}_{\mathrm{k}}^{\mathrm{f}} \right)^{\top} \left(\mathbf{Y}_{\mathrm{k}}^{\mathrm{f}} \left(\mathbf{Y}_{\mathrm{k}}^{\mathrm{f}} \right)^{\top} + \mathbf{E}_{\mathrm{k}}^{\mathrm{o}} \left(\mathbf{E}_{\mathrm{k}}^{\mathrm{o}} \right)^{\top} \right)^{-1}, \tag{3.42}$$

where the matrix $\boldsymbol{\Theta}_{\mathrm{k}}^{\mathrm{f}}$ plays the same role for the parameters as the matrix $\mathbf{X}_{\mathrm{k}}^{\mathrm{f}}$ defined in (3.33) for the states. We then have:

$$\boldsymbol{\Theta}_{\mathrm{k}}^{\mathrm{f/a}} = \frac{\left[\boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{f/a,(1)}} - \overline{\boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{f/a}}}, \cdots, \boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{f/a},(N_{\mathrm{e}})} - \overline{\boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{f/a}}} \right]}{\sqrt{N_{\mathrm{e}} - 1}} \in \mathbb{R}^{N_\theta \times N_{\mathrm{e}}} \tag{3.43}$$

with

$$\overline{\boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{f/a}}} = \frac{1}{N_{\mathrm{e}}} \sum_{i=1}^{N_{\mathrm{e}}} \boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{f/a,(i)}} \tag{3.44}$$

Once new values of the model's parameters are inferred, we can update the state by EnKF (see Sec. 3.5.5.2). The Dual Ensemble Kalman filter allows performing a

recursive parametric inference/state-estimation using the information from the ensemble members. The algorithm that we use is given in Sec. A.3. The Dual EnKF was recently employed to combine data assimilation with machine learning to obtain data-driven flow models (Kumar, 2021).

3.5.6 Applications of Sequential DA algorithms for fluid mechanics problems: a critical review

The KF algorithm has been introduced and special emphasis has been placed on its characteristics. We have seen that the canonical form of the KF is seldom used in realistic flow problems due to its high computational cost. Moreover, the KF is tailored for linear systems. Extensions adapted to non-linear applications exist, e.g., the extended Kalman Filter, but their computational cost remains too high.

On the other hand, ensemble Kalman filters offer an interesting alternative to the classical filters. They are particularly well suited for non-linear multiscale systems. In this version of the Kalman filter, the error statistics are no longer propagated via a covariance matrix, but instead approximated through several perturbed trajectories. In other words, the error in the estimate is obtained from a finite ensemble of possible model realizations. On many problems, an ensemble size of roughly 100 members is enough to accurately capture error propagation characteristics, which makes this estimation strategy extremely tractable for high-dimensional problems. Below, we review some applications of the Kalman Filter (and its multiple versions) in the analysis of flows.

$\hookrightarrow$***Flow control***

In a flow control problem, having accurate information about the state of the flow is essential for designing a successful control strategy. Bewley and Liu, 1998 and Bewley et al., 2001 explored the application of modern linear control theory to fluid mechanics problems. They showed that estimator-based control strategies outperform classical (proportional) control approaches. Högberg et al., 2003 investigated the

ability of constant-gain linear control strategies to delay transition to turbulence in channel flows. They introduced a state estimator in the form of an extended Kalman Filter which was combined with the controller to obtain a wall-information-based linear compensator. The fundamental idea behind their approach was to express the incompressible Navier-Stokes equations as a dynamical system where linear state-space control theory can be applied.

Linear model-feedback control problems based on limited and noisy measurements involve two independent sub-problems: (i) the state feedback control problem; (ii) the state-estimation problem. Hœpffner et al., 2005 focused on developing effective state-estimation strategies for flow control. Starting from the linearized version of the Navier-Stokes equations, they used the Kalman filter and its extended version to estimate the state of a laminar wall-bounded flow from noisy skin-friction and pressure measurements. Building on this research work, Chevalier et al., 2006 extended the estimation method to the turbulent channel flow problem, with the objective to develop an estimator capable of providing an accurate complete representation of the flow from wall measurements. Finally, Colburn et al., 2011 investigated the use of the ensemble version of the Kalman filter (EnKF) for state-estimation purposes.

$\hookrightarrow$***Hybrid Simulation***

In the 2000s, there was an interest in the development of *hybrid simulation* techniques (see Nisugi et al., 2004; Suzuki, Ji, et al., 2009, 2010; Suzuki, Sanse, et al., 2009 for example), which attempted to combine wind-tunnel data (PIV) with CFD simulation. In short, experimental and CFD data were *fused* through a weighted averaged between the two sources of information. The capability of this technique was demonstrated for low-Reynolds flows but, its applicability to high-Reynolds flows was shown to be problematic due to the lack of temporal and spatial resolution of the PIV and the noise in the observation. The combination strategies used in the early hybrid simulation efforts are not so far from the actual framework of the KF. The only difference lies on the calculation of the weights, which in the first case is obtained heuristically, while in the second case is calculated in a rigorous probabilistic

fashion, accounting for the stochastic nature of both sources of information. Suzuki, 2012 explored the possibility of using the KF to calculate the optimal weight. The authors were inspired by the state-estimation framework introduced by Bewley's group in flow control but, adopted a reduced-order strategy to reduce the costs related to the forecasting of the error covariance matrix.

$\hookrightarrow$***Parameter estimation***

Other authors explored the parameter estimation capabilities of the EnKF. Kato and Obayashi, 2013, used this method to optimize the coefficients of the Spalart-Allmaras turbulence model in the case of a flow over a flat plate. In a later study, Kato et al., 2015 used the EnKF to estimate Mach number, angle of attack and turbulent viscosity wind-tunnel corrections for transonic flows around different airfoils.

In recent years, some research groups have started to tackle turbulence modelling optimization through DA. Xiao et al., 2016 introduced a novel Bayesian framework for uncertainty quantification and reduction in RANS simulation. More specifically, a compact form of the Reynolds stress tensor is adopted, achieving a parametrized reduced-order representation of the term. Uncertainty is thereafter injected in the coefficients of the reduced-order model, and an iterative ensemble Kalman filter (IEnKF) is used to update these coefficients with available observation. The approach adopted is completely non-intrusive in the sense that, the uncertainty in the Reynolds Stress tensor is propagated to the state of the system through the model. However, no actual direct correction of the state is performed. The uncertainty of the global system can then be tracked solely through the distributions of the reduced-order model coefficients and their impact on the state.

$\hookrightarrow$***Adding additional physical constraints to the filtering procedure: Regularized EnKF***

The application of the KF techniques to inverse problems is usually not an easy task due to the issue of ill-posedness. In variational methods, this can be solved straightforwardly by imposing additional constraints to the cost function. This is

more problematic for KF-based methods. X.-L. Zhang et al., 2020 derived a version of the ensemble Kalman filter where additional constraints (regularization) can be embedded in the calculation of the Kalman update. An improvement in the capabilities of the estimator to infer parameters was observed when the regularized version of the estimator was compared to the traditional one. Later, X.-L. Zhang et al., 2021 explored the capabilities of the regularized ensemble Kalman filter to reconstruct turbulent mean flows from disparate data.

↪***Kalman Filter applications to unsteady 3D flows***

Sequential filtering applications to 3D flows are scarce due to the extremely high computational costs. Most of the research efforts put into sequential data assimilation methods are focused on developing strategies that allow to reduce the cost of the estimator. In this sense, Meldi and Poux, 2017 developed a sequential assimilation strategy integrating a reduced-order KF in the PISO algorithm of a segregated solver for incompressible flows. This technique allows to sequentially estimate the state of the flow accounting for the prescribed level of confidence in the model and the experimental data. Flow velocities are directly corrected with the Kalman filter. However, the authors exploit the intrinsic structure of the PISO algorithm and integrate the Kalman update just before the resolution of the Poisson equation. Local corrections are then propagated everywhere in the domain through instantaneous redistribution of pressure. With these reduced-order approach, the authors managed to achieve state-estimation of a 3D mixing-layer flow. Later, Meldi, 2018, performed a detailed analysis of the sensitivity of the performance of the previously developed reduced-order KF estimator to the time density of available observation. The results obtained showed that the minimum update frequency necessary to reach state-estimation convergence is related to the characteristic advection time.

Other EnKF applications in related fields such as combustion and wildfire propagation have also been explored (Labahn et al., 2019; Rochoux et al., 2015). The usage of the EnKF in the geosciences is vast, but we do not treat the applications in this domain in detail. The interested reader is referred to Carrassi et al., 2018;

Houtekamer and Zhang, 2016 for a complete overview on the subject.

$\hookrightarrow$***Conclusion***

Sequential assimilation methods are far less explored than variational strategies in the fluid mechanics community. The computational power required for the propagation of the uncertainty can easily limit their application for complex flow problems. However, sequential methods are much easier to implement than their variational equivalents. They also allow for assessing the uncertainty in the estimate, which is crucial in any attempt to predict future states. The ensemble version of the KF is clearly the most appropriate approach for practical applications, since it allows the analysis of non-linear systems and can be computationally tractable for very large problems. However, there is still a long way to go before achieving fully operational ensemble-based estimators:

- As seen in the literature review, filtering applications to 3D flow are extremely scarce. EnKF approaches are less computationally expensive than classical KF ones, but sample-size associated costs can still be limiting. EnKF applications to high-fidelity simulations (unsteady phenomena, for example) are nowadays untreatable.

- The Kalman state update can be problematic in terms of conservativity (respect of the Navier-Stokes equation) and smoothness of the solution. When observation is integrated in a system through the Kalman filter, the estimate usually does not comply with the original equations. In the context of high-fidelity Navier-Stokes simulation, this can produce spurious effects that might destabilize the numerical solution.

  Sequential data assimilation methods are particularly well-appropriated for unsteady flow phenomena. The only needed information to update the state of a flow with new available information is its previous analysis state (the previous instant where observation was used to update the state). This constitutes a significant advantage over variational approaches from an operational and storage

point of view.

For the reasons mentioned above, this research work is focused entirely on the development of sequential data assimilation methods for unsteady flow. More specifically, a new version of the ensemble Kalman filter is presented, where we tackle computational cost issues and smoothness of the solution.

Chapter 4

Multigrid Ensemble Kalman Filter

Despite the game-changing advantage that EnKF offers for the analysis of large-scale dynamical systems, the use of a sufficiently large ensemble (usually 60 to 100 members are required for convergence (Asch et al., 2016)) may still be prohibitive for advanced applications. In the following, we present an EnKF strategy which relies on the generation of the ensemble members on a subspace (*i.e.*, coarser mesh) of the original model. To do so, we exploit the multiple levels of resolution naturally used by the multigrid procedure (introduced in Sec. 2.6.1) for the time advancement of the flow. For the classical case of the FAS two-grid multigrid algorithm, which employs two levels of resolution (coarse and fine), the ensemble members calculated on the coarse mesh level are run with a single high-refinement simulation, which is updated using the coarse mesh assimilation results. For this reason, the computational costs and the memory storage of the physical variables are dramatically reduced. For sake of simplicity, the procedure is here detailed just for the FAS two-grid multigrid algorithm.

The developed estimation algorithm combines the classical EnKF with features of the multigrid strategy. It is thus called the Multigrid Ensemble Kalman Filter (MGEnKF). A first simplified sketch of the algorithm is presented in Fig. 4-1, where the different operations run on the fine grid and the coarse grid are illustrated. There are two distinct procedures that are run on the coarse grid of the MGEnKF algorithm: an *outer loop*, where the observations used are obtained from an external source of

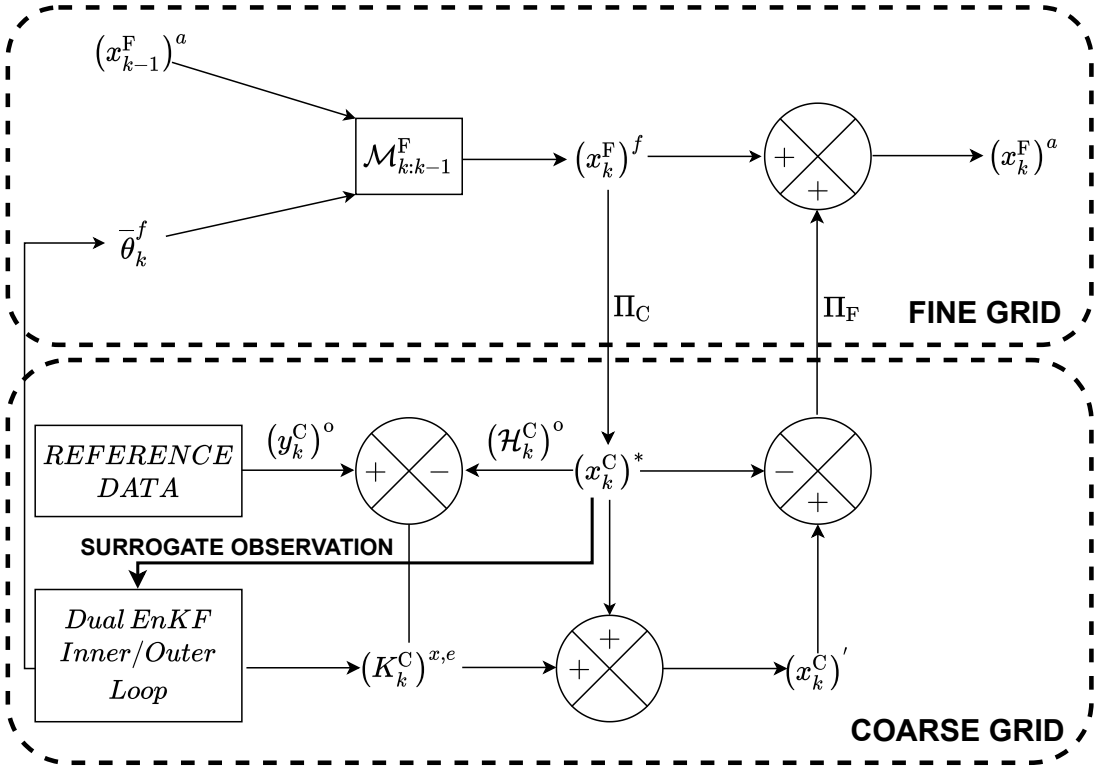


Figure 4-1: Schematic representation of the Multigrid Ensemble Kalman Filter (MGEnKF). Two different levels of representation (fine and coarse grids) are used to obtain an estimation for the main simulation running on the fine grid. The inner loop and outer loop procedures (see below for description) are performed on the coarse grid.

information, and an *inner loop*, where fine grid solutions projected onto the coarse grid are used as surrogate observations. The complete general algorithm for MGEnKF is structured in the following operations:

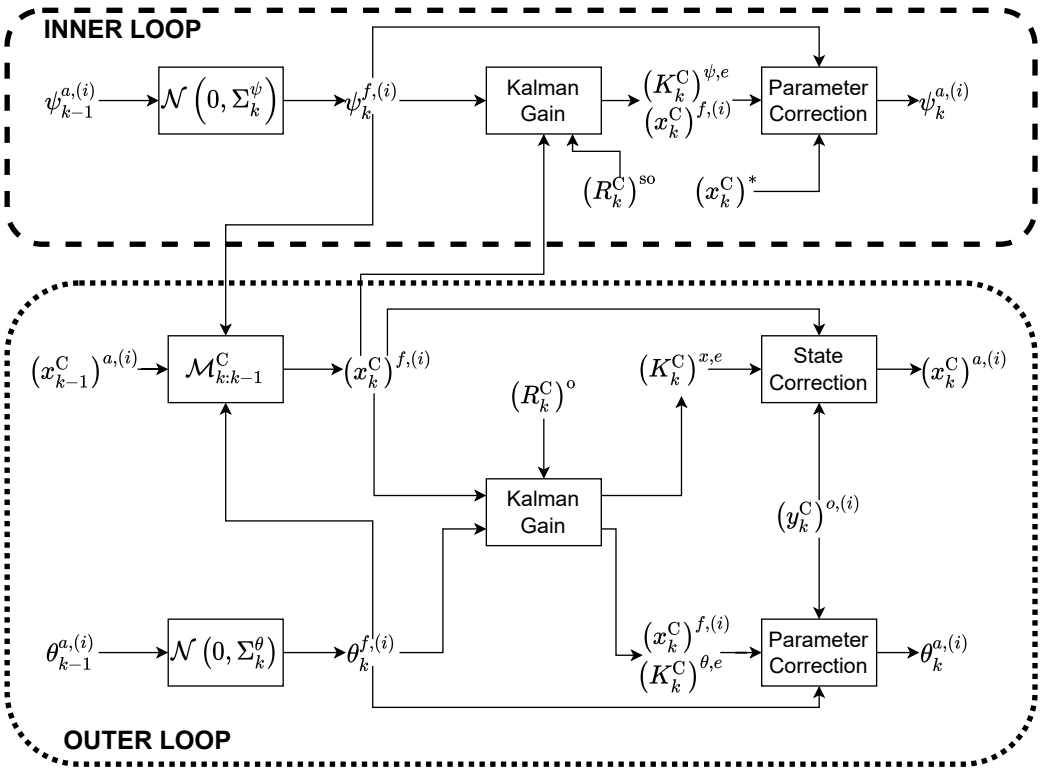


Figure 4-2: Schematic representation of the box *Dual EnKF* in Fig. 4-1. Two optimization procedures, referred to as *inner loop* and *outer loop*, are sequentially performed. Essentially, the outer loop corresponds to the classical Dual EnKF algorithm, while the inner loop is a classical EnKF procedure where only parameter estimation is performed.

1. **Predictor step**. The initial solution on the fine grid $\left(\mathbf{x}_{k-1}^{\mathrm{F}}\right)^{\mathrm{a}}$ is used to calculate a forecast state $\left(\mathbf{x}_{k}^{\mathrm{F}}\right)^{\mathrm{f}}$

$$\left(\mathbf{x}_{k}^{\mathrm{F}}\right)^{\mathrm{f}} = \mathcal{M}_{k:k-1}^{\mathrm{F}}\left(\left(\mathbf{x}_{k-1}^{\mathrm{F}}\right)^{\mathrm{a}}, \overline{\theta_{k}^{\mathrm{f}}}\right) \tag{4.1}$$

where $\mathcal{M}_{k:k-1}^{\mathrm{F}}$ is the model used on the fine grid and $\overline{\theta_{k}^{\mathrm{f}}}$ is a set of free parameters describing the setup of the model on the fine grid. Each member i of the

ensemble calculated on the coarse grid is also advanced in time

$$\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{f,(i)}} = \mathcal{M}_{k:k-1}^{\mathrm{C}}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a,(i)}}, \theta_k^{\mathrm{f,(i)}}\right) + \mathcal{C}_{k:k-1}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a,(i)}}, \psi_k^{\mathrm{f,(i)}}\right), \tag{4.2}$$

where $\mathcal{M}_{k:k-1}^{\mathrm{C}}$ is the coarse grid model parameterized by $\theta_k^{\mathrm{f,(i)}}$, while $\mathcal{C}_{k:k-1}$ is an additional correction term included to compensate the loss of resolution due to calculations on the coarse grid (similar to the strategy used by Brajard et al., 2021). This additional model, whose structure has to be determined, is driven by the set of free parameters $\psi_k^{\mathrm{f,(i)}}$.

2. **Projection on the coarse grid & *inner* loop**. $\left(\mathbf{x}_k^{\mathrm{F}}\right)^{\mathrm{f}}$ is projected on the coarse grid space via a projection operator Π_{C}, so that $\left(\mathbf{x}_k^{\mathrm{C}}\right)^*$ is obtained, i.e.,

$$\left(\mathbf{x}_k^{\mathrm{C}}\right)^* = \Pi_{\mathrm{C}}\left(\left(\mathbf{x}_k^{\mathrm{F}}\right)^{\mathrm{f}}\right). \tag{4.3}$$

   In this step, surrogate observation, $\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{SO}}$, is extracted from $\left(\mathbf{x}_k^{\mathrm{C}}\right)^*$, with an observation operator $\left(\mathcal{H}_k^{\mathrm{C}}\right)^{\mathrm{SO}}$: $\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{SO}} = \left(\mathcal{H}_k^{\mathrm{C}}\right)^{\mathrm{SO}}\left(\mathbf{x}_k^{\mathrm{C}}\right)^*$. The observation operator, $\left(\mathcal{H}_k^{\mathrm{C}}\right)^{\mathrm{SO}}$, determines the region of the solution obtained on the fine grid that is observed. Thereafter, the surrogate observation is used in the *inner* loop. Here, the EnKF is used as a *parameter estimation only* scheme, i.e., the ensemble states $\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{f,(i)}}$ are not modified, but the free parameters $\psi_k^{\mathrm{f,(i)}}$ are optimized to obtain values $\psi_k^{\mathrm{a,(i)}}$. This optimization targets an improvement of the prediction of the ensemble members simulated on the coarse grid via an update of the term $\mathcal{C}_{k:k-1}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a,(i)}}, \psi_k^{\mathrm{f,(i)}}\right)$.

3. ***Outer* loop**. If external observation $\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{O}}$ is available, the ensemble forecast $\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)}$ is corrected with the standard Dual EnKF procedure to obtain $\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{a},(i)}$ as well as an update for the parameters $\theta_k^{\mathrm{a},(i)}$.

4. **Determination of the state variables on the coarse grid**. In this step, the physical state of the main simulation is updated on the coarse grid. This solution, which will be referred to as $\left(\mathbf{x}_k^{\mathrm{C}}\right)'$, is obtained by classical iterative

procedures on the coarse grid using the initial solution $(\mathbf{x}_k^{\mathrm{C}})^*$ if observations are not available. On the other hand, if observations are available, the Kalman gain matrix, $(\mathbf{K}_k^{\mathrm{C}})^{x,\mathrm{e}}$, determined in the framework of EnKF for the ensemble members is used to determine the coarse grid solution $(\mathbf{x}_k^{\mathrm{C}})^{'}$ through a KF operation, i.e.,

$$(\mathbf{x}_k^{\mathrm{C}})^{'} = (\mathbf{x}_k^{\mathrm{C}})^* + (\mathbf{K}_k^{\mathrm{C}})^{x,\mathrm{e}} \left[(\mathbf{y}_k^{\mathrm{C}})^{\mathrm{O}} - (\mathcal{H}_k^{\mathrm{C}})^{\mathrm{O}} \left((\mathbf{x}_k^{\mathrm{C}})^* \right) \right]. \tag{4.4}$$

5. **Final iteration on the fine grid**. The fine grid state solution $(\mathbf{x}_k^{\mathrm{F}})^{'}$ is determined using the results obtained on the coarse space: $(\mathbf{x}_k^{\mathrm{F}})^{'} = (\mathbf{x}_k^{\mathrm{F}})^{\mathrm{f}} + \Pi_{\mathrm{F}} \left((\mathbf{x}_k^{\mathrm{C}})^{'} - (\mathbf{x}_k^{\mathrm{C}})^* \right)$. The state $(\mathbf{x}_k^{\mathrm{F}})^{\mathrm{a}}$ is obtained from a final iterative procedure starting from $(\mathbf{x}_k^{\mathrm{F}})^{'}$, following the multigrid approach.

A detailed representation of the *Dual EnKF* box shown in Fig. 4-1 is provided in Fig. 4-2. The functioning of the MGEnKF is illustrated in Fig. 4-3. At the beginning of the assimilation procedure, a significant mismatch between coarse grid ensemble and main fine grid simulation is expected. In the first phases of the assimilation process, the algorithm may not perform accurately, since the analysis phase is penalized by the high numerical errors present in the coarse grid ensemble realizations. As the assimilation progresses and the model correction parameters, ψ, are optimized, the mismatch between fine grid and coarse grid models is expected to be reduced, giving rise to an increased global accuracy of the estimator.

This algorithm has been specifically conceived to reduce the computational costs associated with the classical EnKF approach for large scale problems. For this, we combine a multigrid framework, frequently encountered in flow solvers, and a Dual Ensemble Kalman filter. This method falls into the class of *multilevel* techniques that aim to improve the estimation of statistics of expensive numerical simulations by considering different levels of resolution – in time or in space – of the same set of equations. In multilevel Monte Carlo applications, a few high-resolved solutions are combined with a larger number of low-resolution data (see Fossum et al., 2020; Hoel et al., 2016; Law et al., 2020; Siripatana et al., 2019, for some applications).

Instead of considering additional simulation models for the same set of equations and several resolutions, it is also possible to reduce the variance of Monte Carlo methods by considering different sets of equations (surrogate models at different levels of accuracy). This approach, called *multifidelity*, has recently been used with POD Galerkin reduced-order models (Popov et al., 2021) and for uncertainty quantification (Gorodetsky et al., 2020).

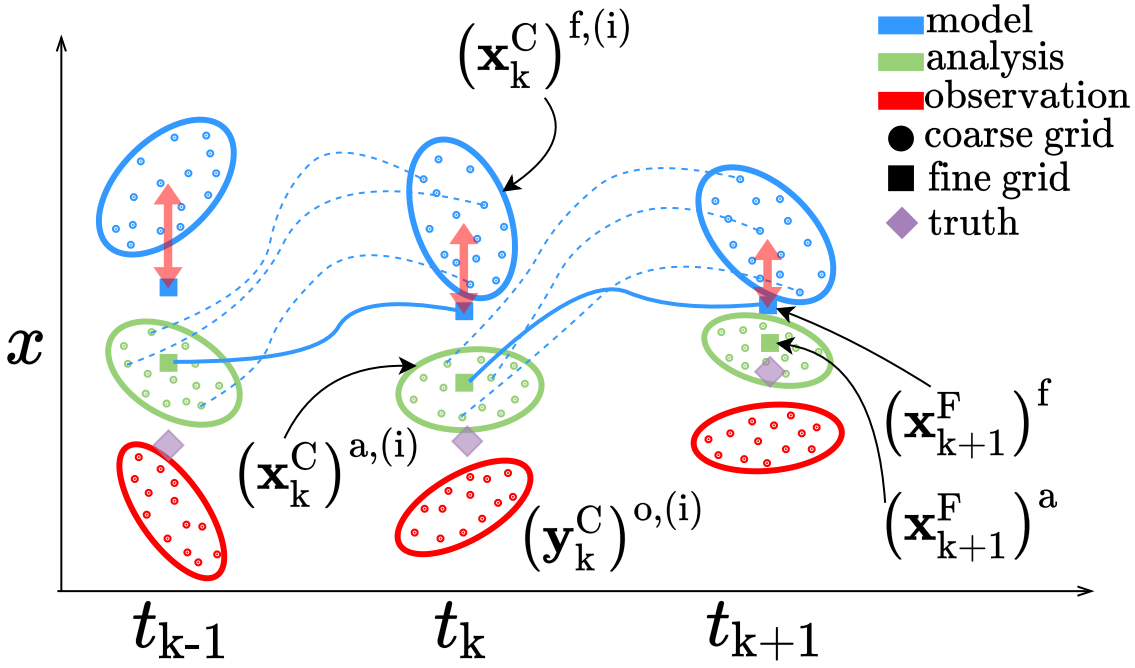


Figure 4-3: Schematic representation of the MGEnKF in a DA problem.

Two important features must be discussed:

- The recursive structure of the algorithm allows for integration of iterative corrections for non-linear systems (Sakov et al., 2011) as well as hard constraints (see the discussion in the introduction of (Nachi, 2007; Simon & Chia, 2002; X. Zhang et al., 2020)) to respect the conservativity of the model equations. However, these corrections may result in an increase of the computational resources required. Here, the multigrid algorithm itself is used for regularization (i.e., for smoothing the discontinuities in the physical variables produced by the update via Kalman Filter) of the flow. If an intentionally reduced tolerance is imposed in the iterative steps 4 and 5, the final solution will keep memory of the

features of the state estimation produced in step 3. The iterative resolution will smooth the estimation via the state transition model $\mathcal{M}$, which will perform a natural regularization of the flow. Clearly, if a reduced tolerance is imposed, the final solution will not necessarily respect the conservativity constraints of the model equations. However, one can argue that complete conservativity is not an optimal objective in this case if the model state at the beginning of the time step is not accurate.

- The state obtained on the fine grid is used to improve the accuracy of the calculation of the ensemble members via a second nested EnKF application. This nested cycle uses as observation the sampled data from the fine grid level prediction (variable $(\mathbf{x}_k^{\mathrm{C}})^*$) to infer the parametric behaviour of an ad-hoc modelling term, $\mathcal{C}_{k:k-1}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a},(\mathrm{i})}, \psi_k^{\mathrm{f},(\mathrm{i})}\right)$, included in the time advancing model for the ensemble members (4.2). This second, internal EnKF procedure called *inner loop* is used only to infer the parametric description of the coefficients $\psi_k^{\mathrm{f},(\mathrm{i})}$, and no modification to the state is performed, in order to avoid the collapse of the solutions of the ensemble members. A similar procedure, although not in the framework of multigrid applications, has been very recently proposed byBrajard et al., 2021. In their work, the correction model, $\mathcal{C}_{k:k-1}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a},(\mathrm{i})}, \psi_k^{\mathrm{f},(\mathrm{i})}\right)$, attempts to reduce the difference between the state obtained using two models with different accuracy. For application to turbulent flows, one could envision for example to run a Large Eddy Simulation (LES) ensemble on the coarse grid level and to use DNS results on the fine grid level to infer the behaviour of subgrid scale modelling for the LES.

The advantages of our strategy with respect to classical approaches based on EnKF may be summarized in the following points:

- The RAM requirement necessary to store the N_{e} ensemble members during the assimilation is usually moderate. The reduction in computational costs is driven by N_{e} and by the size of the coarse variables. To illustrate this, let us consider the case of a simple two-level geometric multigrid approach for a

3D test case with a constant coarsening ratio, $r_{\mathrm{C}} = 4$, and a sample size of $N_{\mathrm{e}} = 100$. Each ensemble member is then described by $4^3 = 64$ times less mesh elements than the single simulation on the fine grid. If one considers that one main simulation and 100 ensemble members are run simultaneously, and if the RAM requirement is normalized over the main simulation, this implies that R_{RAM}, the non-dimensional RAM requirement, is equal to $1 + 100/64 = 2.56$. In other words, the total cost in RAM is increased to just 2.56 times the cost of the simulation without EnKF. For $r_{\mathrm{C}} = 8$, the normalized RAM requirement is $R_{\mathrm{RAM}} = 1 + 100/8^3 = 1.195$, thus, just a 20% increase in RAM requirements. This is clearly orders of magnitude more advantageous than a fine-grid classical EnKF application with $N_{\mathrm{e}} = 100$, since in this case $R_{\mathrm{RAM}} = N_{\mathrm{e}} = 100$.

- The computational cost relative to the ensemble forecast on the coarse grid can typically become less important than the cost of the single simulation retained on the fine grid, depending on the r_c value. Considering that the ensemble members in the coarse grid and the simulation over the fine grid are running simultaneously, communication times are optimized.

- Owing to the iterative procedures of steps 4 and 5, regularization of the final solution is naturally obtained.

- The algorithm is here described and tested in the framework of geometric multigrid, but it can actually be integrated within other algorithmic structures, such as the algebraic multigrid. It can also be used with conventional solvers on multiple grids; however, in this case regularization is not obtained unless specific state corrections are included.

The specificity of MGEnKF is that it is tailored for the simulation of compressible flows, for which spurious oscillations produced by the KF procedure may be responsible for irreversible numerical instabilities. Fig. 4-4 illustrates the kind of unwanted effects that the brute correction of the fine grid could produce. The potential spurious oscillations can be smoothed through the multigrid strategy.

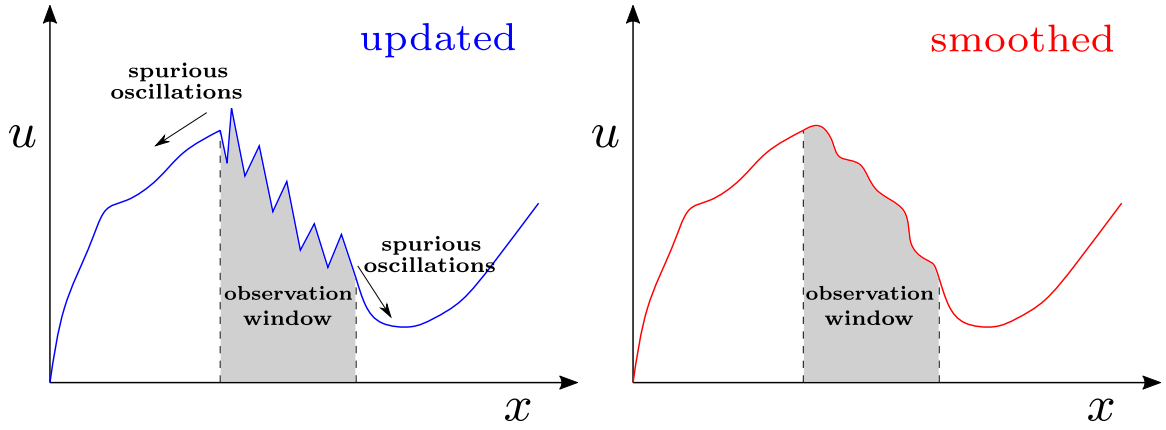


Figure 4-4: The updated solution on the fine grid of the MGEnKF approach does not respect the original model equations, thus breaking conservativity. The objective is to *smooth* this state correction through the multigrid approach.

This general algorithm may be easily tailored accounting for the complexity of the test case investigated, in particular for the requirements of iterative loops on both the coarse grid level and the fine grid level. The algorithm that we used to validate our approach is described in A.4.

Chapter 5

Validation of the MGEnKF

5.1 Introduction

In the previous chapter, we have presented the multigrid ensemble Kalman filter (MGEnKF). This strategy manipulates data using several meshes with different resolutions, exploiting the natural multilevel structure of multigrid solvers. In this chapter, we show that the main advantage of this MGEnKF is that a good level of accuracy of the DA procedure (comparable to classical application of the EnKF) is conserved with a significant reduction of the computational resources required. In the case of the classical FAS two-level multigrid algorithm, the sources of information operating in the MGEnKF are the following:

- *One main simulation* whose final solution at each time step is provided on the *fine* level of the grid.
- *An ensemble of low-resolution simulations*, which are performed at the *coarse* level of the grid.
- Some *observation* which is provided locally in space and time in the physical domain.

The MGEnKF is applied to the analysis of different test cases. Several dynamical systems of increasing complexity were chosen in order to highlight different properties

of the algorithm. All the test cases performed to assess the performance of the estimator are twin DA experiments with a very similar structure:

- A flow model is used with a specific parametrized inlet condition to generate a true state. Synthetic observation is generated from this true state by adding random Gaussian perturbations.
- In the actual DA experiment, the flow is initialized at a state which is not the true one. We further assume that some parameters of the inlet condition are unknown. The objective of the estimator is then twofold: to assimilate the synthetic observations to correct the state of the flow and to infer the unknown inlet parameters.

This chapter is organized as follows:

- In Sec. 5.2, a 1D twin-experiment is performed with the advection equation as a model for the time-evolution of the system. This simplified flow model allows to comprehensively analyse the performance of the estimator in different scenarios. The reader may also find this section very instructive for understanding how the MGEnKF works.
- In Sec. 5.3, the model used is the 1D Burgers' equation. This prototype model is fundamental in numerical fluid mechanics research, since, in 1D, it is equivalent to the Navier-Stokes equation but without the pressure gradient. It contains thus the full convective non-linearity of the equations, as well as viscosity effects. Moreover, analytical solutions for this equation can be obtained for some cases, and these are essential for validating numerical approaches.
- In Sec. 5.4, the model chosen for the twin-experiment is the 1D Euler equations. In contrast to the Burgers' equations, in the Euler equations the flow is considered to be inviscid. However, there are additional complexities related to the effects of the pressure gradient. In contrast to the twin-experiments performed in the previous two sections, in this section we attempt to estimate a time-varying parameter related to the inlet boundary condition.

- Finally, in Sec. 5.5, a twin experiment is performed where the model is the 2D compressible Navier-Stokes equations. Similarly to the test performed in the previous section, a time-varying inlet forcing parameter is estimated from noisy observations.

In the Sec. 5.2 and Sec. 5.3, an extensive analysis is performed on two different test cases to assess the effects of the *inner loop* over the global accuracy obtained via the MGEnKF. While the accuracy of the numerical model employed to obtain the predicted states for the ensemble members is directly affected by the *outer loop*, further significant improvement is expected with the application of the *inner loop* for two main reasons. The first one is that the usage of surrogate observation from the main simulation is *consistent* with the numerical model used for time advancement on the different refinement levels of the computational grids. Therefore, biases that can affect data assimilation using very different sources of information (such as experiments and numerical results) are naturally excluded. One can also expect a faster rate of convergence owing to this property. The second valuable feature of the *inner loop* is that, as the whole physical state of the main simulation is known on the fine grid level, the surrogate observation can be sampled everywhere in the physical domain. One of the main problematic aspects when assimilating experimental results in numerical models is that, often, the placement of sensors is affected by physical limitations which can preclude the sampling in highly sensitive locations. This problem is completely bypassed in the *inner loop*, where the user can arbitrarily select the number and the location of sensors.

5.2 One-dimensional advection equation

The estimator is assessed in a 1D test case where the time evolution of the flow is modelled using the advection equation. The sensitivity of the MGEnKF to the performance of the *inner loop* introduced in Chapter 4 is analysed. More precisely, the *inner loop* is here used to optimize the behaviour of the numerical model when used on a coarse grid, as discussed in Sec. 5.2.1. The optimization is performed over

an ensemble of parameters that characterize a spatially variable model correction term.

5.2.1 Optimization of numerical integration schemes on a coarse mesh

An original feature of the MGEnKF algorithm presented in Chapter 4 is that the solution advanced at the most refined level of the grid can also be used as surrogate observation to optimize the parameters of the numerical schemes. In this section, we present the principles guiding the choice of the numerical scheme and associated parametrization retained to control the numerical errors on coarse meshes (shape of $\mathcal{C}$ in the MGEnKF algorithm). Let us consider the prototype 1D linear advection equation of a scalar quantity u advected with the constant velocity c:

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0 \tag{5.1}$$

For simplicity of presentation, we restrict ourselves to using an explicit finite difference scheme on four-point stencils (second or third order accurate) for the case $c > 0$. The spatial discretization is performed on a Cartesian mesh with a constant size Δx. Δt is the time step and $\sigma = c\Delta t/\Delta x$ the CFL number. $u_{\mathrm{j}}^{\mathrm{k}}$ represents the discrete numerical solution at the spatial location $x_{\mathrm{j}} = (j-1)\Delta x$ at time $t = k\Delta t$. The following considerations can be extended quite easily to non-linear systems with time and space varying advection velocity, irregular mesh or higher-order schemes.

A general one-parameter family of second order accurate schemes (see for example Hirsch, 2007 p. 364) may be defined on a backward upwind stencil $(j-2, j-1, j, j+1)$ as:

$$\begin{aligned} u_{\mathrm{j}}^{\mathrm{k}} = \; u_{\mathrm{j}}^{\mathrm{k\text{-}1}} - \frac{\sigma}{2}\left(u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} - u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}}\right) + \frac{\sigma^2}{2}\left(u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} - 2u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}}\right) \\ +\delta\left(-u_{\mathrm{j\text{-}2}}^{\mathrm{k\text{-}1}} + 3u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}} - 3u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}}\right) \end{aligned} \tag{5.2}$$

The first line of (5.2) corresponds to the standard centred Lax Wendroff scheme. The second line can be interpreted as the discretization of an additional dispersion

term of the form $\delta\Delta x^3 \frac{\partial^3 u}{\partial x^3}$. The expression (5.2) is compatible with the linear advection equation (5.1) discretized at precision order proportional to $(\Delta x^2, \Delta t^2)$. If one retains the two first dominant error terms in the combination of Taylor expansions corresponding to these discrete terms forming scheme (5.2), the following equivalent differential equation is obtained:

$$\begin{aligned} \frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = & \frac{c}{6}\left[6\delta - \sigma\left(1-\sigma^2\right)\right]\Delta_x^2 \frac{\partial^3 u}{\partial x^3} \\ & -\frac{c}{8}\left[\delta\left(1-2\sigma\right) + \frac{\sigma^2}{4}\left(1-\sigma^2\right)\right]\Delta_x^3 \frac{\partial^4 u}{\partial x^4} + \mathcal{O}\left(\Delta_x^4\right) \end{aligned} \tag{5.3}$$

Equation (5.3) reveals that the dominant numerical error of (5.2) is dispersive with an error proportional to Δx^2 and that a less dominant (third-order) diffusive error term occurs. The expressions of these errors show that their relative level varies both as a function of σ and δ. In order to keep the MGEnKF algorithm sufficiently flexible and general, we will not search to optimize the parameter σ. Therefore, it is chosen to target only an optimization for δ with a strategy suitable for any value of σ. This choice allows the user to set the value of σ based on practical constraints, such as for synchronizing simulations with available observation data. The level of the dominant error can be controlled through the parameter δ. By setting in particular $\delta = \sigma(1-\sigma^2)/6$, the dominant error term cancels out, and the scheme becomes third order accurate in space with now a diffusive dominant error term proportional to Δx^3. Other values of δ maintain the formal second order accuracy, but can induce significantly different effective evolution of the numerical errors. An illustration of this is shown in Fig. 5-1 where the diffusion and dispersion errors are represented. These errors are quantified through a classical Fourier analysis, through which the expression of the complex gain factor $G = \mathrm{Re}(G) + \jmath\,\mathrm{Im}(G)$ is extracted as a function of the phase angle $\phi = m\Delta x$ with m corresponding to a spatial wavenumber. It is worth recalling that the phase angle $\phi = m\Delta x = 2\pi/\lambda$, where λ is the spatial wavelength, can also be written as $\phi = 2\pi/(N-1)$, where N represents the number of points used to discretize the signal over λ. To ensure stability, the modulus of the gain factor $|G|$ should remain less than 1 for the whole range of ϕ present in the signal being advected. However,

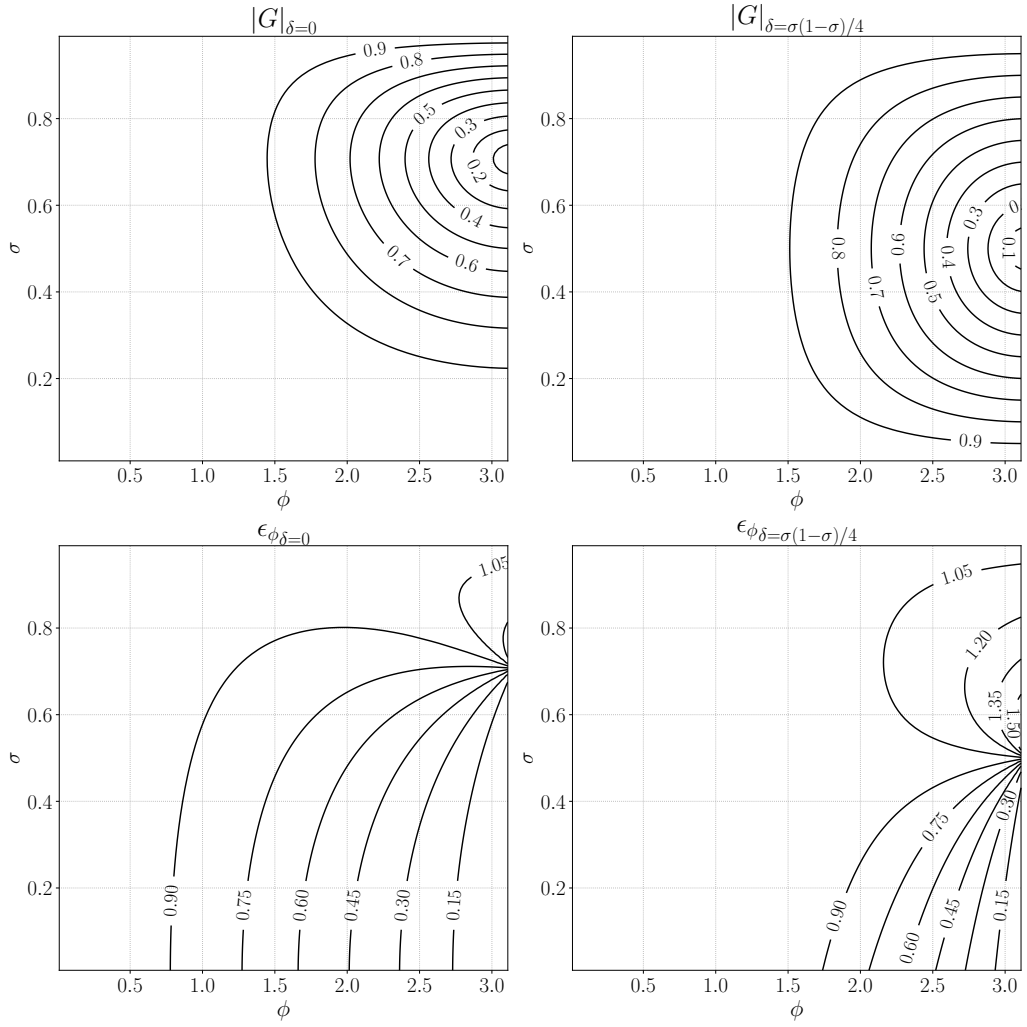


Figure 5-1: Comparison of diffusive ($\epsilon_D = |G|$, top) and dispersive (ϵ_ϕ, bottom) errors for two different a priori parametrisations of δ. $\delta = 0$ (left column) corresponds to the Lax Wendroff scheme and $\delta = \sigma(1-\sigma)/4$ (right column) to Fromm's scheme.

$1-|G|$ represents the level of numerical diffusion. This quantity might be minimized to avoid artificial decrease of wave amplitude components. The dispersive error is here characterized by $\epsilon_\phi = \arctan(-\mathrm{Im}(G)/\mathrm{Re}(G))/\sigma\phi$, which corresponds to the spurious multiplicative factor affecting the expected phase velocity of wave components. A good numerical scheme should keep the value of ϵ_ϕ as close as possible to unity to limit phase advance or delay observed for $\epsilon_\phi > 1$ or $\epsilon_\phi < 1$, respectively. Figure 5-1 clearly shows the role played by δ on the diffusion and dispersion errors. The case $\delta = 0$ (Lax Wendroff scheme) is characterized by a dominant phase lag within the stability bounds $0 < \sigma < 1$. The choice $\delta = \sigma(1-\sigma)/2$ (not shown) corresponds to the Beam and Warming scheme, which yields a dominant phase advance error for $0 < \sigma < 1$. The case $\delta = \sigma(1-\sigma)/4$ corresponds to the famous Fromm's scheme, which compensates to some extent the phase errors of the two aforementioned schemes for a wide range of σ. Indeed, we notice in Fig. 5-1 that isolines of ϵ_ϕ lower than unity are significantly shifted towards the higher values of ϕ, indicating that dispersive error levels can be a priori significantly reduced in the intermediate range of ϕ corresponding to practical simulation cases. However, the reduction of diffusive errors, as illustrated with $|G|$ is far less efficient, in particular for high values of σ. The diffusive errors are even seen to increase for lower values of σ and high values of ϕ.

We have to keep in mind that the MGEnKF algorithm employs ensemble members which have to be generated using relatively coarse grids (thus high values of ϕ) for which both the dispersive and diffusive errors are likely to be important. The aforementioned observations of the non-monotonic and uncorrelated evolutions of $|G|$ and ϵ_ϕ suggest that adjusting only δ is not sufficient to allow a satisfactory control of both kind of errors at the same time. An optimization of δ made for reducing the dispersive error could undesirably deteriorate the diffusive behaviour of the scheme. This justifies that in the optimization strategy that we consider in Sec. 5.2, we add to (5.2) an additional correction of the same order as the dispersive correction term in factor of δ. This correction is chosen as being consistent with $\alpha\Delta x^2 \frac{\sigma^2}{2}\frac{\partial^2 u}{\partial x^2}$. With negative values of α, this term will be expected to add an anti-diffusive behaviour, counteracting the diffusion error intrinsically associated with the scheme (5.2). The

combined use of both correction terms is thus expected to allow a more relevant separated control of both dispersion and diffusion errors.

As previously observed in Fig. 5-1, the properties of the numerical scheme significantly vary as function of ϕ. It is therefore far from being evident that considering δ and α as constant optimization parameters is sufficient to reduce the numerical error over a wide range of ϕ scales. In view of considering complex (spectrally richer) solutions and extending the use of this scheme to non-linear models, possibly leading to spatially evolving frequency content, it is thus also chosen to consider spatially varying functions for δ and α instead of constant values. This variability, which will be represented by expressing these parameters via spatial expansions of polynomials, will allow for local numerical optimization on coarse meshes.

Summary: numerical scheme retained to perform coarse-grid simulations

Following the a priori analysis given all along this section, the following numerical scheme is finally retained:

$$\begin{aligned} u_{\mathrm{j}}^{\mathrm{k}} = \; & u_{\mathrm{j}}^{\mathrm{k\text{-}1}} - \frac{\sigma}{2}\left(u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} - u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}}\right) + (1+\alpha)\,\frac{\sigma^2}{2}\left(u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} - 2u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}}\right) \\ & + (\delta+\gamma)\left(-u_{\mathrm{j\text{-}2}}^{\mathrm{k\text{-}1}} + 3u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}} - 3u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}}\right) \\ = \; & \mathcal{M}_{k:k-1}\left(u;\sigma,\delta\right) + \mathcal{C}_{k:k-1}\left(u;\sigma,\alpha,\gamma\right) \end{aligned} \tag{5.4}$$

Here, one can see that the optimization of δ is not performed directly, but via a parameter γ which measures the deviation of the optimized dispersion coefficient from the constant value $\delta = \sigma(1-\sigma^2)/4$ proposed by Fromm. This choice has been performed to provide a clear separation between the dynamical model $\mathcal{M}$ and the correction model $\mathcal{C}$ when comparing (5.2) and (5.4). The variability in space of the coefficients γ and α is obtained expressing them in terms of Legendre Polynomial expansions truncated to the order n:

$$\gamma\left(x\right) = \gamma_0 P_0\left(x\right) + \gamma_1 P_1\left(x\right) + \cdots + \gamma_{\mathrm{n}} P_{\mathrm{n}}\left(x\right), \tag{5.5}$$

and

$$\alpha(x) = \alpha_0 P_0(x) + \alpha_1 P_1(x) + \cdots + \alpha_n P_n(x). \tag{5.6}$$

Preliminary tests showed that 4-th order representation ($n = 4$) is satisfactory for the cases considered in this study. The inner loop will be used to optimize the expansion coefficients for a total of 10 parameters (five expansion coefficients γ_i and five expansion coefficients α_i). The values for these parameters could possibly be constrained during the inner loop optimization in order to accept only values leading to stable solutions. In particular, the extended stability constraint for the present scheme in absence of additional anti-diffusive correction is imposed, which reads as:

$$\gamma(1 - 2\sigma) + \frac{1}{4}\sigma^2(1 - \sigma^2) \geq 0 \tag{5.7}$$

5.2.2 Set-up of test case and test solutions on coarse meshes

The set-up of the test case representing the one-dimensional linear advection equation (see Eq. (5.1) in Sec. 5.2.1) is now presented. The constant advection velocity is set to $c = 1$. Preliminary numerical tests are carried out with the scheme presented in Eq. (5.4) and by setting a priori $\alpha = 0$ and $\gamma = 0$ (Fromm scheme). The initial condition is set to $u(x, t = 0) = c$ everywhere in the physical domain. A Dirichlet time-varying condition is imposed at the inlet:

$$u(x = 0, t) = c\,(1 + \theta \sin(2\pi t)), \tag{5.8}$$

where θ represents the amplitude of a sinusoidal perturbation of period $\Delta T = 1$ and is set to $\theta = 0.015$. The outlet boundary condition is extrapolated from the last points of the domain using 4-th order Lagrange polynomials. The simulations are performed over a computational domain of size $0 \leq x \leq 10$ in $L_0 = c\Delta T$ units. Three different levels of mesh refinement (moderate to low) are chosen for these tests. The resolution is chosen to be of practical interest for usage for the ensemble members in the MGEnKF algorithm. The mesh size is set with constant values of

$\Delta_x = 0.0625$, 0.1 and 0.125, respectively. This corresponds to 16, 10 and 8 discrete nodes per characteristic length L_0 or equivalently, phase angles around $\phi = 0.4$, 0.6 and 0.8. According to Fig. 5-1, relatively moderate error levels can be observed with these resolution levels, and their accumulation during the signal advection is expected to become significant. Preliminary simulations are also performed using different CFL numbers σ. The results, which are shown in Fig. 5-2, are compared with the *true* (exact) known solution. One can see that, with the exception of the finer mesh resolution and higher values of σ, the numerical solutions are rapidly affected by the accumulation of diffusive and dispersive errors. In particular, a significant amplitude reduction and phase advance can be observed. It is worth recalling that these errors could be naturally eliminated in the present case for the specific choice of $\sigma = 1$. However, this constraint is not necessarily compatible with practical needs associated with the numerical simulations running within the MGEnKF algorithm (reduced σ required to ensure stability with more complex boundary conditions, signal synchronization with observation, and so on).

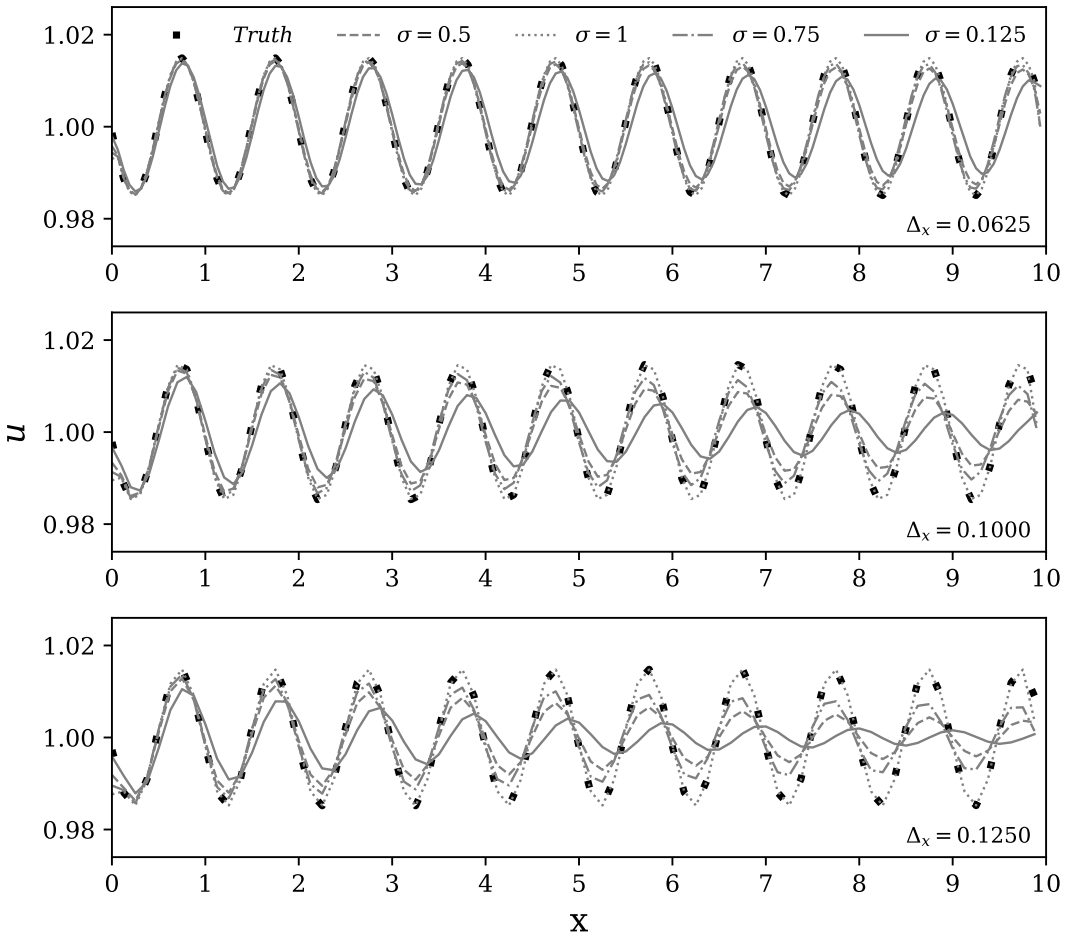


Figure 5-2: Preliminary simulations for the 1D advection equation. Solutions at $t = 10$ for different grid refinement and σ values are compared with the *true* state.

5.2.3 Performance of the MGEnKF algorithm without the *inner loop*

The performance of the MGEnKF algorithm without the *inner loop* may be severely degraded by the numerical errors induced by the use of coarse grids for the ensemble members. This can be shown with a simple twin-experiment where observation is accessible relatively far from the inlet and the DA tool attempts to estimate the inlet parameter θ. Data assimilation is performed with the following conditions:

- Observations are generated from the analytical solution with $\theta = 0.015$ on the space domain $[3, 4]$ and on the time window $[0, 390]$. The sampling frequency is

set so that approximately 15 observation updates per characteristic evolution time L_0/c are obtained, for a total of ≈ 6000 DA analysis phases. Also, the time origin for the sampling is shifted of ten characteristic times so that the state for $t = 0$ is *fully developed*, i.e., the initial condition $u(x, 0) = c$ is completely advected outside the computational domain. For simplicity, we assume that the observations and the coarse-grid ensemble are represented on the same space. Therefore, $\mathcal{H}^{\mathrm{o}}_{\mathrm{k}} \equiv \mathcal{H}^{\mathrm{o}}$ is a subsampling operator independent of time, retaining only the points comprised in the coarse space domain [3, 4]. The observations are artificially perturbed using a constant in time Gaussian noise of diagonal covariance $\mathbf{R}^{\mathrm{o}}_{\mathrm{k}} \equiv \mathbf{R}^{\mathrm{o}} = 2.25 \cdot 10^{-6}\mathbf{I}$. This choice has been performed for every test case following the recommendations of Tandeo et al., 2020, which extensively investigated the sensitivity of the EnKF to the noise/uncertainty in the model and in the observation.

- The *model* used in the MGEnKF approach is represented by i) a main simulation with a resolution of $\Delta_x = 0.0125$ (i.e., 80 mesh elements per L_0) and ii) ensemble simulations performed on coarse grids. Three different runs of the MGEnKF are performed using three different mesh resolutions (Δ_x equal to 0.125, 0.1 and 0.0625) for the ensemble members and also imposing different values for σ. The *model* employs fixed parameters $\alpha = 0$ and $\gamma = 0$ (Fromm scheme) for all cases (the inner loop is deactivated and $\mathcal{C}_{k:k-1}(u; \alpha, \gamma) = 0$). The size of the ensemble is set to $N_{\mathrm{e}} = 100$. The amplitude of the sinusoidal inlet perturbation θ for the ensemble simulations is considered to be unknown. It is initially assumed to be described by Gaussian distribution $\theta \sim \mathcal{N}(0.025, \mathbf{Q}_\theta)$, with $\mathbf{Q}_\theta(t = 0) = 2.5 \cdot 10^{-7}\mathbf{I}$. For the main simulation run on the fine grid, the mean value of the Gaussian distribution, *i.e.,* $\theta = 0.025$ is initially imposed. These values are significantly far from $\theta = 0.015$ used with the analytical solution. This choice allows analysing the rate of convergence of the optimization procedure. The initial condition $u(x, t = 0) = c$ is used for the main simulation on the fine-grid as well as for the coarse ensemble simulations. It is worth recalling that, in such

a case, at $t = 0$, the true state exhibits a very different solution. This choice allows checking the robustness of the algorithm during the transient solution and ascertain the correct evolution of the first state estimation stages when the solution of the model may be very different from the observations.

The time-evolution of the estimation of θ is shown in Fig. 5-3. First, one can see that the estimation procedure starts for $t = 10$. This choice is consistent with the positioning of the sensors for observation, which are located in the middle of the computational domain ([3, 4]).

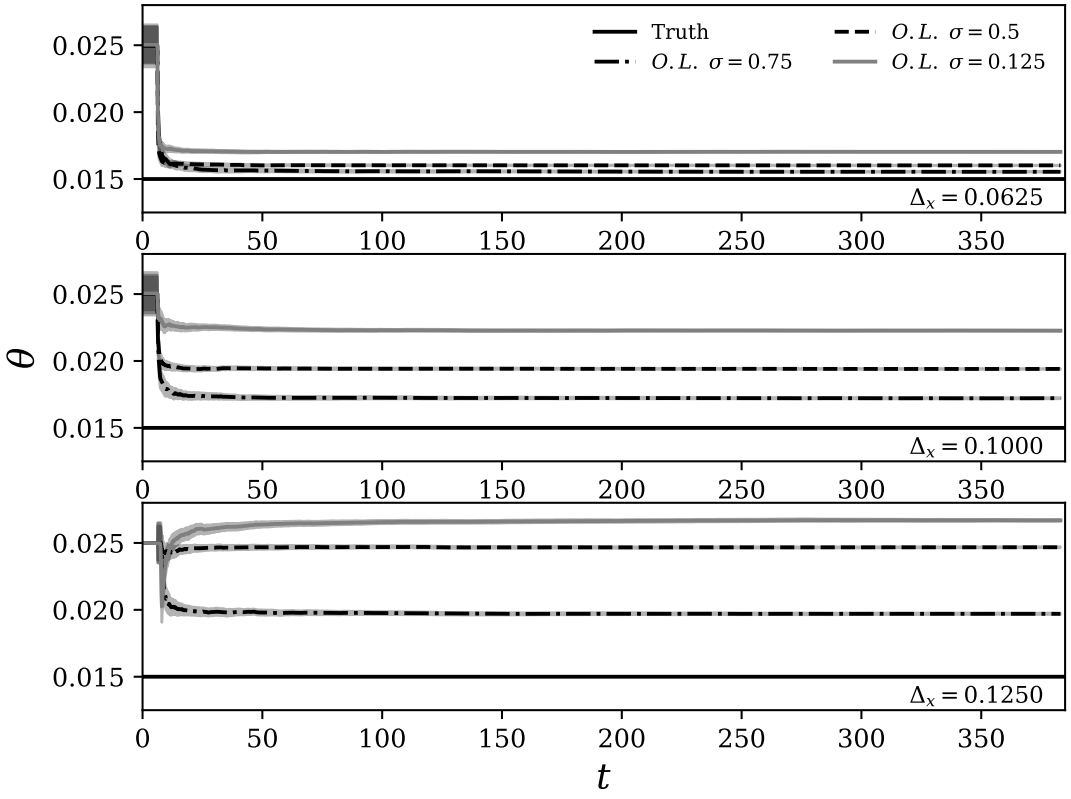


Figure 5-3: Estimation history of the inlet parameter θ using the MGEnKF without the *inner loop*. The DA method is performed using three different mesh resolutions for the ensemble members and varying the parameter σ.

The estimation of θ is progressively degraded as the refinement is decreased, with the most accurate results obtained with the finest mesh refinement $\Delta_x = 0.0625$ for any given σ. Concerning the influence of the CFL number σ, the model prediction is

generally more accurate as $\sigma \to 1$. One can also see that progressively larger deviations are observed varying σ for coarser meshes. Therefore, the increased dispersive and diffusive errors associated with lower CFL numbers are the cause for the large deviations observed when $\sigma = 0.125$. θ is naturally over-predicted in this case as the calculations of the model in the sampling region are dominated by numerical errors; the amplitude of the sinusoidal wave imposed at the inlet is numerically *diffused* and *dispersed* between $0 \leq x \leq 3$ and the MGEnKF estimator obtains the optimized value for θ which compensates for the mismatch between model and reference in the sampling region. This can be clearly observed in Fig. 5-4, where the ensemble coarse-grid estimation is shown.

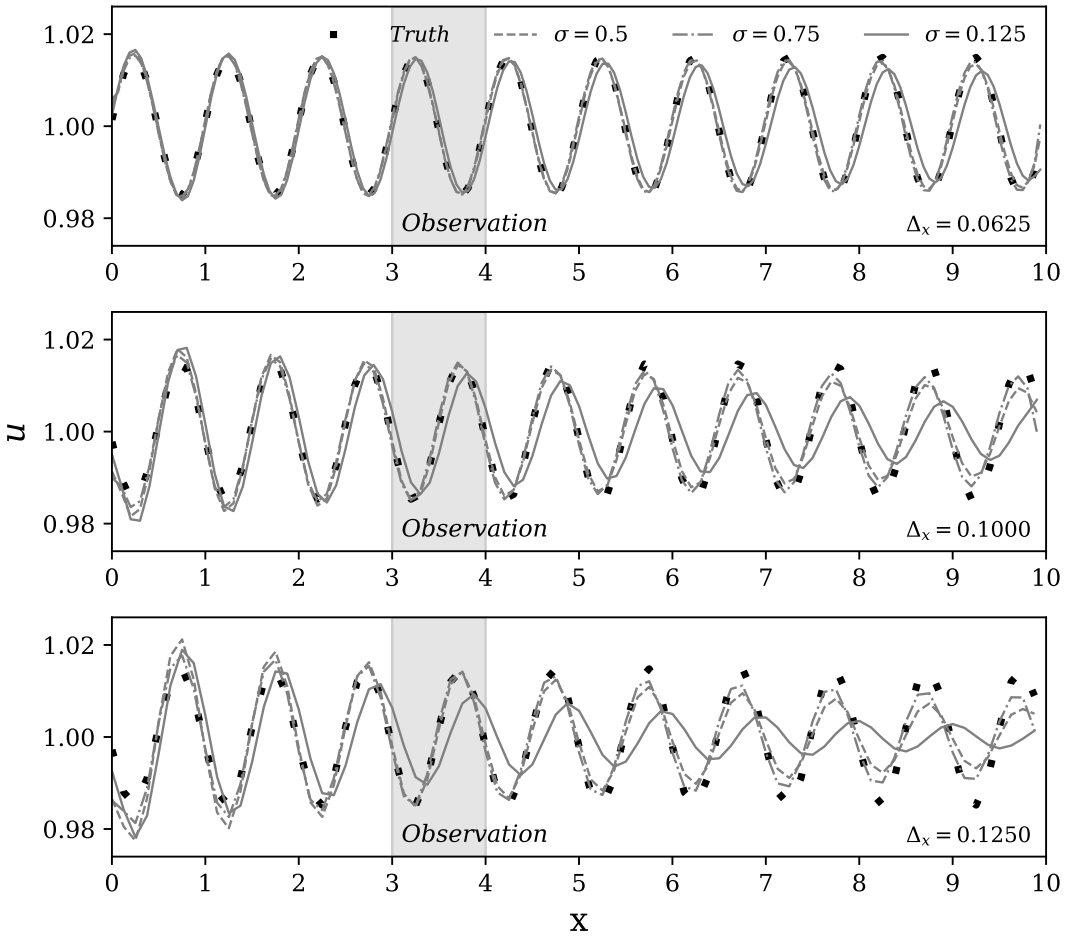


Figure 5-4: State estimation obtained for the ensemble members via MGEnKF without the *inner loop* for the linear advection equation test case. Comparisons with the exact solution are shown for $t = 300$ for different grid refinement levels and σ values.

5.2.4 Performance of the MGEnKF with the *inner loop*

In this section, the complete MGEnKF scheme is used to study the same DA problem investigated in Sec. 5.2.3. This analysis allows to unambiguously identify the contribution of the *inner loop* for the optimization of the ensemble members running on the coarse grid level.

The main modification when compared with the previous analysis is that now the parameters $\gamma(x)$ and $\alpha(x)$ in Eq. 5.4 are considered to be unknown space varying model parameters which will be optimized using the *inner loop*. The additional term related to the correction model $\mathcal{C}_{k:k-1}$ can be explicitly written as $\alpha\frac{\sigma^2}{2}\left(u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} - 2u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}}\right) + \gamma\left(-u_{\mathrm{j\text{-}2}}^{\mathrm{k\text{-}1}} + 3u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}} - 3u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}}\right)$. In particular, the optimization will target the values of the Legendre polynomial expansion coefficients γ_i and α_i introduced in Eq. 5.5 and Eq. 5.6 using the fine-grid state as surrogate observation.

The MGEnKF thus performs two optimization procedures within the analysis phase, one in the *inner loop* and a second one in the *outer loop*:

1. In the *inner loop*, optimization of the polynomial expansion coefficients α_i and γ_i to reduce the discrepancy between the low-fidelity (ensemble members) and high-fidelity (main simulation) models.

2. In the *outer loop*, optimization of the amplitude of the inlet perturbation θ.

The coefficients α_i and γ_i are initially described by Gaussian distributions $\alpha_{\mathrm{i}} \sim \mathcal{N}(0, \mathbf{Q})$, $\gamma_{\mathrm{i}} \sim \mathcal{N}(0, \mathbf{Q})$ with $\mathbf{Q}(t=0) = 9\cdot10^{-8}\mathbf{I}$ and $i = 0, 1, 2, 3, 4$. The information from the entire fine-grid domain is available. Therefore, $\mathcal{H}_{\mathrm{k}}^{\mathrm{SO}} \equiv \mathcal{H}^{\mathrm{SO}}$ is a subsampling operator independent of time retaining all the points comprised in the coarse space domain $[0, 10]$. Preliminary tests showed that, in order to improve the performance of the *inner loop*, the surrogate observation from the fine grid should be perturbed using a constant in time Gaussian noise of covariance $\mathbf{R}_{\mathrm{k}}^{\mathrm{SO}} \equiv \mathbf{R}^{\mathrm{SO}} = 1\cdot10^{-8}\mathbf{I}$. The value here used for $\mathbf{R}^{\mathrm{SO}}$ is orders of magnitude lower than the observation covariance matrix $\mathbf{R}^{\mathrm{O}}$. Therefore, one can consider the surrogate observation as a quasi-exact observation. It should be noted that the EnKF procedure in the *inner loop* only op-

timizes the parameters affecting the model term $\mathcal{C}$ included in the low-fidelity model, and no state estimation is performed on the coarse ensemble within this phase.

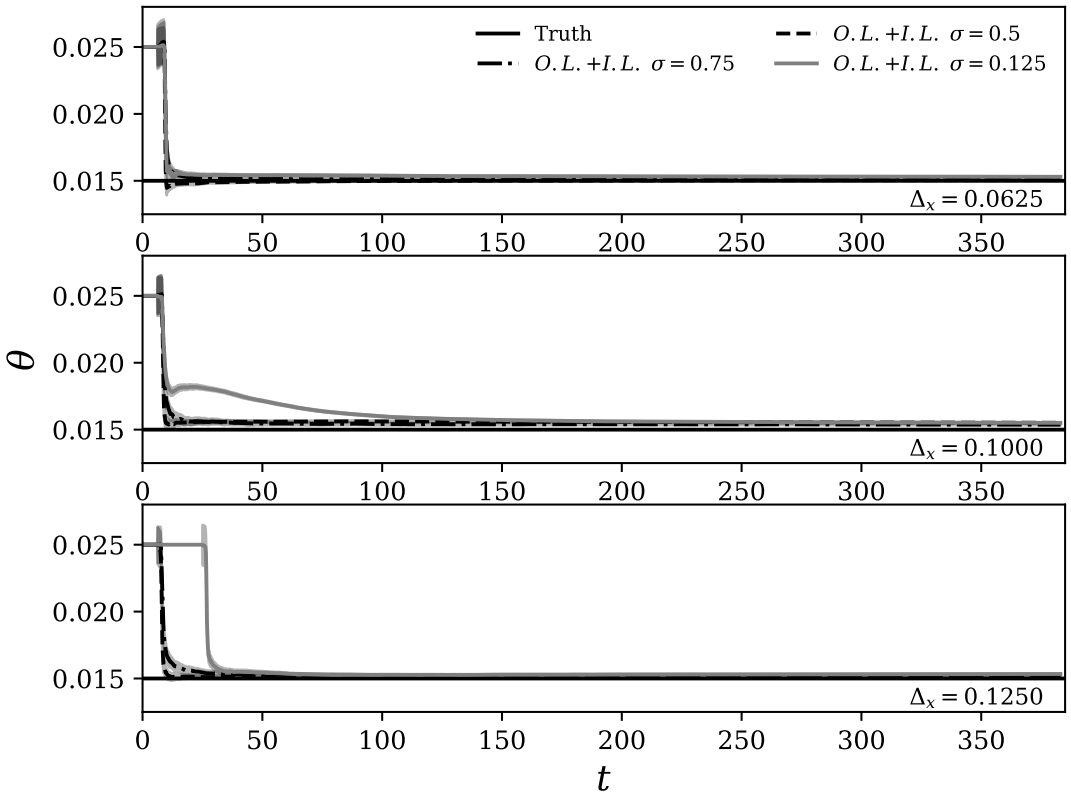


Figure 5-5: Estimation history of the inlet parameter θ using the complete MGEnKF. The DA method is performed using three different mesh resolutions for the ensemble members and varying the parameter σ.

The estimation of the parameter θ using the complete MGEnKF is shown in Fig. 5-5. As in Sec. 5.2.2, runs have been performed using three levels of mesh refinement for the ensemble members. The first ten characteristic times of the experiment are used to initialize the tuning of the parameters α and γ (i.e., *outer loop* initially deactivated). For $t > 10$, both optimization procedures are performed. During the very first phases of the assimilation process ($10 < t < 20$) the estimated value of θ gets within the 5% error margin when compared to the *truth* and no degradation is observed for every mesh refinement/σ combination investigated. The convergence is noticeably slower when $\sigma = 0.125$ for $\Delta_x = 0.100$ and $\Delta_x = 0.125$, where the numerical errors in the initial phase are the largest. For the worst case scenario

($\sigma = 0.125$, $\Delta_x = 0.125$) an initial phase of 40 characteristic times was required to obtain converged results for the inner loop, which delayed the start of the outer loop. Overall, the complete MGEnKF outperforms the version without the *inner loop*. This result highlights the complementary features of the two optimization strategies to obtain an accurate representation of the flow.

The optimization performed in the inner loop is now analysed in detail. The parameter $\alpha(x)$ is shown in Fig. 5-6 for $t = 300$. As expected, one can see that α exhibits negative values, which approach zero with increasing mesh resolution and higher σ values. The numerical diffusion observed in the model when computed on coarser meshes increases, thus the optimization procedure provides an anti-diffusive contribution.

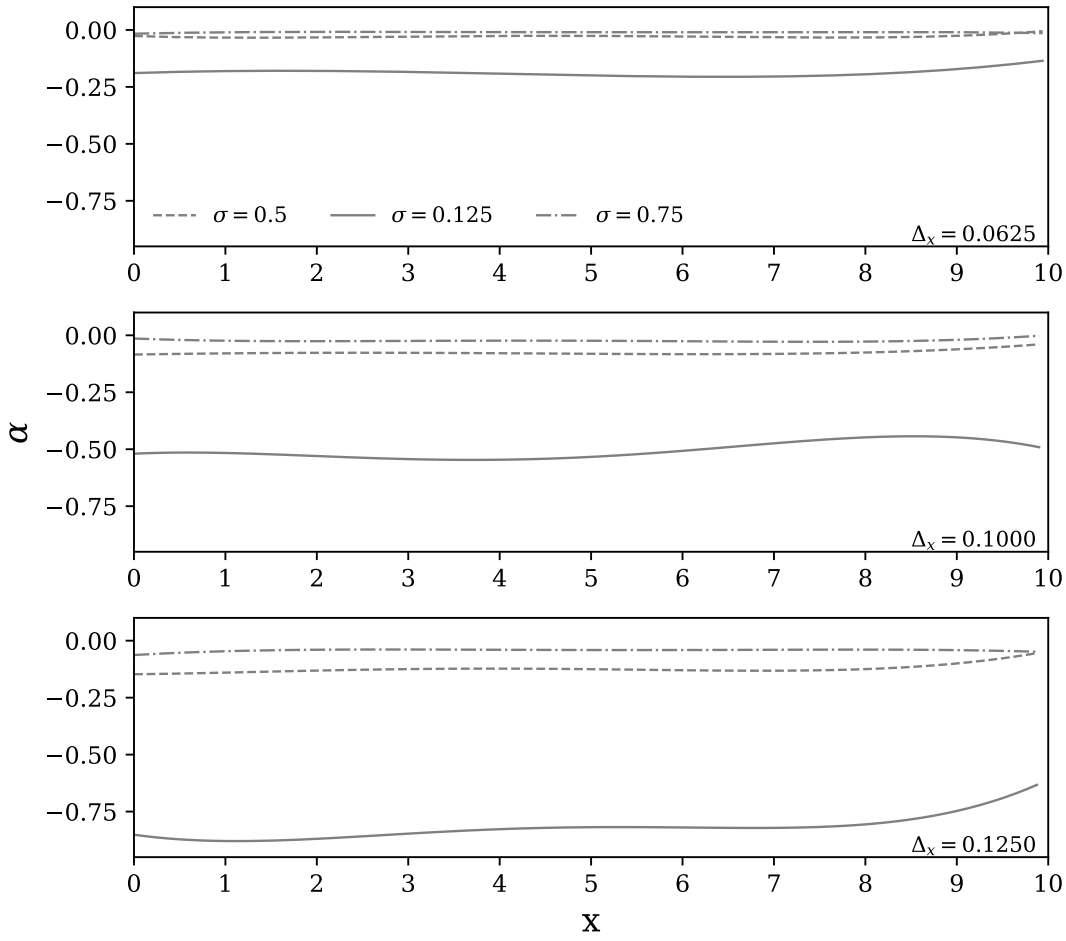


Figure 5-6: Values of the parameter α obtained via the *inner loop*. Results are shown for different grids and σ values for a simulation time $t = 300$.

The spatial distribution of the sum of the parameters $\gamma+\delta$, representing the total dispersion of the scheme (Eq. (5.4)) is presented in Fig. 5-7 for $t = 300$. For every case analysed, one can remark that $\gamma+\delta$ tend to converge towards the value for which the scheme in Eq. (5.4) becomes third order accurate, that is $\delta + \gamma = \sigma\left(1-\sigma^2\right)/6$ (0.0547, 0.0625 and 0.0205 for σ equal to 0.75, 0.5 and 0.125 respectively). This result is expected since this particular value cancels out the dominant dispersive error in the schemes, as shown in (5.3).

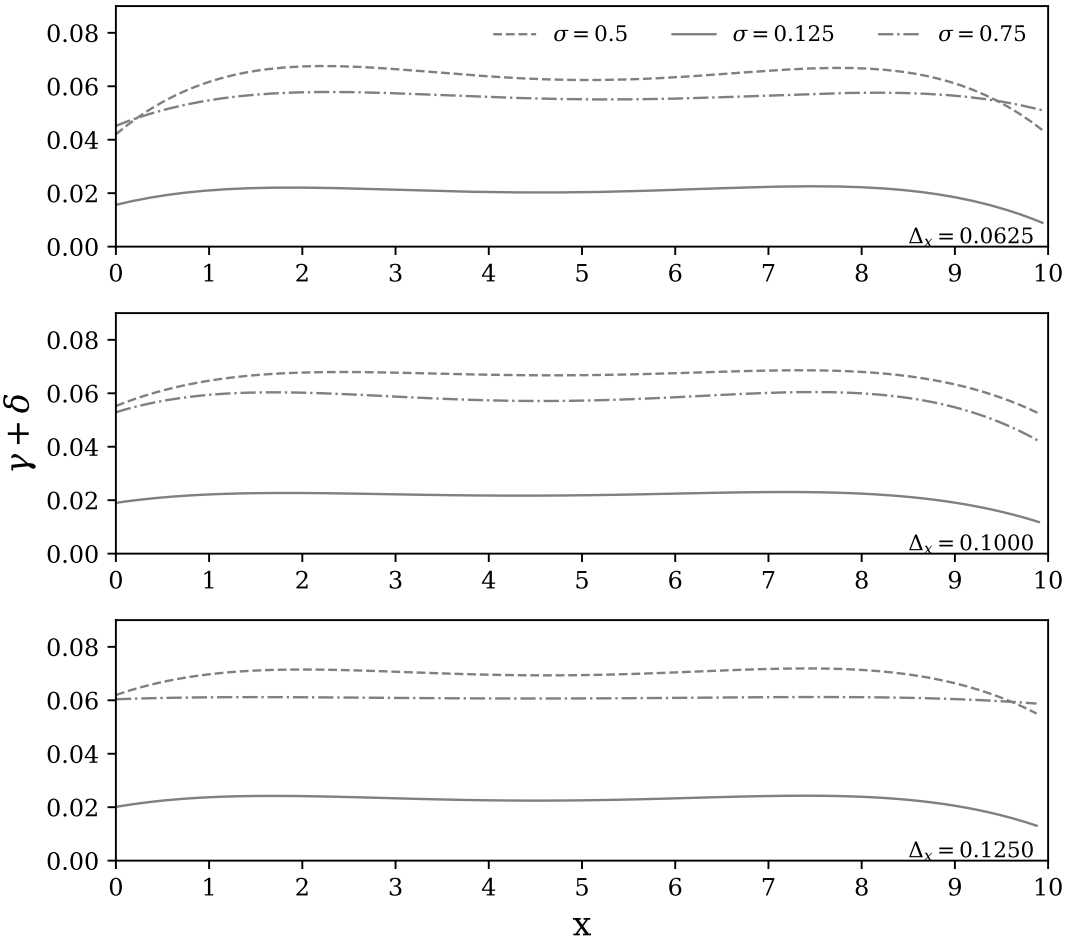


Figure 5-7: Values of the parameter $\gamma+\delta$ obtained via the *inner loop*. Results are shown for different grids and σ values for a simulation time $t = 300$.

Finally, results for the ensemble members are shown in Fig. 5-8 for $t = 300$. One can see a marked improvement when these results are compared to the ones shown in Fig. 5-4. The estimation of the parameter θ is clearly much more accurate (5%

error) and there is virtually no difference in prediction between the *truth* and model used on the coarse grids. This result has been obtained owing to the suppression of the numerical diffusion and dispersion errors via *inner loop*, which proved to be efficient for every configuration analysed (wide range of phase angle/CFL numbers). DA analyses considering more complex inlet conditions (multiple frequencies θ_i) have been performed to assess the method. The results, which are not presented here for the sake of brevity, show similar accuracy.

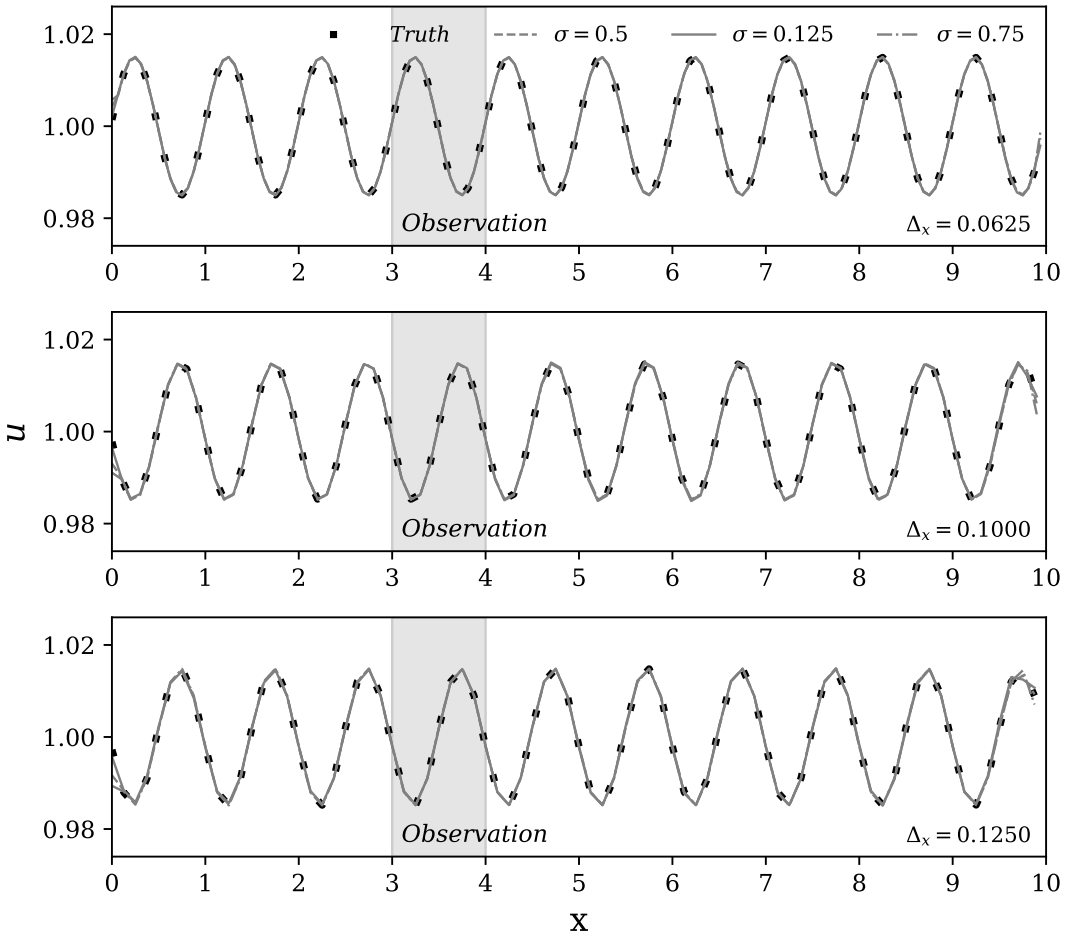


Figure 5-8: Solutions provided by the ensemble members in the complete MGEnKF. Results, which are obtained for different meshes and values for σ, are compared with the *true* state for $t = 300$.

5.3 Application: one-dimensional viscous Burgers' equation

Let us now consider the non-linear and viscous 1D Burgers equation:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2} \tag{5.9}$$

where x is the spatial coordinate, u the velocity and ν the kinematic viscosity. Considering a centred difference scheme for both the convection and the diffusion term over a uniform grid with mesh size Δx, and an explicit forward first order scheme for the time derivative, one obtains:

$$u_{\mathrm{j}}^{\mathrm{k}} = u_{\mathrm{j}}^{\mathrm{k\text{-}1}} - u_{\mathrm{j}}^{\mathrm{k\text{-}1}}\frac{\Delta_t}{2\Delta_x}\left(u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} - u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}}\right) + \nu\frac{\Delta_t}{{\Delta_x}^2}\left(u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} - 2u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}}\right) \tag{5.10}$$

where Δ_t represents the time step.

Similarly to what was done in Sec. 5.2, the performance of the MGEnKF is here studied for this test case. However, owing to the non-linear behaviour of this flow configuration, a model of the numerical error associated with the discretization process cannot be derived from the dynamic equation. Thus, for this case, the optimization by the inner loop is performed using the same model, $\mathcal{C}$, employed in Sec. 5.2. While this model has not been derived for the dynamic equations (5.9)-(5.10), one can assess the degree of precision attained in reducing amplitude and phase errors.

A numerical experiment for this test case is first performed using a high-resolution mesh to obtain a reference solution and to generate observation for the MGEnKF application. A Dirichlet time-varying condition is imposed at the inlet:

$$u(x=0,t) = u_0\left(1 + \theta\sin\left(2\pi t\right)\right), \tag{5.11}$$

where u_0 is the mean characteristic velocity of the flow and is set to $u_0 = 1$. θ represents the amplitude of a sinusoidal signal whose period is $\Delta T = 1$. The am-

plitude parameter has been set to $\theta = 0.2$ in order to observe significant non-linear effects with the Reynolds number chosen for this application, which will be discussed in the following. The outlet boundary condition is extrapolated from the nearest points to the outlet using 4-th order Lagrange polynomials. The initial condition is $u(x, t = 0) = u_0$ everywhere in the physical domain. The Reynolds number is set to $Re = \frac{u_0 L_0}{\nu} = 200$, where $L_0 = u_0 \Delta T$ is the mean wave-length of the signal and the characteristic length of the system. All physical lengths characterizing the system are normalized by L_0. The simulation is performed over a computational domain of size 10 length units, and the origin of the system is set so that $0 \leq x \leq 10$. The mesh resolution is set to 64 discrete nodes per L_0, for a total of 640 mesh elements. A constant time step Δt is set so that the mean CFL number is $CFL = \frac{u_0 \Delta_t}{\Delta_x} = 0.025$, which is small enough to guarantee a stable numerical evolution of the system.

The predicted solution using this model is referred to as the *true* state of the system. This state is first compared with the prediction obtained via a low-fidelity model, which is identical to the reference simulation but uses only 8 nodes per length L_0, for a total of 80 mesh elements. The comparison of the two solutions, which is shown in Fig. 5-9, clearly indicates that the lack of mesh resolution is responsible for important errors in the amplitude and in the phase of the velocity signal. In

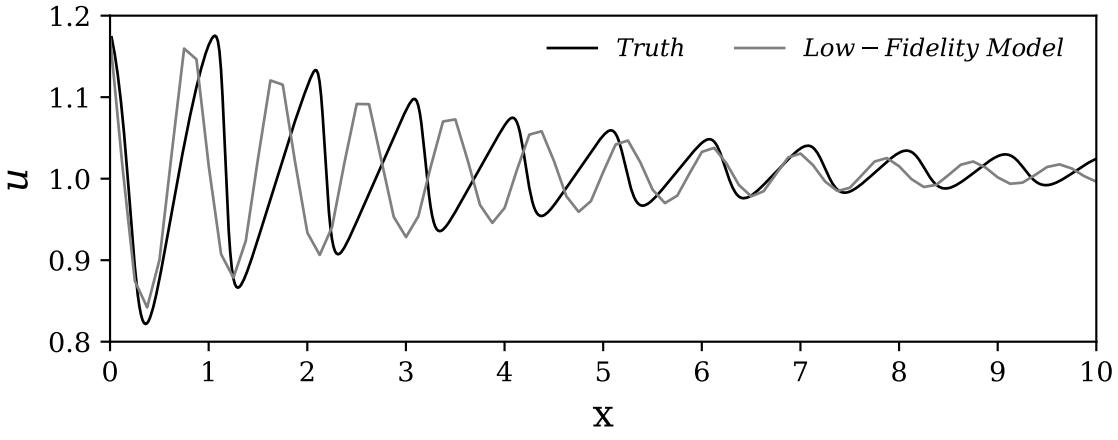


Figure 5-9: Instantaneous solution of the 1D Burgers' equation at $t = 100$. Solutions obtained via a very refined simulation (Truth, black line) and using a coarse grid (Low-Fidelity Model, gray line) are compared.

particular, the main source of numerical error appears to be of dispersive nature. More specifically, the time period of a full oscillation of the velocity field is significantly shorter when compared with the reference simulation. Diffusion errors are also visible, although their magnitude is smaller. The combination of these two sources of error severely affects the representation of the non-linear phenomena at play. In fact, in the reference simulation, one can see that non-linear dynamics are strong enough to sensibly deform the sinusoidal profiles imposed at the inlet. On the other hand, the state predicted via the low-fidelity model does not show marked deformations of the velocity profile, suggesting that non-linear effects are poorly represented.

In order to perform an extensive test of the performance of the MGEnKF strategy, two different runs are performed. The first one includes an outer loop only, while the second one performs the complete inner loop and outer loop scheme. This comparison allows assessing the impact of the inner loop performance over the optimization of the global coefficients of the simulation. Similarly to what was proposed in Section 5.2, the numerical scheme presented in (5.10) is modified to introduce *model correction* terms:

$$\begin{aligned} u_{\mathrm{j}}^{\mathrm{k}} =& u_{\mathrm{j}}^{\mathrm{k\text{-}1}} - u_{\mathrm{j}}^{\mathrm{k\text{-}1}} \frac{\Delta_t}{2\Delta_x} \left(u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} - u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}} \right) + \nu \frac{\Delta_t}{{\Delta_x}^2} \left(u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} - 2u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}} \right) + \mathcal{C}_{k:k-1}(u, \alpha, \gamma) \\ =& u_{\mathrm{j}}^{\mathrm{k\text{-}1}} - u_{\mathrm{j}}^{\mathrm{k\text{-}1}} \frac{\Delta_t}{2\Delta_x} \left(u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} - u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}} \right) + (1+\alpha)\, \nu \frac{\Delta_t}{{\Delta_x}^2} \left(u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} - 2u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}} \right) \\ & + \gamma \left(-u_{\mathrm{j\text{-}2}}^{\mathrm{k\text{-}1}} + 3u_{\mathrm{j\text{-}1}}^{\mathrm{k\text{-}1}} - 3u_{\mathrm{j}}^{\mathrm{k\text{-}1}} + u_{\mathrm{j+1}}^{\mathrm{k\text{-}1}} \right). \end{aligned} \tag{5.12}$$

The model $\mathcal{C}$ here introduced is composed by two correction terms, which are driven by the parameters $\gamma(x,t)$ and $\alpha(x,t)$. γ and α are identically zero in the main simulation of the MGEnKF, while they are optimized in the inner loop for the ensemble members. First, the α parameter controls a diffusive effect/numerical viscosity term. For this reason, local values are bounded to respect the condition, $\alpha(x,t) \geq -1$ i.e., non-physical solutions with negative global viscosity are excluded. Fig. 5-9 shows that grid coarsening is responsible for an over estimation of diffusive effects. Therefore,

one should expect to observe a convergence of the parameter $\alpha(x,t)$ towards negative values. The second model term $\gamma\left(-u_{\text{j-2}}^{\text{k-1}}+3u_{\text{j-1}}^{\text{k-1}}-3u_{\text{j}}^{\text{k-1}}+u_{\text{j+1}}^{\text{k-1}}\right)$ mimics the effects of a dispersion term of the form $\left(\gamma\Delta_x{}^3u_{xxx}\right)$ (Hirsch, 2007). This term was used in the previous section to correct the dispersive errors observed in the advection equation. The time evolution of $\alpha(x,t)$ and $\gamma(x,t)$ is taken into account by the MGEnKF itself, as the parameters are updated at each inner analysis phase. On the other hand, the space variability of the two parameters is obtained by expressing them in terms of a Legendre Polynomial expansion. Similarly to what was done for the linear advection case presented in Section 5.2, the expansion is truncated to $n=4$, i.e., a 4-th polynomial order. This implies that the optimization performed in the *inner loop* targets the value for the ten model coefficients γ_{i} and α_{i}.

The performance of the estimators (inner and outer loop, outer loop only) is assessed via the following data-assimilation experiment:

- The observations are sampled each 160 time steps of the reference simulation run on the space domain $[3,4]$ (64 sensors) and on the time window $[0,240]$. Considering the value of the time step Δt employed for the investigation, this implies that approximately 20 analysis phases per characteristic time evolution ΔT are performed. The sampling of the reference simulation is performed starting from a *fully developed* state for $t=0$. This is easily done owing to the periodic characteristics of the inlet. For the sake of simplicity, we assume that the observations and the coarse-grid ensemble are represented on the same space. Therefore, $\mathcal{H}_{\text{k}}^{\text{o}}\equiv\mathcal{H}^{\text{o}}$ is a sub-sampling operator independent of time retaining only the points comprised in the coarse space domain $[3,4]$. The observations are artificially perturbed by adding a constant in time Gaussian noise of diagonal covariance $\mathbf{R}_{\text{k}}^{\text{o}}\equiv\mathbf{R}=4\cdot10^{-4}\mathbf{I}^{\text{o}}$.

- The *model* realizations consist of a main simulation (run on the same mesh used for the reference simulation) and an ensemble of $N_{\text{e}}=100$ coarse simulations (8 mesh elements per wavelength L_0) which are run using the numerical scheme in equation (5.12). The initial condition $u(x,t=0)=u_0$ is imposed for each sim-

ulation. The outer loop provides an optimization for the value of the parameter θ driving the inlet condition. The initial condition for this parameter for each simulation is provided in the form of a Gaussian distribution $\theta \sim \mathcal{N}(0.15, \mathbf{Q}_\theta)$, with $\mathbf{Q}_\theta(t = 0) = 6.25 \cdot 10^{-4}\mathbf{I}$. As previously stated, the inner loop of the MGEnKF optimizes the polynomial expansion coefficients α_i and γ_i controlling the behaviour of the model terms introduced in the dynamic equations. Here, the physical state predicted by the main simulation is used as surrogate observation for this step. Initial values of the coefficients are described by Gaussian distributions $\alpha_\mathrm{i} \sim \mathcal{N}(0, \mathbf{Q})$, $\gamma_\mathrm{i} \sim \mathcal{N}(0, \mathbf{Q})$ with $\mathbf{Q}(t = 0) = 9 \cdot 10^{-8}\mathbf{I}$ and $i = 0, 1, 2, 3, 4$. The surrogate observation is here represented by the projection of the complete state predicted on the fine-grid by the main simulation over the coarse grid. Therefore, $\mathcal{H}^{\mathrm{SO}}_\mathrm{k} \equiv \mathcal{H}^{\mathrm{SO}}$ is a subsampling operator independent of time retaining all the points comprised in the coarse space domain $[0, 10]$. The surrogate observation sampled from the fine-grid is further randomized using a constant in time Gaussian noise of covariance $\mathbf{R}^{\mathrm{O}}_\mathrm{k} \equiv \mathbf{R}^{\mathrm{O}} = 1 \cdot 10^{-8}\mathbf{I}$ in order to improve the convergence of the inner loop optimization procedure. As in Sec. 5.2, one can see that $\mathbf{R}^{\mathrm{O}}$ is orders of magnitude smaller than the observation covariance matrix $\mathbf{R}$. The inner loop only optimizes values for α_i and γ_i and no update of the state estimation is here performed. Also, the inner loop is not performed at each outer loop analysis phase, but it is instead performed around twice per characteristic time (i.e., once every ten outer loop analyses).

The time-evolution of the estimation for the parameter θ is shown in Fig. 5-10. The accuracy of the complete MGEnKF scheme is excellent, while a significant error ($\theta = -0.2$) is observed for the simplified MGEnKF using the outer loop only. The phase error due to grid coarsening observed in Fig. 5-9 for the model is responsible for this important mismatch, and the reason is clear when analysing the instantaneous physical state in Fig. 5-11. In fact, the cumulative loss of phase of the model in the region $3 \leq x \leq 4$, which includes the observation, is approximately π. Thus, this error induces a bias in the estimation of the inlet parameter θ, which compensates the error in the observation region but provides massive errors outside of it.

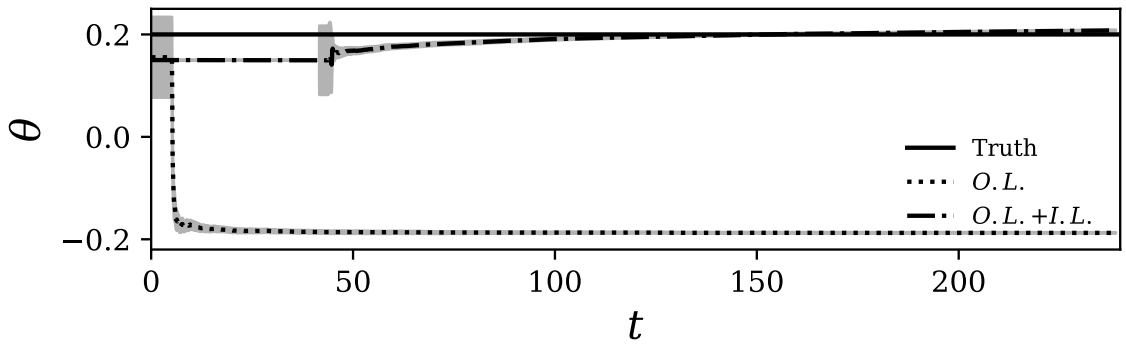


Figure 5-10: Evolution in time of the parameter θ during the outer loop optimization via MGEnKF. Results obtained from the DA complete model (outer plus inner loop, dot-dashed line) and the simplified DA model (outer loop only, dotted line) are compared with the exact result (black line).

On the other hand, the optimization via inner loop of the coefficients α_{i} and γ_{i} allows obtaining a precise representation of the flow field in the whole physical domain and not only in the observation region. The physical state obtained in Fig. 5-11 by the complete MGEnKF scheme is in better agreement with the *truth*, and the non-linearity of the flow is adequately captured despite the significant difference in resolution between the reference simulation and the ensemble members.

The performance of the inner loop is now investigated via the analysis of the model parameters α and γ. In figure 5-12 and 5-13 spatial distributions for the two parameters at $t = 240$ are shown. It will also be shown later that time variations are weak for these quantities, so the results presented can be considered as mean values for α and γ as well. As expected, α exhibits negative values to compensate the higher numerical diffusion due to the coarser grid. However, the condition $\alpha > -1$ is strictly respected in the whole physical domain. The spatial distribution for γ is quasi constant and equal to ≈ 0.004. For the linear advection equation, if $\gamma = \sigma\left(1-\sigma^2\right)/6$, one obtains a third-order accurate scheme on the support $j-2, j-1, j, j+1$, namely the Warming, Kutler, Lomax scheme (Hirsch, 2007). Considering the values for u_0, Δx and Δt used for this analysis, $\gamma = 0.025\left(1-0.025^2\right)/6 \approx 0.004$. Thus, the optimized value for γ obtained via the inner loop is close to the value that provides maximum accuracy in the linear advection case.

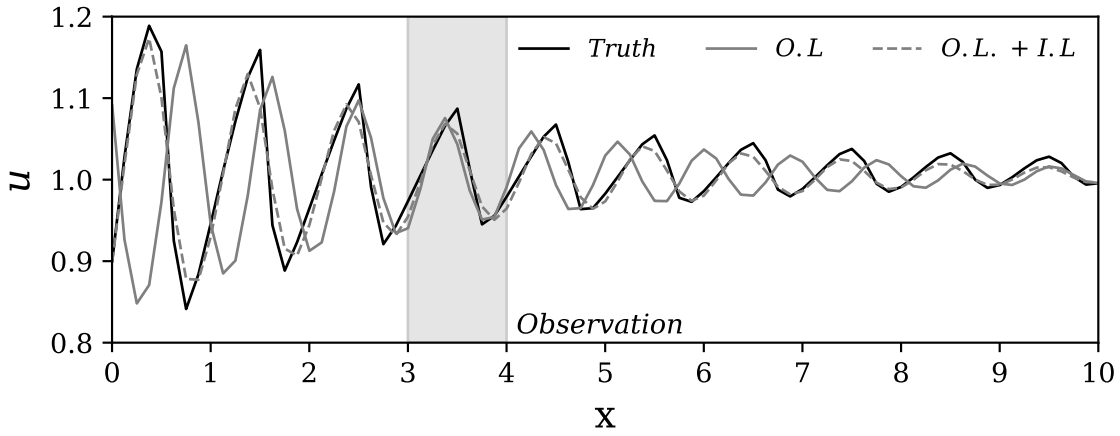


Figure 5-11: State estimation results for the 1D Burger's test case, projected on the coarse grid for $t = 240$. The projected true state (black line) is compared with results obtained via the MGEnKF complete model (outer plus inner loop, gray dotted line) and the simplified MGEnKF model (outer loop only, gray line).

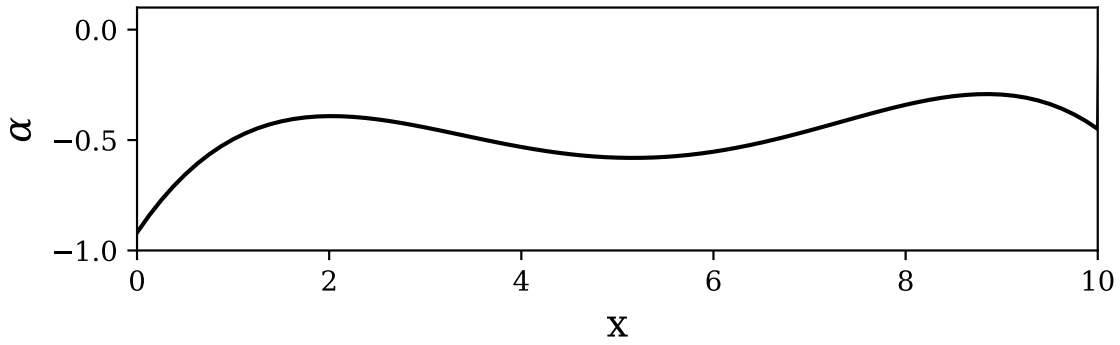


Figure 5-12: Instantaneous space distribution of the model parameter α determined via inner loop optimization. The results shown correspond to a simulation time of $t = 240$.

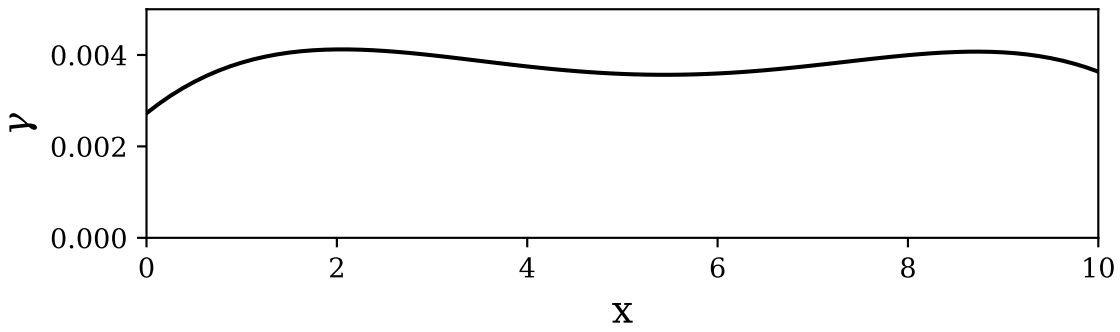


Figure 5-13: Instantaneous space distribution of the model parameter γ determined via inner loop optimization. The results shown correspond to a simulation time of $t = 240$.

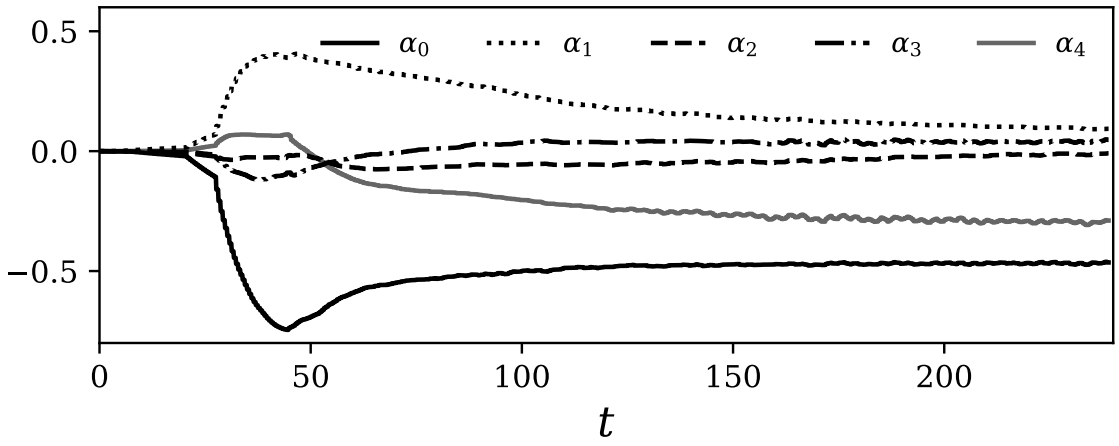


Figure 5-14: Estimation history of the Legendre Polynomial coefficients α_{i}.

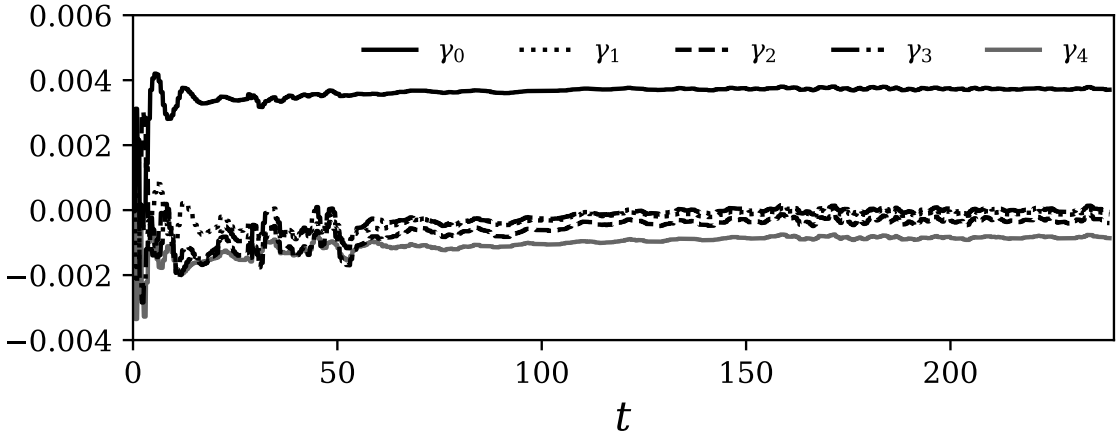


Figure 5-15: Estimation history of the Legendre Polynomial coefficients γ_{i}.

More information about the numerical models can be drawn by the analysis of the time evolution of the coefficients α_{i} and γ_{i} in Fig. 5-14 and 5-15. First, one can see that the zero-order contributions α_0 and γ_0 are the most important in terms of magnitude. In addition, one can also see a small but non-zero evolution in time of such coefficients. This observation is tied to the interactions between the optimization procedures performed in the inner loop and in the outer loop, whose results interact for non-linear phenomena. This is also the reason the rate of convergence of α, which is strictly connected to θ, is slower.

Multiple strategies for the complete MGEnKF scheme have been tested, varying the starting time of the inner and outer loop. It has been observed that the global optimization converges faster, and it is more robust if a first phase using the inner loop only is followed by a second phase where both inner loop and outer loop are applied. This procedure allows to *train* the model used for the ensemble members to perform similarly to the model employed on the fine grid for the main simulation. Therefore, the following classical DA optimization represented by the outer loop converges more rapidly to the targeted behaviour provided by the observation. This initial phase of training, which has been here performed for $t \in [0, 40]$, is particularly important if the values of the parameters driving the correction model used for the ensemble members are unknown, like for the present analysis.

5.4 Acoustic propagation of sinusoidal wave

The MGEnKF strategy is now applied to a more complex physical system, namely the inviscid one-dimensional Euler equations:

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho u)}{\partial x} = 0 \tag{5.13}$$

$$\frac{\partial (\rho u)}{\partial t} + \frac{\partial ((\rho u) u)}{\partial x} + \frac{\partial p}{\partial x} = 0 \tag{5.14}$$

$$\frac{\partial \left(\rho E\right)}{\partial t} + \frac{\partial \left(\left(\rho E\right) u\right)}{\partial x} + \frac{\partial (p u)}{\partial x} = 0 \tag{5.15}$$

where ρ is the density, u is the velocity, p is the pressure and E is the total energy per unit mass. In this case, viscous effects are absent, but acoustic propagation affects the evolution of the flow. The equations are discretized in space using a second-order finite difference centred scheme. A first-order explicit Euler scheme is used for the time integration method. After discretization, a representation similar to (2.39) is obtained. A centred sixth-order numerical filter is included to damp numerical spurious oscillations (Bogey & Bailly, 2004). We specifically analyse the acoustic propagation of a sinusoidal wave with a time-varying amplitude. To do so, a Dirichlet time-varying velocity condition is imposed at the inlet:

$$u(x=0,t) = u_0\left(1+\theta(t)\sin(2\pi f_c t)\right) \tag{5.16}$$

The value of u_0 is set in order to impose an inlet Mach number $M = \frac{u_0}{a} = 0.4$, where a is the speed of the sound. The amplitude of variation in θ is sufficiently low to allow a flow evolution mainly driven by acoustic phenomena. The inlet velocity perturbation creates an acoustic wave that is transported along the domain with a speed equal to $u_0 + a$. The characteristic velocity and length scales are $u_c = u_0 + a$ and L_c, which is the wavelength of the signal imposed at the inlet. The characteristic time of the system is defined as $t_c = L_c/u_c$.

The sinusoidal behaviour of the velocity at the inlet is characterized by a constant frequency $f_c = 1/t_c$. However, the amplitude of the sinusoidal wave is driven by the time-varying parameter $\theta(t) = \theta_0\left(1+\sin\left(2\pi\frac{f_c}{10}t\right)\right)$, where θ_0 is a constant. At the inlet, we set $\rho(x=0,t) = \rho_0$ and $E(x=0,t) = E_0 = e + 0.5u_0^2$, where e is the internal energy per unit of mass. By definition, $e = C_v T_0$ where C_v is the heat capacity at constant volume and T_0 the initial temperature of the flow. The outlet boundary condition is extrapolated from the nearest points to the outlet using 4-th order Lagrange polynomials. The initial condition imposed at $t = 0$ is $u(x, t=0) = u_0$, $\rho(x, t=0) = \rho_0$ and $E(x, t=0) = E_0$ everywhere in the physical domain and for all the simulations (fine-grid and coarse-grid ensemble members). The fluid is considered an ideal gas with $C_v = 0.7171$, $\gamma = 1.4$, $\rho_0 = 1.17$ and $T_0 = 300$ in

S.I. units.

The computational domain has been set to a size of $L_x = 10$. A uniform mesh distribution is used for every calculation. 80 mesh elements are used to discretize the characteristic length L_c for a total of $N_x = 800$ elements in the domain. Finally, the normalized value of Δt is set to $\Delta t = 0.0006$.

A preliminary simulation is performed for $\theta_0 = 0.015$ (true state). A flow visualization of the wave patterns is shown in Fig. 5-16 at $t = 17.3$. The fully developed state obtained at $t = 10$ is used to initialize a new simulation from $t = 0$. This simulation is run for a total time of $T_{\mathrm{DA}} = 110$. The fine grid data are projected on the coarse grid and sampled to obtain observations of ρu in the space region $x \in [0, 1]$. These observations are artificially perturbed using a Gaussian noise of variance $\mathbf{R}^{\mathrm{o}} = 0.09\mathbf{I}$. The observation operator $\mathcal{H}^{\mathrm{o}}$ selects the points in the coarse grid where ρu is available.

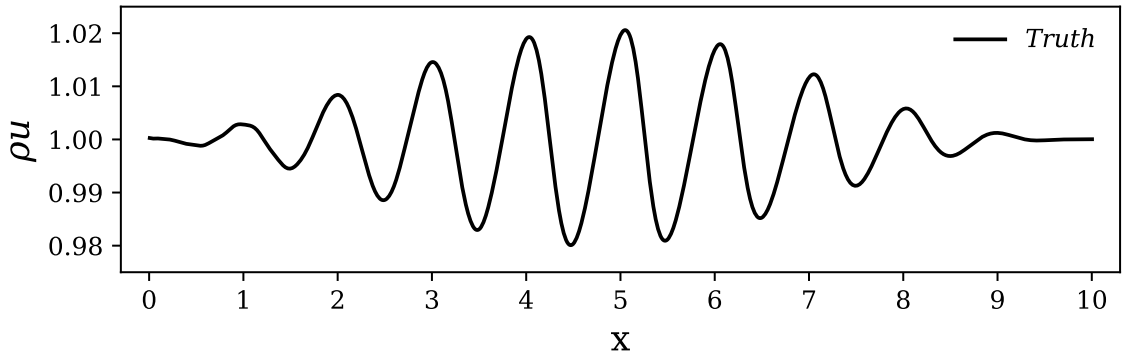


Figure 5-16: Solution ρu of the inviscid one-dimensional Euler equations at $t = 17.3$ for $\theta_0 = 0.015$ (*true* state)

.

In contrast to the DA experiment undertaken in Sec. 5.2 and Sec. 5.3, in Sec. 5.4 and Sec. 5.5 the analysis is focused on the performance of the *outer loop*. This choice was performed due to the central contribution of this loop to the global data assimilation strategy. For this reason, the *inner loop* was suppressed in order to obtain an unambiguous assessment of this element of the MGEnKF. Moreover, the models used in these last two sections are significantly more complex than the prototype Burgers and advection equations used in the first DA experiments performed; they include a coupling between velocity and pressure which complicates the determination

of the model correction term $\mathcal{C}$. This aspect is still under research.

The assimilation is composed of the main simulation, which is run on the fine grid, and $N_{\mathrm{e}} = 100$ simulations on the coarse-grid level. In this section, a coarsening ratio ($r_{\mathrm{C}} = \Delta x^{\mathrm{F}}/\Delta x^{\mathrm{C}}$) of $r_{\mathrm{C}} = 4$ is considered. The estimator is used to dynamically track the value of the parameter θ, which evolves in time. No a priori knowledge about the behaviour of the parameter is used. A similar analysis using a classical Kalman smoother was recently proposed by Mons et al., 2016.

For each coarse grid simulation of the estimator, θ is initially assumed to be a random Gaussian phenomenon $\theta \sim \mathcal{N}(0, \mathbf{Q}_\theta)$. The initial value of the covariance is $\mathbf{Q}_\theta(t = 0) = 6.4 \times 10^{-5}$. The value initially imposed at the inlet of the fine-grid simulation is $\theta = 0$, while random values are selected for each ensemble member on the coarse grid level following the normal distribution introduced above. The variance of the parameter θ for the ensemble members is artificially increased, as in the classical Dual EnKF algorithm. As described in A.3, we add to the estimated parameter of each member of the ensemble a Gaussian noise of covariance $\mathbf{\Sigma}_k^\theta = 10^{-10}\mathbf{I}$, which is reminiscent of the strategy used by Moradkhani et al., 2005. Extensive numerical tests have been performed and a sensitivity of the results to $\mathbf{\Sigma}_k^\theta$ has been observed. The value chosen for $\mathbf{\Sigma}_k^\theta$ has been set to avoid the collapse of the ensemble members over the state estimation while keeping the noise level for θ moderately low.

Let f_a be the number of analysis phases per characteristic time of simulation t_c, *i.e.* $f_a = t_c/t_a$. In the following, three different values of f_a are investigated: $f_a = 2$, 10, 55.

The estimator is run for a total simulation time $T_{\mathrm{DA}} = 110$, which encompasses 220 to 6000 DA analysis phases, depending on the value of f_a. At the end of each analysis, the mean value and the variance of the amplitude θ are updated following the Dual EnKF technique.

The results for the estimation of the time-varying parameter θ are reported in Fig. 5-17. For clarity, only the results for the assimilation window $t \in [40, 70]$ are shown. The precision of the parametric inference is measured via the relative error η defined as $\eta_k = \dfrac{\overline{\theta_k^a} - \theta_k^{\mathrm{true}}}{\max_k |\theta_k^{\mathrm{True}}|}$, where $\overline{\theta_k^a}$ is the average (first central moment) of the

Gaussian random value θ and θ_k^{True} is the *true* value of θ.

The time evolution of θ is correctly estimated for the three values of f_a. This is an important result, considering that no a priori information was provided for the evolution of this parameter. A more detailed analysis reveals a lag in the parameter estimation. The application of a simple Kalman filter seems to be responsible for this result, while a Kalman Smoother (KS) should have been used to obtain a better synchronization. However, considering that the implementation of a KS is straightforward in this case and that observation is always provided close to the inlet, we considered that the increase in computational resources required by the KS was not needed. We find that the lag increases when a relatively small number of DA analyses is done. One can see that the prediction is significantly degraded for $f_a = 2$, while similar results are obtained for $f_a = 10, 55$. This observation is quantified by the time evolution of the relative error η, which is significantly large for $f_a = 2$. In addition, θ tends to be generally underestimated (around $10 - 20\%$) when it reaches its maximum value. This result is arguably associated with the under-resolution of the coarse level of the grid, where the gradients of physical variables are calculated with lower accuracy.

Now, results dealing with the state estimation are discussed. The predicted physical variable ρu, normalized over the initial value $\rho_0 u_0$, is shown in Fig. 5-18, 5-19 and 5-20 for $f_a = 2, 10$ and 55, respectively. For $f_a = 2$, the state estimation is significantly distant from the truth. It appears that the field correction applied via the Kalman gain is not able to compensate the poor estimation of θ. However, accurate results are observed for $f_a = 10$ and 55. Even though the value of the parameter θ is not exact, the state estimation including the correction via Kalman gain is very precise. For the case $f_a = 55$, almost no discernible difference is observed between the state estimation and the truth.

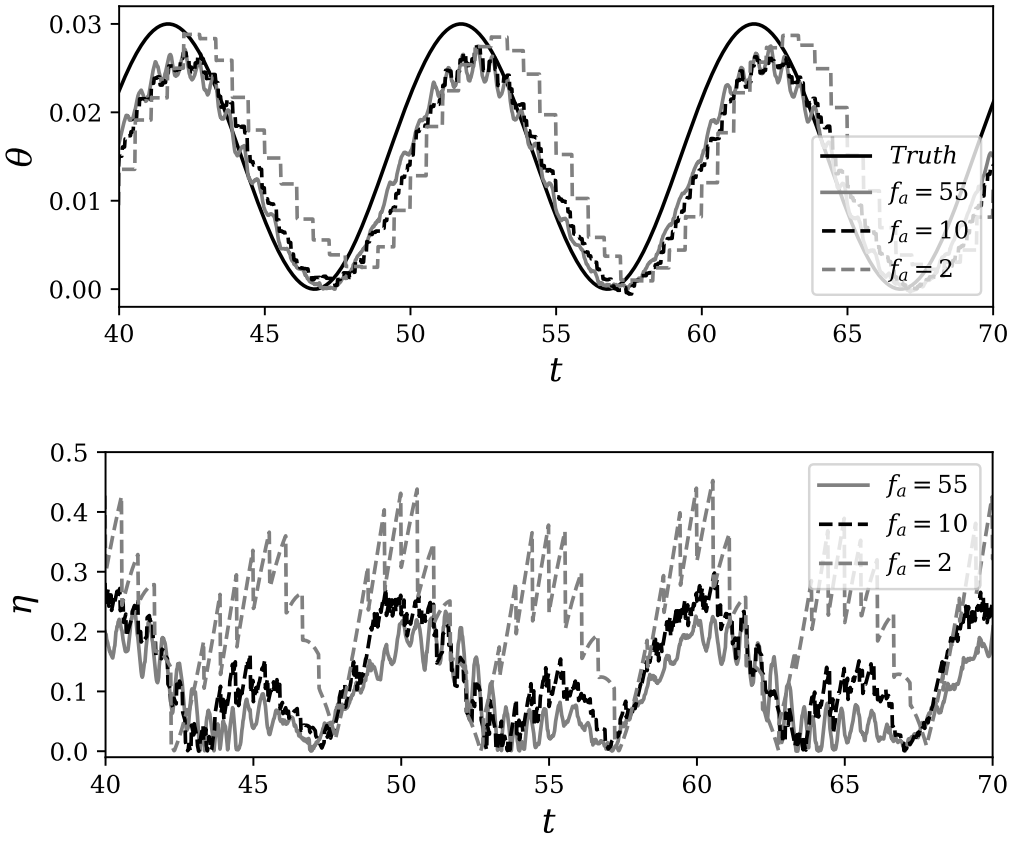
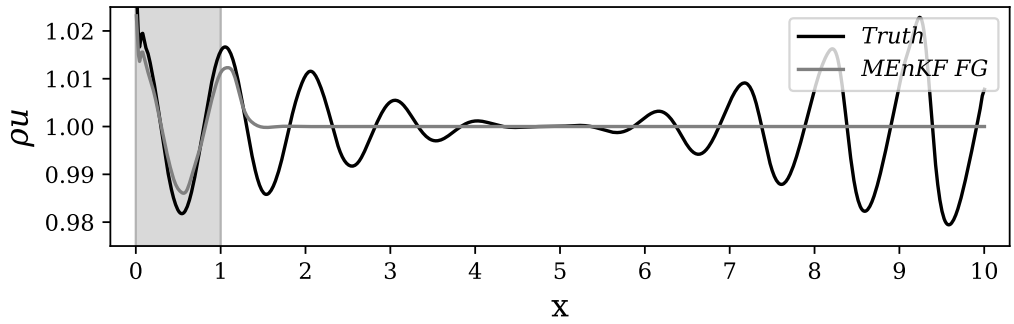
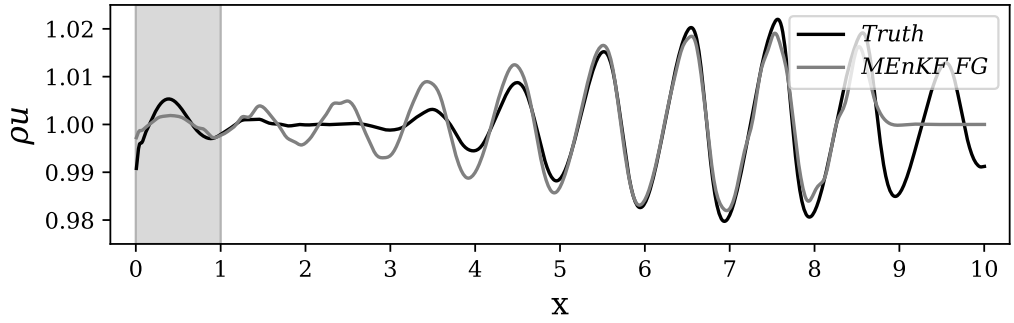


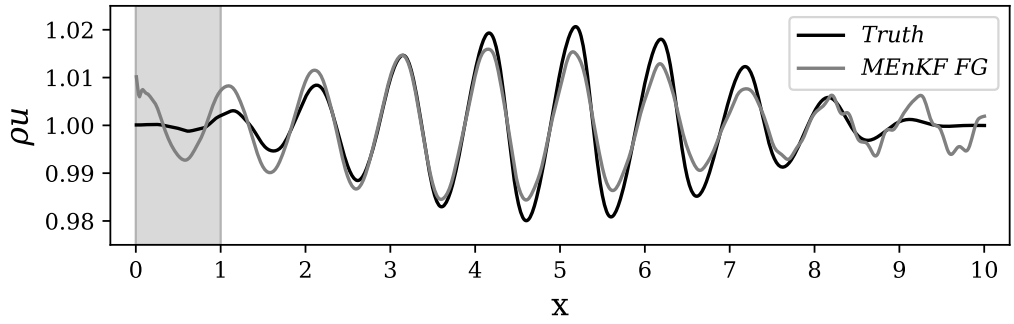
Figure 5-17: Time estimation of the parameter θ driving the amplitude of the sinusoidal acoustic wave for the assimilation window $t \in [40, 70]$. In the top image, results are shown for $f_a = 2, 10, 55$ and compared to the true value of θ. In the bottom image, the relative error η quantifying the parametric inference is shown.



(a) $t = 1.23$

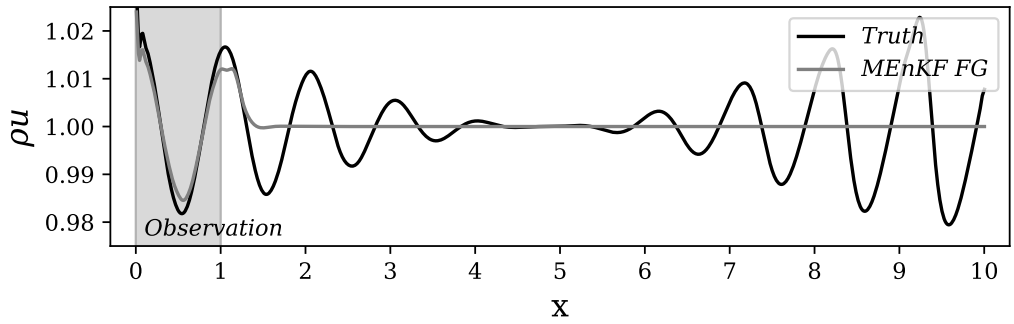


(b) $t = 8.32$

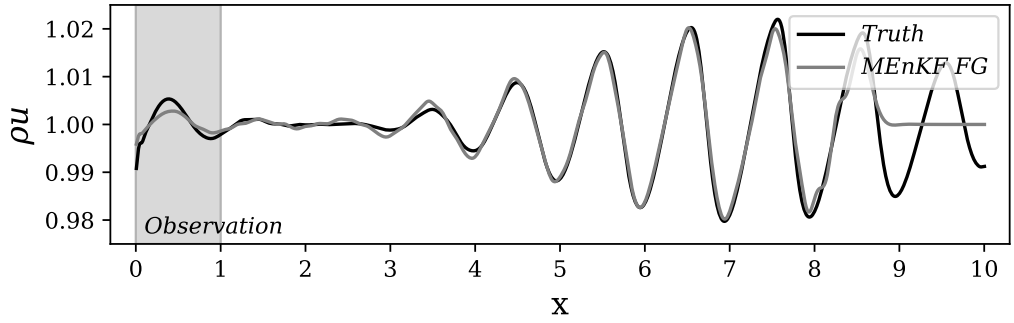


(c) $t = 16.30$

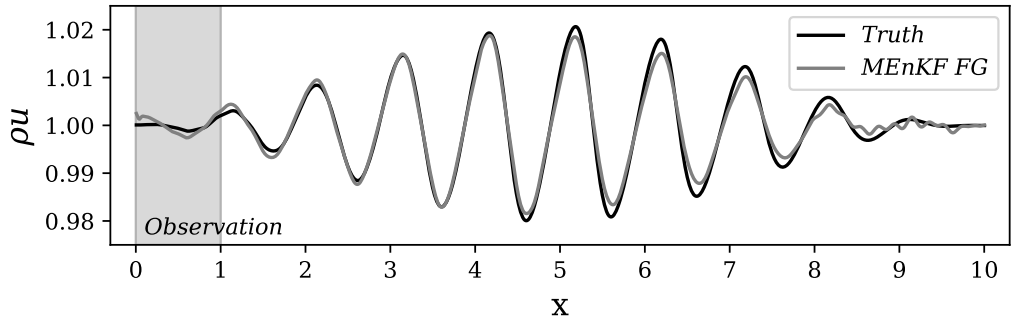
Figure 5-18: Estimations by MGEnKF of the momentum ρu normalized by $\rho_0 u_0$ for $f_a = 2$ at $t = 1.23$ (a), $t = 8.32$ (b) and $t = 16.30$ (c). Times are given in t_c units. The grey shaded area corresponds to the observation window.



(a) $t = 1.23$

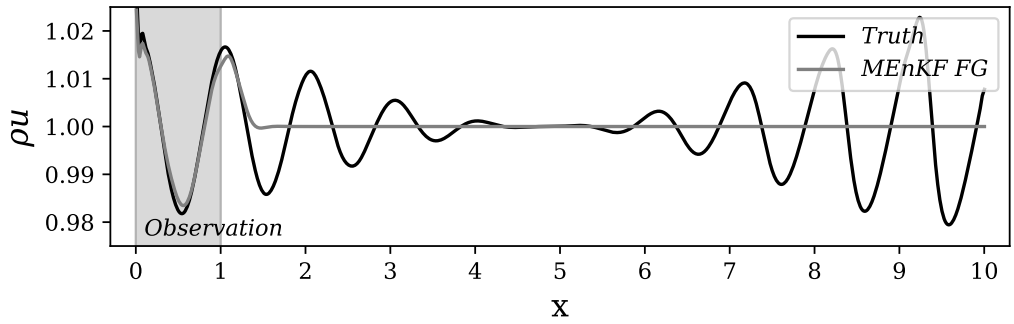


(b) $t = 8.32$

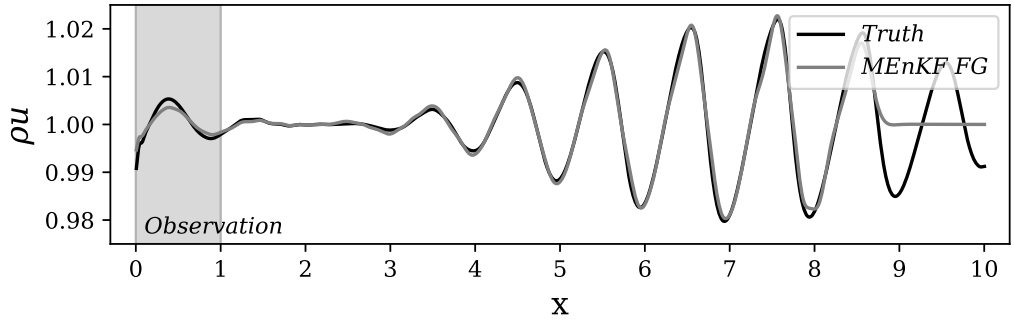


(c) $t = 16.30$

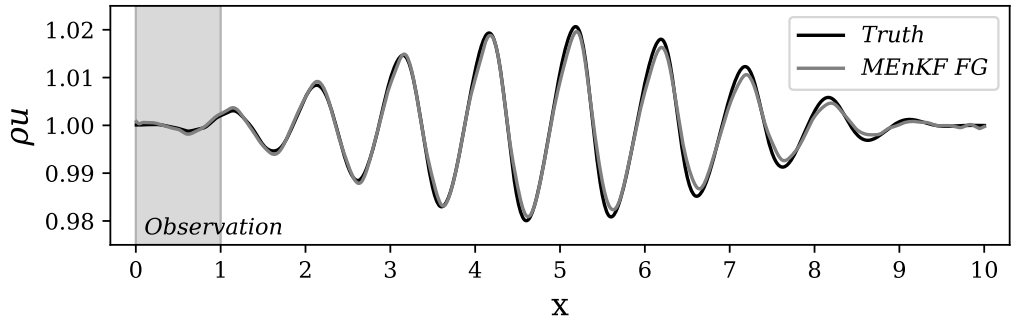
Figure 5-19: Estimations by MGEnKF of the momentum ρu normalised by $\rho_0 u_0$ for $f_a = 10$ at $t = 1.23$ (a), $t = 8.32$ (b) and $t = 16.30$ (c). Times are given in t_c units. The grey shaded area corresponds to the observation window.



(a) $t = 1.23$



(b) $t = 8.32$



(c) $t = 16.30$

Figure 5-20: Estimations by MGEnKF of the momentum ρu normalized by $\rho_0 u_0$ for $f_a = 55$ at $t = 1.23$ (a), $t = 8.32$ (b) and $t = 16.30$ (c). Times are given in t_c units. The grey shaded area corresponds to the observation window.

At last, the relative Root Mean Square Error (RMSE) defined as

$$\mathrm{RMSE}(k) = \sqrt{\frac{\int_x \left[\left((\rho u)_k^{\mathrm{F}}\right)^{\mathrm{a}}(x) - \left((\rho u)_k^{\mathrm{F}}\right)^{\mathrm{True}}(x)\right]^2 \mathrm{d}x}{\int_x \left[\left((\rho u)_k^{\mathrm{F}}\right)^{\mathrm{True}}(x)\right]^2 \mathrm{d}x}} \tag{5.17}$$

is shown in Fig. 5-21. The error achieves a quasi-constant asymptotic behaviour after a complete propagation of the signal in the physical domain ($t \approx 10t_c$). As expected, a low global error is obtained for the cases $f_a = 10$ and $f_a = 55$. On the other hand, the error for the $f_a = 2$ case is around $2 - 3$ times larger. The very small difference in performance between the cases $f_a = 10$ and $f_a = 55$ can be interpreted as a sign of convergence of the procedure. Another interpretation may be that the leading contribution of the error corresponds to the statistical error from using only 100 ensemble members in the estimation. We can however note that similar results were previously observed in three-dimensional simulations Meldi, 2018.

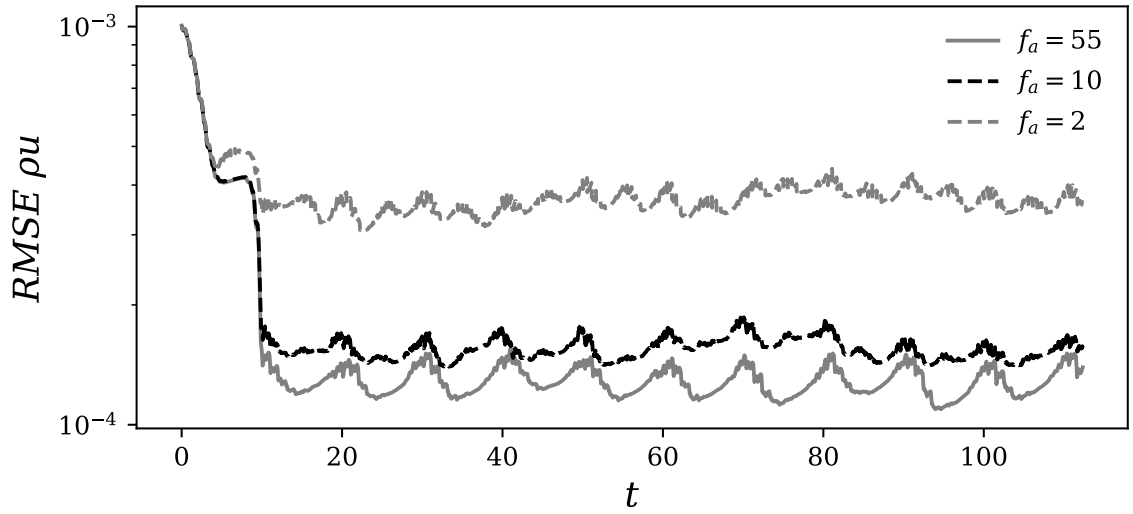


Figure 5-21: Time evolution of the RMS error of ρu for $f_a = 2$, 10, 55.

5.5 Spatially evolving compressible mixing layer

In this section, we consider the compressible Navier-Stokes equations in a two-dimensional physical domain:

$$\frac{\partial \rho}{\partial t} + div(\rho \mathbf{u}) = 0 \tag{5.18}$$

$$\frac{\partial (\rho \mathbf{u})}{\partial t} + div(\rho \mathbf{u} \otimes \mathbf{u}) = -\mathbf{grad}\, p + \mathbf{div}\, \overline{\overline{\tau}} \tag{5.19}$$

$$\frac{\partial (\rho E)}{\partial t} + div(\rho E \mathbf{u}) = -div(p\mathbf{u}) + div(\overline{\overline{\tau}}\mathbf{u}) + div\,(\lambda(T)\mathbf{grad}\, T) \tag{5.20}$$

where ρ is the density, $\mathbf{u}$ is the velocity (components u in the streamwise direction and v in the normal direction), p is the pressure, E is the total energy per unit of mass, $\overline{\overline{\tau}}$ is the tensor of the viscous constraints and T is the temperature. To obtain the representation given by (2.39), the equations are discretized using the finite difference method. Second-order centred schemes are used for the derivatives in space and a first-order scheme for the time integration. A centred sixth-order numerical filter is included to damp numerical spurious oscillations (Bogey & Bailly, 2004).

The two-dimensional spatially evolving mixing layer at Re = 100 is here investigated. For this value of Reynolds number, the flow exhibits unsteady features. It can be shown (Ko et al., 2008; Meldi et al., 2012) that the characteristics of the mixing layer are strongly affected by the inlet and, in particular, by imposed *ad hoc* time perturbations. The computational domain has been set to a size of $14L_c \times 6L_c$ in the streamwise direction x and normal direction y, respectively. The characteristic length L_c, which is taken as reference length from now on, is given by $L_c = A\delta_0$, where δ_0 is the initial vorticity thickness imposed at the inlet. The value of the parameter A is set in order to represent the most unstable wavelength determined by Linear Stability Theory (LST). At Re = 100, we have $A = 14.132$ (McMullan et al., 2007). The mesh resolution in the horizontal direction is constant for $x \in [0; 10]$. The size of the elements is $\Delta x = \frac{\delta_0}{8}$. For $x \geq 10$, a sponge zone is established with a coarsening ratio between successive elements which increases from 1.025 to 1.04. The resolution in the normal direction is constant and equal to $\Delta y = \frac{\delta_0}{20}$ for $-0.18 \leq y \leq 0.18$. Outside

this zone, the mesh elements increase in size moving away from the centreline with a constant coarsening ratio of 1.01.

The Reynolds number of the flow is calculated as $\mathrm{Re} = (U_1 - U_2)\delta_0/\nu = 100$ with asymptotic velocities set to $U_1 = 173.61$ and $U_2 = 104.17$. These values correspond to a Mach number $\mathrm{Ma} = 0.5$ and $\mathrm{Ma} = 0.3$, for each stream, respectively. The kinematic viscosity and thermal diffusivity of the flow are considered to be constant, and their value is fixed to $\nu = 1.568 \times 10^{-5}$ and $\alpha = 22.07 \times 10^{-7}$, respectively. Finally, the flow is considered to be a perfect gas with $\gamma = 1.4$ and $C_v = 0.7171$. All these quantities are expressed in S.I. units. The inlet boundary condition is taken from Ko et al., 2008. For the velocity field, one has:

$$U_{\mathrm{in}}(y,t) = \frac{U_1 + U_2}{2} + \frac{U_1 - U_2}{2} \tanh\left(\frac{2y}{\delta_0}\right) + U_{\mathrm{Pert}}(y,t) \qquad -3 < y < 3 \tag{5.21}$$

$$V_{\mathrm{in}}(y,t) = 0 \tag{5.22}$$

where U_{in} is the streamwise velocity at the inlet and V_{in} is the normal velocity. U_{in} is estimated as a classical hyperbolic tangent profile plus a time-varying perturbation component:

$$U_{\mathrm{Pert}}(y,t) = \sum_{i=1}^{N_{\mathrm{in}}} \epsilon_i(t) \frac{U_1 + U_2}{2} [f_i(y) \sin(\omega_i t)], \tag{5.23}$$

where N_{in} is the total number of perturbation modes and ϵ_i quantifies the magnitude of each mode. The function $f_i(y) = \cos\left(4n_i \dfrac{y}{\delta_0}\right) h(y)$ controls the shape of the perturbation of the inlet velocity profile in the normal direction. The role of $h(y) = 1 - \tanh^2\left(\dfrac{2y}{\delta_0}\right)$ is to damp the perturbation component moving away from the centreline. The wavelength parameters n_i are tuned according to the LST results. In the following, we consider $N_{\mathrm{in}} = 1$, *i.e.*, the inlet perturbation consists of a single mode. In the numerical tests, we follow Ko et al., 2008 and set $n_1 = 0.4\pi$ and $\omega_1 = 1/t_c$, where $t_c = 2L_c/(U_1 + U_2)$ is the average advection time. The inlet density is set to be constant so that $\rho_{\mathrm{in}} = 1.177$, as well as the temperature $T_{\mathrm{in}} = 300$ in S.I. units. The inlet total energy per unit mass E_{in} is calculated as $E_{\mathrm{in}} = e + 0.5\,(U_{\mathrm{in}}^2 + V_{\mathrm{in}}^2)$, where the internal energy e is defined as $e = C_v T_{\mathrm{in}}$. The

outlet boundary conditions for all the variables present in the state vector are extrapolated from the nearest points to the outlet using 4-th order Lagrange polynomials. The zero gradient boundary condition is imposed on the transverse sides of the domain (at $y = -3$ and $y = 3$).

In this section, the parameter θ of the model corresponds to the single parameter ϵ_1, which is the time-dependent variable governing the amplitude of the perturbation. A reference simulation is run where ϵ_1 varies in time following a sinusoidal form: $\epsilon_1(t) = \epsilon(1 + \sin(\omega_\epsilon t))$. The values of the numerical parameters characterizing the perturbation are $\epsilon = 0.15$ and $\omega_\epsilon = 0.62\omega_1$. At $t = 0$, the variables of the fine grid and the coarse grid ensemble are initialized in the physical domain using the values imposed at the inlet. This implies that the physical quantities can exhibit initial variations in the y direction, while their value in the streamwise x direction is constant. For the case of the streamwise velocity, the velocity perturbation $U_{\text{Pert}} = 0$ at $t = 0$, see (5.23). A flow visualization of ρv at $t = 10$ is shown in Fig. 5-22 for this reference simulation. One can clearly observe the emergence of coherent structures with complex pairing patterns.

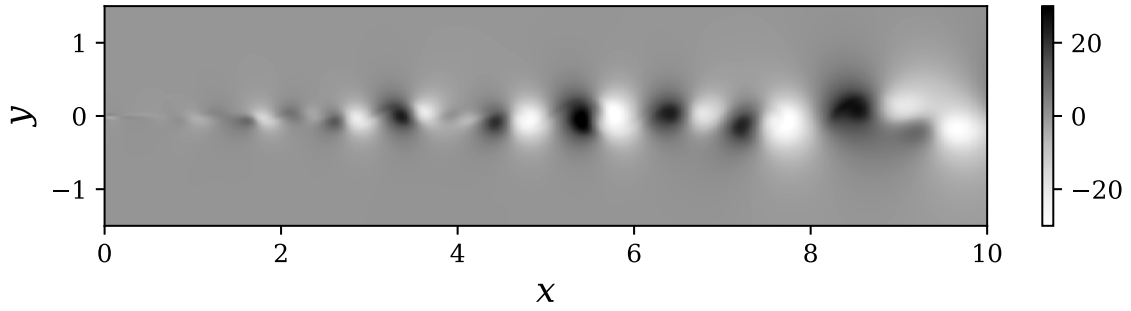


Figure 5-22: Visualization of the normal momentum ρv (S.I. units) for the 2D compressible Navier-Stokes equation. Reference simulation at $t = 10$ for a time-varying value of ϵ_1.

The DA procedure is performed using the following elements:

- The observation is sampled from the reference simulation shown in Fig. 5-22, which is run for a total simulation time of $T_{\text{DA}} = 40$ in t_c units. This value corresponds to four complete advections in the whole physical domain. A fully

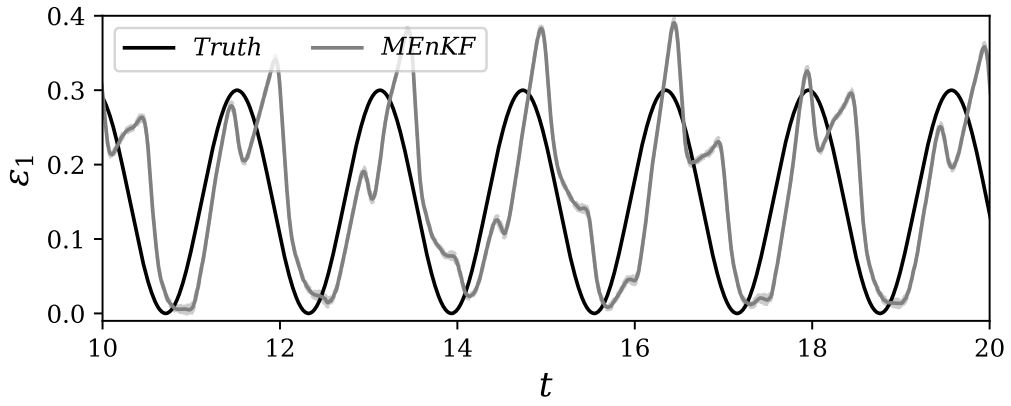
developed state obtained from a prior simulation at $t = 10$ is used to initialize the simulations at $t = 0$. Data are projected on the coarse grid and sampled every 30 time steps in the region $x \in [0, 0.55]$ and $y \in [-0.16, 0.16]$. Considering the results obtained in Sec. 5.4, the number of analysis phases per characteristic time of simulation ($f_a = 75$) is chosen sufficiently high to assure a good estimation. The observations are made from the instantaneous fields ρu and ρv. The data used as observation are artificially perturbed using a Gaussian noise of covariance $\mathbf{R}^{\text{o}} = \mathbf{I}$. The observation operator $\mathcal{H}^{\text{o}}$ acts as described in Sec. 5.4.

- The model is the discretized version of the system given by (5.18) - (5.20). The features of the fine mesh level were previously introduced. For the coarse grid level, a homogeneous coarsening ratio $r_{\text{C}} = 4$ is employed. The initialization strategy used for the coarse ensemble simulations is identical to the one described above for the fine-grid reference case. We consider that no prior information is available on the time evolution of ϵ_1. At $t = 0$, this coefficient is fixed to be a random Gaussian value $\epsilon_1 \sim \mathcal{N}(0, \mathbf{Q}_a)$ where the initial value of $\mathbf{Q}_a(t = 0) = 0.0625$. The value imposed on the main fine-grid simulation at $t = 0$ is $\epsilon_1 = 0$, while random values are imposed for each ensemble member on the coarse grid level. The size of the ensemble is $N_{\text{e}} = 100$.

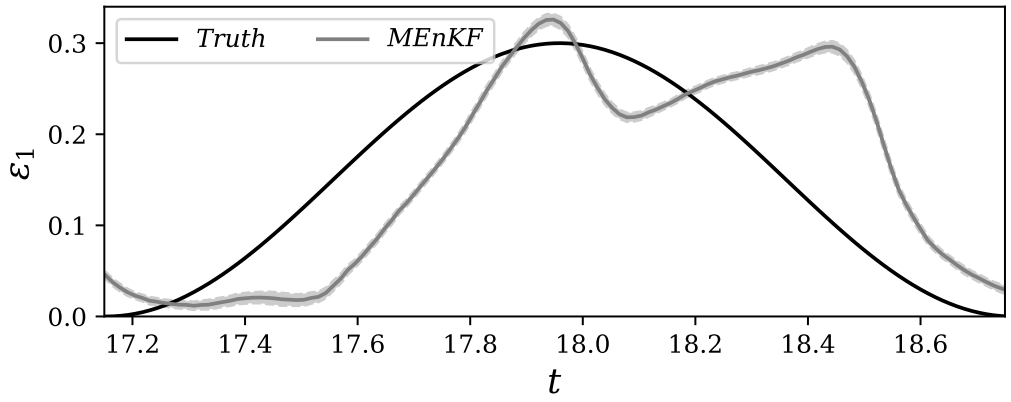
The estimation algorithm is run over a time window equal to $T_{\text{DA}} = 40$ which encompass roughly 3200 DA analysis phases. At the end of each analysis, the mean value and the variance of the coefficient ϵ_1 are updated following the Dual EnKF technique (Moradkhani et al., 2005), similarly to what was done in Sec. 5.4.

The time evolution of the estimated value of ϵ_1 is reported in Fig. 5-23. The overall sinusoidal trend is generally respected, although a relatively small phase lag is visible. This lag does not appear to be larger than the one previously observed for the one-dimensional case based on the Euler equation. This lag is due to the usage of a Kalman Filter approach instead of a Kalman Smoother, as discussed in Sec. 5.4. However, in this case, some overprediction of the parameter is locally observed in

time, which was not obtained for the wave propagation test case.



(a) Estimation history



(b) Zoom

Figure 5-23: Time evolution of the inferred values of ϵ_1 for the time-varying reference case. (a) Large time window. (b) Zoomed region. The shaded area represents the 95% credible interval for the estimated parameter.

The results obtained for the prediction of the normal momentum ρv are shown in Fig. 5-24. One can see that the combination of parameter and state estimations produces an accurate prediction of the flow. Minor discrepancies are observed when comparing the state estimation with the true state. In particular, the momentum ρv does not exhibit spurious oscillations which could stem from the field correction determined via the Kalman gain. In order to evaluate the respective influence of the parameter estimation step and state estimation phase, a test case is run in which only the parameter estimation is performed. That is, the state estimation obtained on the

coarse-grid level is not included in the steps 4 and 5 of the algorithm presented in Chapter 4. While the results of the parameter estimation are the same for the two cases, one can see in Fig. 5-25 that the prediction is sensibly deteriorated.

This observation is quantified by the evaluation of the relative Root Mean Square Error (RMSE), defined as:

$$\mathrm{RMSE}(k) = \sqrt{\frac{\int_x \left[\left((\rho v)_k^{\mathrm{F}}\right)^{\mathrm{a}}(x) - \left((\rho v)_k^{\mathrm{F}}\right)^{\mathrm{true}}(x)\right]^2 \mathrm{d}x}{\int_x \left[\left((\rho v)_k^{\mathrm{F}}\right)^{\mathrm{true}}(x)\right]^2 \mathrm{d}x}} \tag{5.24}$$

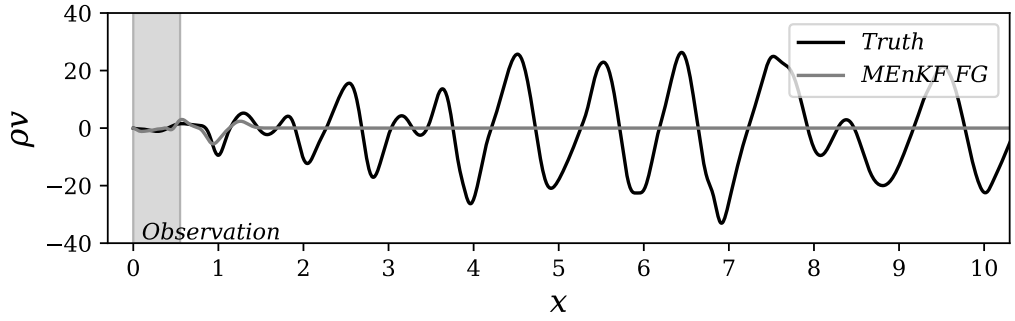
The results, which are shown in Fig. 5-26, indicate that the accuracy of the complete algorithm is higher when compared to the case in which only the parameter estimation is performed. Therefore, the two operations concurrently provide an improvement in the prediction of the flow.

At last, an analysis of the conservativity of the algorithm is performed. As previously discussed, the state estimation obtained via EnKF does not necessarily comply with the dynamical equations of the model. This drawback can be responsible for discontinuities in the physical field, which can significantly affect the accuracy and stability of the global algorithm. The analysis is performed considering an indicator Γ_k^{F} which measures the conservation of the transversal momentum equation (5.19) in discretized form:

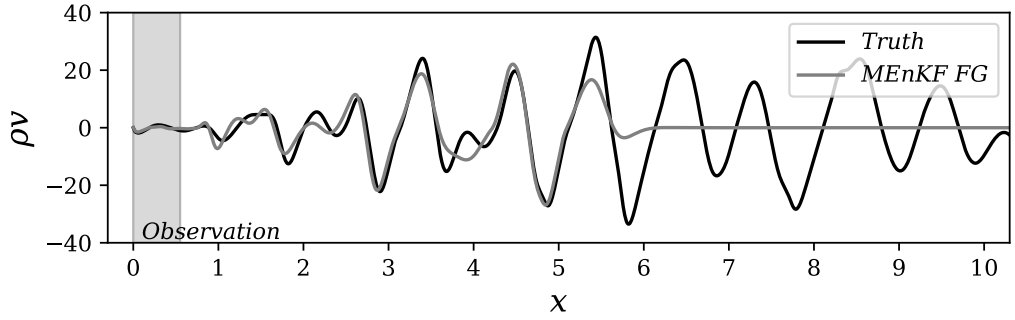
$$\frac{\left((\rho v)_k^{\mathrm{F}}\right)^{\mathrm{a}} - \left((\rho v)_{k-1}^{\mathrm{F}}\right)^{\mathrm{a}}}{\Delta t} - \mathcal{F}_{\rho v}\left(\rho_k, \mathbf{u}_k, p_k, \overline{\overline{\tau}}_k\right) := \Gamma_k^{\mathrm{F}}(x, y), \tag{5.25}$$

where $\mathcal{F}_{\rho v}$ represents the spatial discretization terms in the transversal momentum equation. In the forecast step performed via the model, $\Gamma_k^{\mathrm{F}} = 0$ down to a convergence rate δ which is prescribed. However, the value of Γ_k^{F}, at the end of a time step where a forecast-analysis is performed, is strictly connected with the computational strategy employed. Here, three scenarios are considered for a simplified case where the amplitude of the inlet-perturbation is assumed to be constant ($\epsilon_1 = \mathrm{ct} = 0.15$):

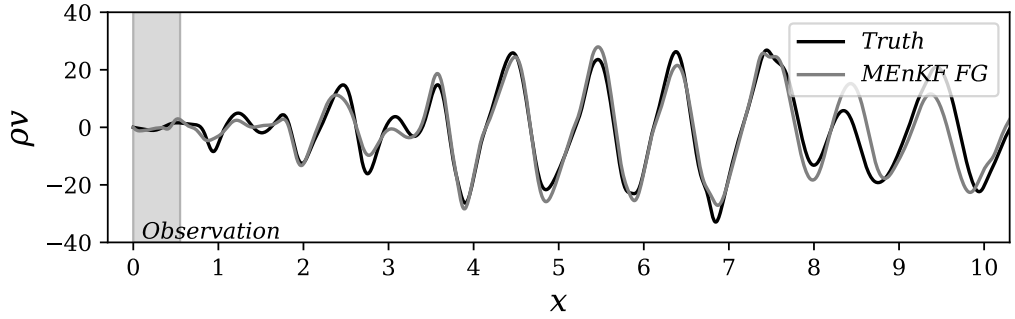
- S_1: A classical Dual EnKF is performed on the coarse grid and a fine-grid correction



(a) $t = 1$

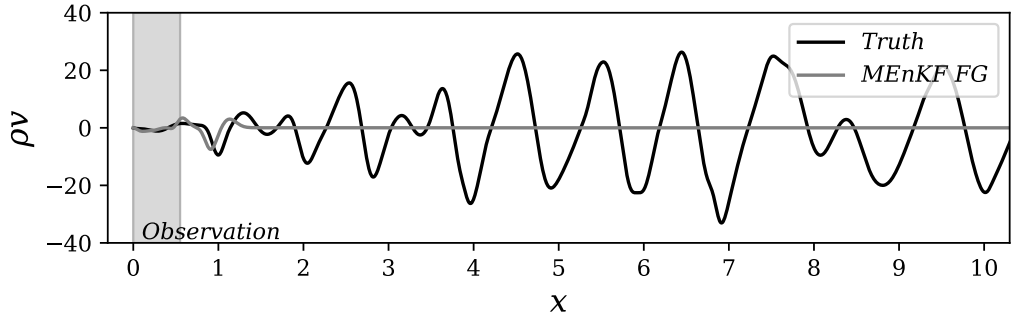


(b) $t = 5$

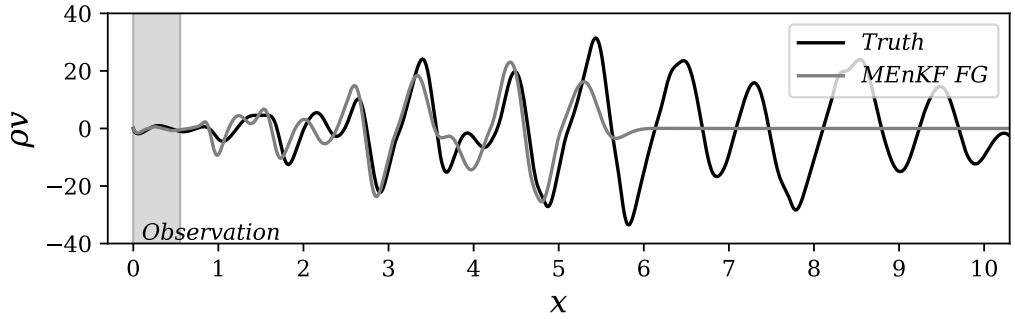


(c) $t = 30$

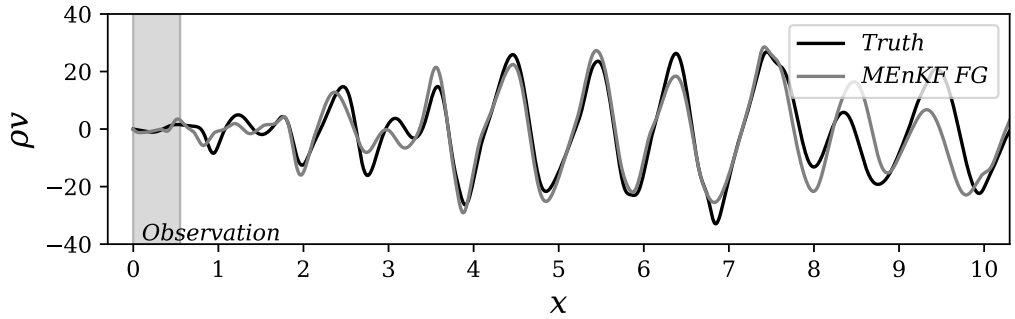
Figure 5-24: Estimations obtained by MGEnKF of the momentum ρv (S.I. units) at the centreline $y = 0$ of the mixing layer. Results at $t = 1$ (a), $t = 5$ (b) and $t = 30$ (c) for the time-varying ϵ_1.



(a) $t = 1$



(b) $t = 5$



(c) $t = 30$

Figure 5-25: Estimations obtained by MGEnKF of the momentum ρv (S.I. units) at the centreline $y = 0$ of the mixing layer. Here, MGEnKF is only used to provide the estimation of ϵ_1. Results at $t = 1$ (a), $t = 5$ (b) and $t = 30$ (c) for the time-varying ϵ_1.

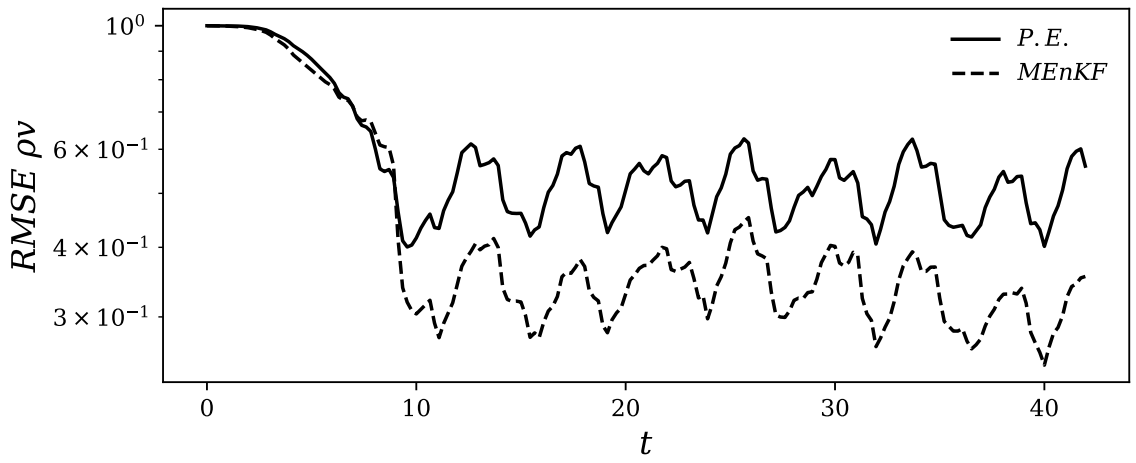


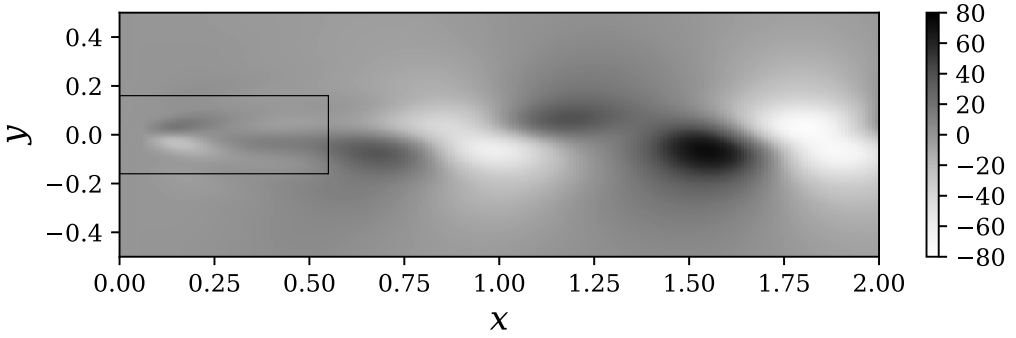
Figure 5-26: Time evolution of the RMS error of ρv for the case of a time-varying inlet parameter ϵ_1. The symbol P.E. corresponds to the case where MGEnKF is only used for the estimation of ϵ_1. The notation MGEnKF corresponds to the standard version of the algorithm, including parameter estimation and physical state correction via Kalman gain.

is obtained through the ensemble statistics. In this scenario, the state estimation obtained in the step 4 of the MGEnKF algorithm presented in Chapter 4 is directly projected in the fine mesh space and used as final solution, thus $(\mathbf{x}_{\mathrm{k}}^{\mathrm{F}})^{\mathrm{a}} = (\mathbf{x}_{\mathrm{k}}^{\mathrm{F}})^{'}$

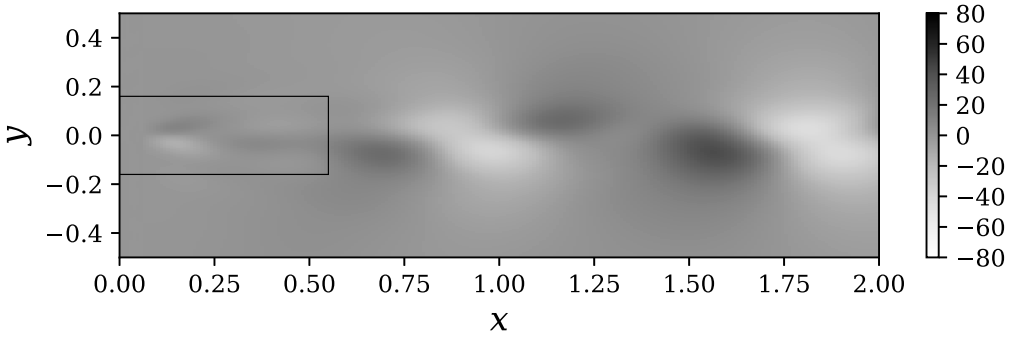
- S_2: A standard MGEnKF algorithm, as described in Chapter 4

- S_3: A MGEnKF algorithm where the ensemble prediction is just used to estimate the unknown parameter of the system. No update of the physical solution is performed using the correction via Kalman gain.

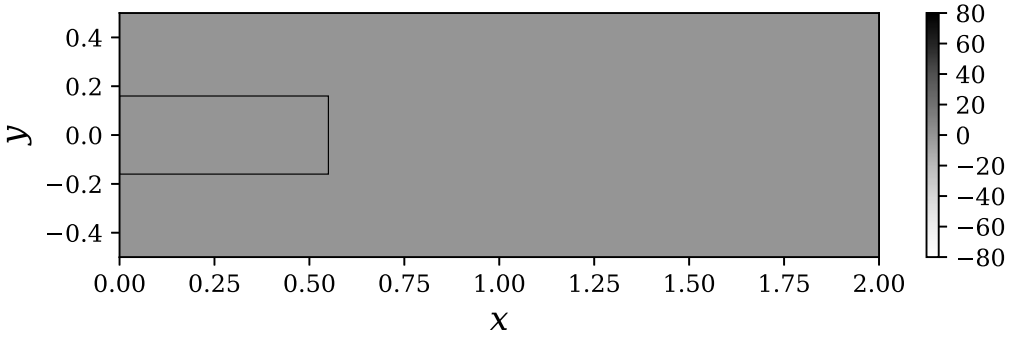
The results are shown in Fig. 5-27 after the first forecast/analysis step. For clarity, we introduce a normalized criterion, $(\Gamma_k^{\mathrm{F}})^* = \frac{\Gamma_k^{\mathrm{F}}}{\Gamma_C}$, where Γ_C is defined as $\max_k \left| \frac{\left((\rho v)_k^{\mathrm{F}}\right)^{\mathrm{a}} - \left((\rho v)_{k-1}^{\mathrm{F}}\right)^{\mathrm{a}}}{\Delta t} \right|$. As expected, $\Gamma_k^{\mathrm{F}} = 0$ everywhere when MGEnKF is only used for the parameter estimation (scenario S_3). In this case, the time advancement of the solution is performed using the model only, which exactly complies with the discretized equation and respects conservativity (up to a convergence error which is negligible). On the other hand, results in Fig. 5-27 (a) show some lack of conservativity in the physical domain for the first scenario. This is also expected, since



(a) Standard Dual EnKF on the coarse mesh.



(b) Standard MGEnKF.



(c) MGEnKF used only for parameter estimation.

Figure 5-27: Analysis of the conservativity of the dynamical model via the normalized quantity $(\Gamma_k^{\mathrm{F}})^*$. Results are shown for three scenarios: (a) the classical Dual EnKF, (b) the classical MGEnKF algorithm and (c) the MGEnKF only used for parameter estimation. The observation region is represented by the rectangle in black lines.

no constraint is imposed to force the Kalman gain correction to comply with the dynamical equations. Finally, results for the MGEnKF are shown in Fig. 5-27 (b). The evolution of $(\Gamma_k^{\mathrm{F}})^*$ is very similar to the results observed for the first scenario. However, one can clearly see that this field appears to be sensibly smoothed out by the multigrid iterative procedures in step 4 and 5 of the MGEnKF algorithm. As previously discussed, complete conservativity starting from an erroneous state at $k-1$ is possibly not an optimal objective, while one wants a regularized solution to avoid affecting the precision of the global calculation. On this last objective, the MGEnKF appears to provide a better result when compared with the classical Dual EnKF, described in the first scenario. Considering also that the MGEnKF showed better accuracy than the algorithm relying on parameter estimation only, one can conclude that the MGEnKF provides an efficient compromise between global accuracy and regularization of the solution. In order to draw more information about this important aspect, the MGEnKF algorithm needs to be tested for the simulation of three-dimensional compressible flows, where the Kalman gain correction may be responsible for important acoustic phenomena which are not observable in 2D and 1D dynamical systems.

5.6 Concluding remarks

The MGEnKF has been assessed for four different test cases using models of increasing complexity. In Sec. 5.2 and Sec. 5.3, the performance of the estimator was assessed using low-complexity flow models. In the DA experiments performed, the influence of the *inner loop* was demonstrated. Using coarser grids for the ensemble simulation can significantly reduce the computational cost of the estimator; however, an important decrease in the accuracy of the ensemble prediction system may be expected. The *inner loop* of the algorithm can be used to counter this unwanted effect by directly acting on the model used on the coarse grid via model correction terms. The basic idea behind this correction procedure is to construct parametrized model correction terms and to *optimize* its coefficients using the *main simulation* as surrogate observation.

The effectiveness of this strategy was proven for test cases where the observation window was situated in the middle of the computational domain. The models used, artificially diffuse and disperse the signal imposed at the inlet as it is propagated downstream. Using the EnKF to estimate the *true* amplitude of this signal from observation can be challenging in this situation and when the *inner loop* was deactivated, the estimator showed poor accuracy, with a general tendency to overestimate the amplitude of the inlet condition. The model correction terms proposed, designed to artificially modify the diffusion and the dispersion of the signal, were successfully optimized with the *inner loop*, achieving a substantial global accuracy gain in the estimation of the inlet amplitude.

The models used in Sec. 5.4 and in Sec. 5.5, introduce additional layers of complexity to the DA experiments performed:

- In Sec. 5.4, the model used is the compressible Euler system of equations. Apart from the non-linear convective transport in the momentum equation, which is already present in Burgers' equation (Sec. 5.3), in the Euler equations the velocity of the flow is coupled with the thermodynamic state of the flow (density and pressure).

- In Sec. 5.5, the flow model used is the fully compressible 2D Navier-Stokes equations, introducing 2D non-linear convective transport and viscosity effects.

The determination of the form of the model correction term $\mathcal{C}$ in this situation is much more challenging, and is currently under research in our research group. For this two test cases, two simplifications have been performed: the inner loop is deactivated, i.e., $\mathcal{C} = 0$ and the observation zone is placed close to the inlet of the computational domain. This allows to reduce the loss of prediction accuracy in the ensemble. The tests performed on the last two sections, mainly concerned state-estimation and time-varying parameter estimation, allowing to assess the performance of the *outer loop* without the *inner loop*. The analysis performed represents the minimum estimation accuracy that one can expect from to achieve from the MGEnKF, and the *inner loop* should only improve the global accuracy of the estimator.

Chapter 6

Application of the MGEnKF to a complex problem: BARC

6.1 Introduction

In Chapter 5 the MGEnKF was validated against 1D and 2D simple flow configurations. The MGEnKF is now used to investigate a complex flow using scale-resolving 3D LES numerical simulation.

Large Eddy Simulation (LES) (Pope, 2000; Sagaut, 2006) is currently one of the most diffused methods for the numerical prediction of turbulent flows, and it is available in both licensed and open-source codes. LES is based on a filtering of the scales of the flow, and it can intrinsically represent unstationary and three-dimensional flows via a time-and-space resolved approach, but with fewer computational resources when compared with DNS. Moreover, it can be easily coupled with RANS methods or wall-functions to alleviate even further the computational needs for near-wall turbulence. These favourable features have extensively promoted its diffusion in the academic community and, more recently, for industrial applications.

However, the main limitation of this technique is that the theoretical framework for which the approach is developed can only be rigorously reproduced for academic test cases (i.e., simple geometries and uniform grids). Approximations may be implemented in more complex cases, which may however lead to unexpected results,

such as lower accuracy of the results with mesh refinement. Numerous works in the literature (Meldi et al., 2012; Meyers & Sagaut, 2006) indicate that these unexpected observations are tied with non-linear interactions between different sources of error, such as explicit and implicit filtering, discretization procedures, quality of the mesh, general set-up of the test case and subgrid scale modelling. These elements are usually governed by numerous free parameters that must be chosen by the user. Extensive analyses using classical trial-and-error approaches are in most cases not feasible due to the large costs of each LES.

Recently, new models for the unresolved stresses in LES have been proposed that aim to overcome the uncertainty issues related to this type of simulation technique. One of such new techniques is the integral length-scale approximation model (ILSA), proposed by Piomelli et al., 2015 and Rouhi et al., 2016. This model aims to provide a grid independent subgrid-scale model. Lehmkuhl et al., 2019 tested the ILSA model for complex geometries. Although the results are promising, the model depends on a parameter that has to be set by the user. The optimal determination of this coefficient requires previous calculations and there is naturally some uncertainty attached to the procedure.

In this chapter, the MGEnKF is used in a practical application of engineering interest, namely the high-Reynolds-number flow around a rectangular cylinder of aspect ratio 5:1, which is the object of an international benchmark (BARC, Bruno et al., 2014). Despite the simple body geometry, the flow dynamics is complex, including flow separation at the upstream edges, mean-flow reattachment on the cylinder side and vortex shedding in the wake.

As shown in the first review work (Bruno et al., 2014), the LES predictions of some flow features and quantities are affected by a significant dispersion and by a high sensitivity to modelling and numerical parameters, such as subgrid scale modelling (Mariotti et al., 2017) and grid refinement (Bruno et al., 2012; Mariotti et al., 2017). More recently, (Rocchio et al., 2020) showed that the introduction of a very small rounding of the upstream edges, such as those due to manufacturing tolerances, has a strong impact on LES predictions.

For all these reasons, this kind of flow is well adapted and challenging for the appraisal of DA procedures. The MGEnKF algorithm is used herein to (i) improve the predictive capabilities of LES carried out on a rather coarse grid, by integrating reference data from a high-fidelity LES, run on a highly refined grid, (ii) to calibrate a parameter which regulates the dissipation introduced by a modal filter that can be considered as a subgrid scale dissipation.

6.2 Numerical and modelling ingredients

We consider the incompressible flow around an elongated rectangular cylinder, whose chord-to-depth ratio is $B/D = 5$, at zero angle of attack. The cylinder upstream edges are rounded with a curvature radius $r/D = 0.0037$, which is the smallest one considered in Rocchio et al., 2020 and corresponds to a value that falls within possible manufacturing tolerances. The numerical simulations are performed using Nek5000, an open-source code based on a high-order spectral element method (Fischer et al., 2008). Each spectral element is rectangular or a suitable coordinate mapping of a rectangle. The basis functions inside the elements are Legendre polynomials of order N for velocity and $N - 2$ for pressure in each direction; $N = 6$ has been used in this work, as in the previous works by Lunghi et al., 2022; Mariotti et al., 2017; Rocchio et al., 2020. Based on the high-order splitting method developed in Maday et al., 1990, a third-order backward finite-difference scheme is used for time advancing. The viscous terms are treated implicitly while the convective terms are explicit, with a third-order forward extrapolation in time.

In Nek5000, when dealing with high Reynolds numbers, the spectral element method can be stabilized by applying a filtering method in the modal space: a low pass filter is employed here as in Mariotti et al., 2017 and Rocchio et al., 2020. At the end of each step of the time integration, the low-pass explicit filter is applied to the velocity field in the modal space. This filter is characterized by a quadratic transfer function, which acts from the mode $N - k_c$ ($k_c = 3$ herein) up to the highest mode

N. The transfer function is written as follows:

$$\begin{cases} \sigma_k = 1 & k < k_c \\ \sigma_k = 1 - w\left(\frac{k-k_c}{N-k_c}\right)^2 & k_c \leq k \leq N \end{cases} \tag{6.1}$$

and, once the cut-off mode (k_c) is set, it can only be tuned through a weighting parameter w. Since the filter acts only at the highest resolved modes, it may be interpreted as a SGS dissipation (Domarazdki, 2010; Mathew et al., 2003). Clearly, by changing the cut-off mode, the behaviour of the stabilizing filter changes as well. However, to limit the number of tunable parameters and to be consistent with the analogous study carried out in Mariotti et al., 2017 for perfectly sharp upstream edges, it is kept constant in this work.

The computational domain is sketched in Fig. 6-1, where x/D, y/D and z/D denote the streamwise, vertical and spanwise directions respectively. The domain is rectangular and it has the same dimensions as the ones used in other LES contributions to the benchmark (Bruno et al., 2012; Bruno et al., 2014; Mariotti et al., 2017; Rocchio et al., 2020). In particular, the cylinder centre is located at $x/D = y/D = 0$ and we have $-15.5 \leq x/D \leq 25.5$, $-15.1 \leq y/D \leq 15.1$ and $0 \leq z/D \leq B$. A uniform velocity profile with no turbulence is imposed at the inlet, no-slip is imposed at the body surface and traction-free boundary conditions are used for the outflow and for the upper and lower boundaries of the domain. Finally, periodicity is imposed in the spanwise direction. The Reynolds number, Re, based on the free-stream velocity and on the cylinder height is $4 \cdot 10^4$.

The spectral discretization is the same used in Mariotti et al., 2017, for the BARC simulations having perfectly sharp upstream edges. In order to represent the curvature radius, the computational grid has been modified in the regions where the upstream edges are placed. To this aim, a transition box, which allows switching from a Cartesian grid to a Cylindrical one, is used to transform each spectral element containing the upstream rounding edges.

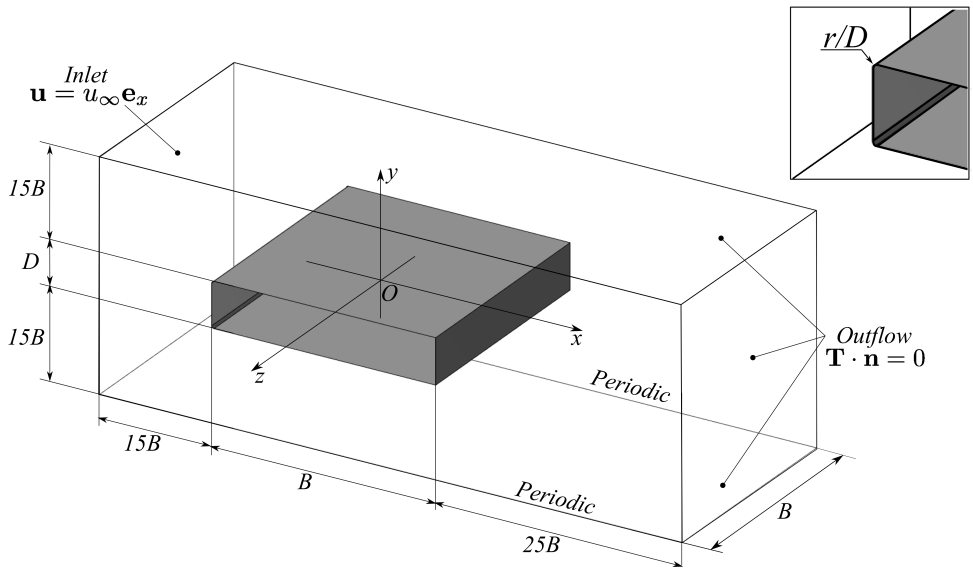


Figure 6-1: Sketch of the computational domain.

6.3 Data-driven enhancement of LES using the MGEnKF

6.3.1 Description of the DA experiment

The MGEnKF method is now used to augment the predictive capabilities of an LES simulation run on a intermediate-level grid (baseline grid) with synthetic observation obtained from a reference LES run on a refined grid (see Table 6.1 for detailed information on the meshing). The reference simulation, which has been run with $w = 0.018$, is considered to be the *true* state of the flow. The main objective of the MGEnKF is to reduce mismatch between the flow predictions obtained with the baseline grids and reference simulation obtained on the refined grid. Details are now provided for each ingredient needed for the application of the Data Assimilation algorithm. Observation is obtained sampling the three components of the instantaneous velocity field of the reference simulation, which is run for a total time of $T = 200t_c$ in advection time units. A cloud of sensors is embedded in the region $-2.25 \leq x/D \leq -1.75$, $0.75 \leq y/D \leq 1$, $1 \leq z/D \leq 4$ (see Fig. 6-2) and data is sampled every $0.1t_c$, for a total of 2000 local snapshots of the instantaneous flow which are captured by the sensors. In order to provide a higher degree of randomization to the sampled data, artificial Gaussian noise of covariance $\mathbf{R}^\circ = \sigma_o^2 I$ is applied to the observation, where

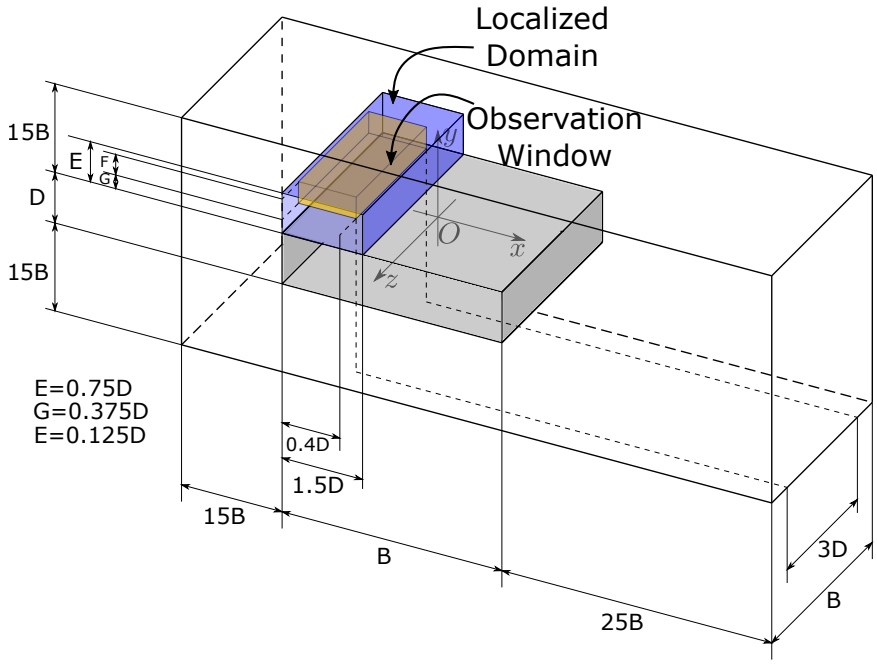


Figure 6-2: Sketch of the computational domain with the localized assimilation region and the observed space.

Grid	$\Delta x/D$	$\Delta y/D$	$\Delta z/D$	Size
Refined	0.0625	0.0625	0.275	≈100M
Baseline	0.125	0.125	0.558	≈10M
Coarse	0.25	0.25	1	≈2M

Table 6.1: Details of the numerical grids used for computation. The total number of elements and the near wall resolution is indicated.

I is the identity matrix. In this test case, σ_o is set to $\sigma_o = 0.02$, and the artificial noise added is considered to be uncorrelated in time, therefore $\mathbf{R}^{\circ}$ is diagonal. This noise improves the rate of convergence of EnKF-based tools, as shown in several research works (Moldovan et al., 2021; Tandeo et al., 2020). The reference simulation, whose results will be used to assess the precision of the DA procedure, will be also referred to as the true state.

The code providing a numerical model for the flow, Nek5000, is embedded within the MGEnKF algorithm with the help of an open-source coupler OpenPALM (Buis et al., 2006; Duchaine et al., 2015). Nek5000 is thus used in a multilevel approach where the main simulation is run on the baseline grid, while the ensemble members are run on the coarse grid. A sketch of the application of the MGEnKF to this specific

test case is shown in Fig. 6-3.

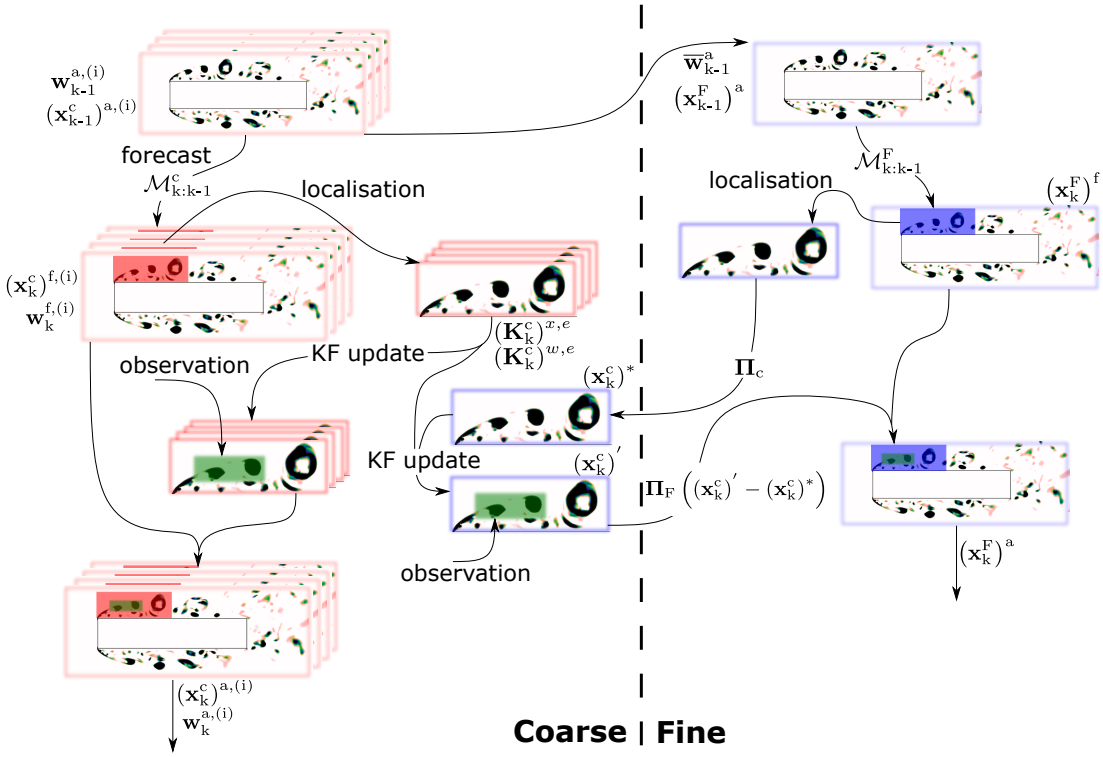


Figure 6-3: Application of the MGEnKF algorithm to the BARC test case.

The main features of the numerical grids used for computation are summarized in Tab. 6.1. One can see that the grid used for the calculations of the ensemble members is around 2 times coarser than the baseline mesh and 4 times coarser than the refined mesh.

The DA experiment, which will be referred to in the following as DA augmented LES, is performed in the following conditions:

- The initial condition imposed to the main and ensemble simulations is obtained from the reference simulation. More specifically, the first time step in the reference simulation where observation data is sampled is interpolated on the baseline grid and the coarse grid, and it is used as initial solution for every simulation running within the MGEnKF. This is performed in order to avoid phase synchronization issues.

- Considering that the outer loop of the MGEnKF works exclusively using the

ensemble members, the observation sampled from the reference simulation is directly interpolated on the collocation points of the coarse grid for sake of simplicity. For this case, the observation is interpolated on a total of 6 spectral elements. The size of the observation vector is thus $\approx 4 \cdot 10^3$.

- An ensemble of size $N_e = 30$ is used in the MGEnKF. Initial values of w for each member of the ensemble $w^{(i)}, i = 1, 2, 3..., N_e$ are sampled from the Gaussian distribution $\mathcal{N}(w_0, R_w)$, where w_0 is the prior value and is set to $w_0 = 0.018$, and $R_w = \sigma_w^2 I$ with $\sigma_w = 0.0001$.

- Observation is sequentially integrated into the system to: i) optimize w, ii) correct the physical state of each member of the ensemble. The DA state update is extended towards the main simulation owing to interpolation functions which connect the coarse and baseline grids. In addition, the parametric description of main simulation is also optimized updating the parameter value with $\mathbb{E}[w]$, i.e. $\overline{w^{(i)}}$.

6.3.2 Simplifications applied to the MGEnKF

A few simplifications have been applied to the classical MGEnKF algorithm to make it more suitable to this HPC application. First, the state vector includes the three components of the velocity field but not the pressure, which is updated via a Poisson equation at the beginning of the following time step. The nature of the Poisson equation allows to rapidly propagate information provided in the DA update (Meldi & Poux, 2017). Therefore, the MGEnKF does not grant conservativity at the end of the analysis phase, but the momentum and mass conservation are respected at the following time step. This simplification allows reducing sensibly the size of the state vector without producing numerical instabilities in the analysis phase, owing to the effect of the Poisson equation. In addition, the effect of the Kalman gain has been *localized*. In practice, a correlation matrix $\boldsymbol{\beta}$ is introduced in (3.39) as follows (Soares et al., 2018):

$$\mathbf{x}_k^{\mathrm{a},(i)} = \mathbf{x}_k^{\mathrm{f},(i)} + \boldsymbol{\beta} \circ \mathbf{K}_k^{\mathrm{e}} \left(y_k^{\mathrm{O},(i)} - \mathcal{H}_k \left(\mathbf{x}_k^{\mathrm{f},(i)} \right) \right), \tag{6.2}$$

where $\circ$ is the Schur product (element-wise product between $\boldsymbol{\beta}$ and $\mathbf{K}_k^{\mathrm{e}}$). $\boldsymbol{\beta}$ represents thus a distance-dependent correlation matrix that can be constructed as (Gaspari & Cohn, 1999):

$$\beta = \exp\left[-\frac{d\left(i,j\right)^2}{2L^2} \right], \tag{6.3}$$

where $d\left(i,j\right)$ represents the euclidean distance between a point where observation is available, i, and a model grid point j, and L is the correlation distance that was set to $L = 0.4D$ in this case. This implies that the matrix $\mathbf{K}_k^{\mathrm{e}}$ is transformed using a spatial Gaussian filter. The product between the correlation matrix $\boldsymbol{\beta}$ and the Kalman Gain erases any spurious long-range statistical correlations (beyond a length L) that might appear. In addition, all the velocity updates via Kalman filter have been set to zero outside a prescribed domain defined by the coordinates: $-2.5 \leq x/D \leq -1$, $0.5 \leq y/D \leq 1.25$, $0 \leq z/D \leq 5$. This strategy is used to avoid spurious correlation to affect the precision of the results, as well as to further reduce the computational costs. Using this strategy, the size of the *localized* state vector updated via data assimilation is $\approx 6 \cdot 10^4$, around 30 times less than the complete state vector on the coarse grid.

The inner loop of the MGEnKF has been deactivated for sake of simplicity.

6.3.3 Results

Starting from the initial solution, coarse grid ensemble and main simulation are sequentially augmented at each analysis phase. This implies an optimization of the parameters $w^{(i)}$ as well as an update for every simulation (main and ensemble members) are performed.

The results obtained via MGEnKF are now analysed. Fig. 6-4 shows the time evolution of w, which is sequentially optimized during the analysis phases. One can

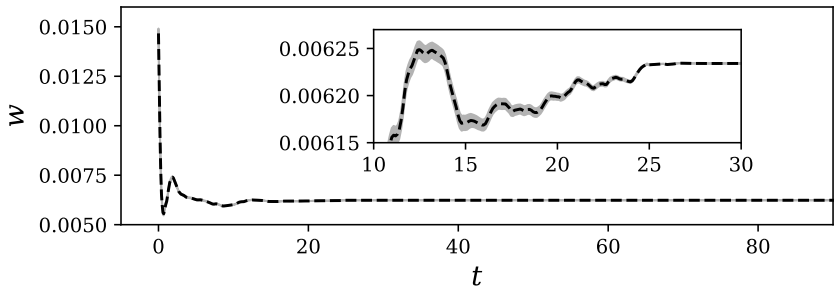


Figure 6-4: Convergence of the MGEnKF optimization of the explicit LES coefficient filter applied to the resolved equations. The coefficient is initialized with the baseline value ($w = 0.018$). The shaded area represents the 95% credible interval.

see that the value of w rapidly moves from the prior and it then exhibits some fluctuations before converging to $w = 0.00624$ in approximately $t = 25t_c$. A progressive reduction of the variance in the estimation (shaded area) is also observed, which is a good indicator of convergence. The optimized value exhibits a large difference when compared with the prior ($w = 0.018$ vs $w = 0.00623$). In order to provide a complete picture of the effectiveness of the method, four LES are compared in the following: the reference LES ($w = 0.018$), the data augmented LES, the optimized LES ($w = 0.00623$), which is an independent simulation run with the optimal value obtained for w from the DA experiment (data augmented LES), and a baseline LES which is identical to the reference simulation but run on a different grid. The reference simulation is run on the refined grid while the three other LES are run on the baseline grid. The MGEnKF is used only on the data-augmented LES, while the optimized LES simulation is just fed with $w = 0.00623$. This allows to compare a simulation where state and parametric estimation is performed with a simulation where only the latter is performed.

The time-averaged flow streamlines provided for the four simulations in Fig. 6-5 seem to indicate that the size of the recirculation regions for the data-augmented LES and the optimized LES is significantly improved when compared with the baseline LES. Similar conclusions can be drawn by the analysis of Fig. 6-6 (a), where results for the mean pressure coefficient, $\langle C_p \rangle$, on the upper surface of the cylinder are provided. The predicted profile for both the DA augmented LES and the optimized LES is

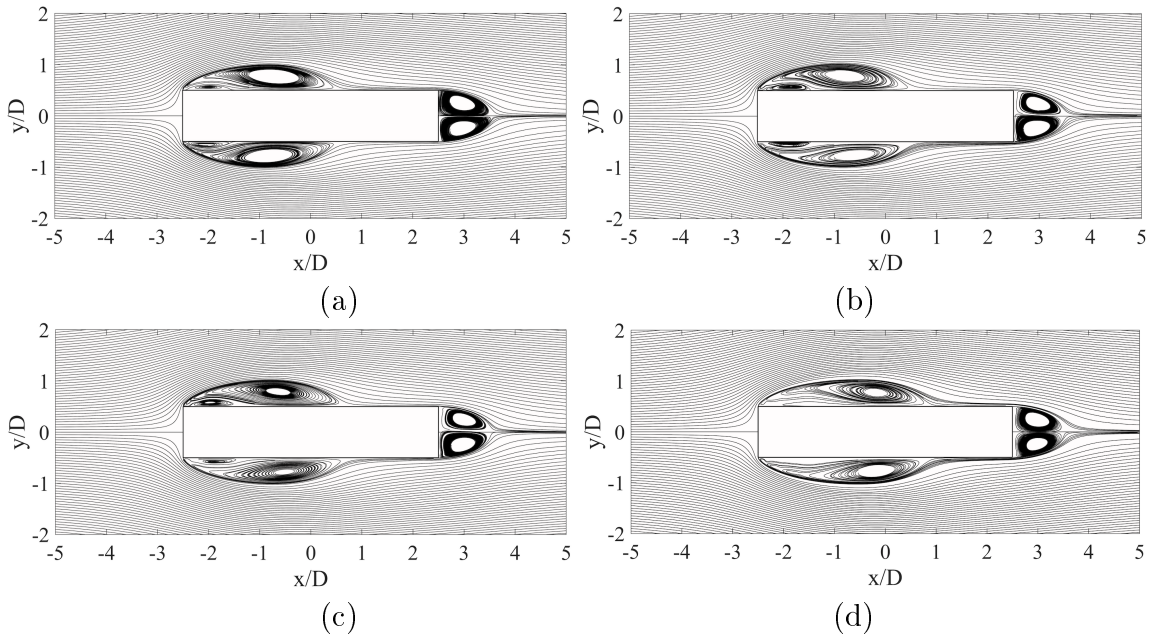


Figure 6-5: Streamlines of the time-averaged flow for (a) the reference LES, (b) the data-augmented LES, (c) the optimized LES and (d) the baseline LES.

significantly closer to the reference solution when compared with the baseline LES.

In addition, Fig. 6-6 (b) also shows an improvement for the prediction of the variance of the pressure coefficient $\sigma(C_p)$. The maximum of this parameter is a measure of the mean point of reattachment of the flow. Despite the observable differences in magnitude, results for the data-augmented LES and the optimized simulation are aligned with the reference results, exhibiting a significant improvement from the baseline LES. One can also see that results from the DA augmented simulation appear to be marginally better than the ones from the optimized simulation, on average. Despite the localization procedures applied to the Kalman gain, it appears that a positive effect of the DA state estimation in terms of improvement of accuracy can be observed on the statistics of the flow.

The results indicate that the prior value of w, when used in combination with the baseline grid, produces excessive dissipation. One could argue that the optimization process within the MGEnKF, reduces the internal dissipation of the LES, by reducing the value of w, in order to compensate the different numerical error associated to the coarsened mesh elements of the baseline grid, when compared with the refined grid

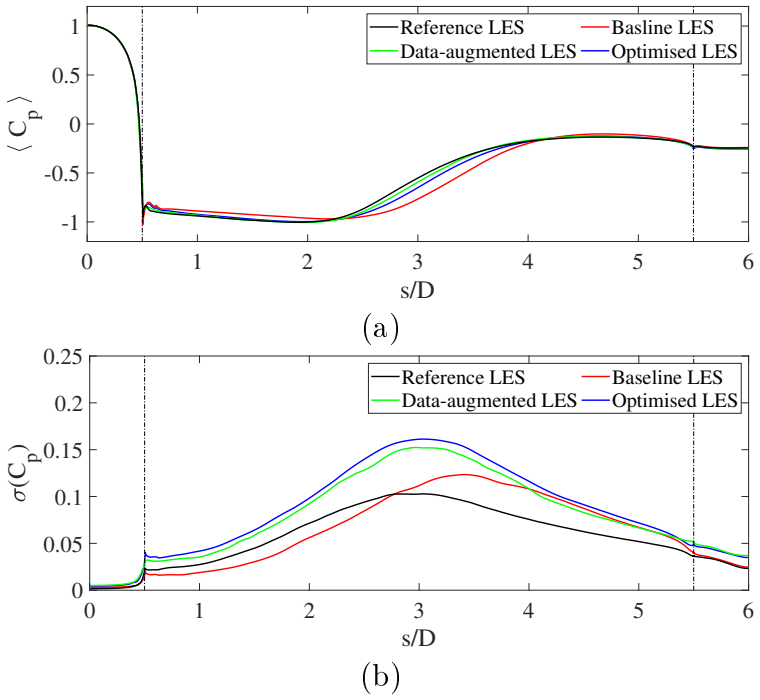


Figure 6-6: (a) Pressure coefficient averaged in time, spanwise direction and between the upper and the lower surfaces, $\langle C_p \rangle$, and (b) its standard deviation, $\sigma(C_p)$, evaluated on the body. Comparison between the results obtained for the reference LES, the data-augmented LES, optimized LES, and baseline LES.

used for the reference simulation.

6.4 Conclusions

The MGEnKF was tested and validated against the international benchmark referred to as BARC (Benchmark of the Aerodynamics of a Rectangular 5:1 Cylinder). This configuration is a simplified model of problems of interest in civil engineering. In the LES setup of this test case, there are numerous free parameters that must be chosen by the user that might have significant impact on the accuracy of the simulation. Extensive analyses using classical trial-and-error approaches are in most cases not feasible due to the large costs of each LES.

A Data Assimilation experiment was performed where the MGEnKF algorithm is used to improve the predictive capabilities of an intermediate-fidelity by noisy data (synthetic observation) from a high-fidelity LES. The ensemble of simulations that are responsible for the estimation process were run on a rather coarse grid. The devised assimilation strategy tackled the optimization of a parameter, w that regulates the numerical dissipation introduced by a modal filter. The working hypothesis was that the additional numerical errors (additional numerical dissipation) that appear on LES run on coarser grids can be reduced by calibrating the artificial dissipation introduced by the solver. Additionally, the statistics obtained from the coarse grid ensemble are used to update the intermediate-fidelity LES with a KF correction.

The performance of the MGEnKF estimator has been assessed by evaluating first and second order statistics. More specifically, pressure distribution on the surface of the body as well as the streamlines of the time-averaged velocity have been compared for four different LES runs: reference LES, data augmented LES, optimized LES, and baseline LES. The optimized and baseline LES simulations are standard runs where the only parameter that changes is w; $w = 0.00623$ in the first case while $w = 0.018$ in the second. The data-augmented LES is the run where the MGEnKF is applied. This estimation strategy allows performing an in-streaming calibration of w together with a state estimation on the intermediate grid, with the synthetic, noisy observation

obtained from reference LES. The estimated flow field obtained with the MGEnKF were the ones where the best agreement with the reference simulation was observed.

Chapter 7

Conclusions and future research

↪***Context and main objectives***

The accurate prediction of turbulent flows is still an open challenge for research in fluid mechanics. Technological barriers concerning computational power restrict the usage of direct numerical simulation tools to low-Reynolds cases, and a revolution in turbulence modelling is not expected in the near future.

In spite of its limitations, CFD has been widely adopted by aircraft and car manufacturers in their design procedures; however, off-design conditions which are usually related to the inherent unstationary features of the flow require more sophisticated tools to either improve the design process or to conceive robust active control techniques.

Tools coming from Estimation Theory are starting to be used by CFD researchers in an attempt to increase the prediction accuracy of already existing codes. The idea behind this class of techniques is to *combine* numerical data with experimental one to improve the estimation. This falls into a broader mathematical domain, which has been developed for decades and is widely used in weather forecast: Data Assimilation. Sequential Data Assimilation techniques, with the Kalman Filter algorithm being the most well known tool, are particularly appropriated for unsteady flows; however, solving a data assimilation problem with a sequential approach is strongly constrained in realistic configurations by the excessive computational costs of the methods based

on the Kalman filter.

The objective of this PhD study was to develop a sequential DA estimator that could be integrated in widely used CFD codes.

↪*An advanced estimation algorithm for assimilation of unsteady flow*

We have developed an estimation algorithm combining the Ensemble Kalman Filter and the multigrid strategy for assimilating unsteady flows. The so-called MGEnKF algorithm (Multi-Grid Ensemble Kalman Filter) exploits physical states obtained on multiple grids of different resolution to perform the state estimation and parametric optimization using EnKF procedures. More precisely, when only two grids are used an ensemble of low-fidelity simulations of the flow is run on a coarse grid level together with a single higher-resolution simulation on the finest mesh level. The state estimation obtained at the coarse level and the associated ensemble statistics are used to filter the finest mesh solution and optimize a set of parameters describing the model (boundary conditions, model parameters, ...). This procedure allows to i) reduce the computational costs of the EnKF and ii) ensure the conservativity and the smoothness of the final solution.

Generally speaking, this method falls into the class of *multilevel* techniques that aim to improve the estimation of statistics of expensive numerical simulations by considering different levels of resolution of the same set of equations.

↪*Validation*

The assessment of the method was performed via the analysis of one-dimensional, two-dimensional and three-dimensional test cases, using different models of increasing complexity. The results show that the MGEnKF can successfully update the state of a system with the available observations to increase the global accuracy of the state prediction provided by the model. In addition, the parametric description of the numerical problem (in terms of prescribed boundary conditions, turbulence closures,

...) can be adequately optimized taking into account the different mesh resolutions employed in the algorithm. Here is a summary of the test cases investigated, with the key findings:

- In Sec. 5.2, the performance of the MGEnKF was assessed in a 1D state/parameter estimation problem. The model chosen was the linear 1D advection equation, and the synthetic observation was generated from a previously run simulation with a known set of parameters. The estimator was used to correct the state and to infer the amplitude of the sinusoidal inlet condition. The analysis focused on the improvement in global performance due to the *inner loop*. This step of the DA strategy targets a systematic improvement of the accuracy of the ensemble members using surrogate information from the main simulation run on the fine grid level. The results indicate the importance of the *inner loop* in improving the performance of the Data Assimilation algorithm. The proposed model correction term, $\mathcal{C}_{k:k-1}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a},(\mathrm{i})}, \gamma_k^{\mathrm{f},(\mathrm{i})}\right)$, has been derived from the exact equation in order to compensate dispersive and diffusive numerical errors. The tests performed showed that this approach can fully correct the numerical errors associated with the coarse grid level where ensemble members are run.

- In Sec. 5.3, the test case reported is similar to the experiment performed in Sec. 5.2, however, the model considered now is the 1D Burgers' equation, which incorporates non-linear convection transport and viscous effects. The dispersive errors observed in the coarse grid ensemble are corrected with part of the model correction terms derived in Sec. 5.2 (dispersive correction term). An additional diffusion term parametrized by a parameter, α, was introduced to control the loss of amplitude of the solution. The improvement in the estimation accuracy due to the use of the *inner loop* is remarkable. However, contrarily to what was observed in the linear advection case, the model correction term here is not able to fully correct the discrepancies due to the numerical error.

- In Sec. 5.4, the MGEnKF is used to track the time evolution of a free parameter for the case of a wave propagation, using a one-dimensional Euler model.

More specifically, a time-varying parameter concerning the inlet condition was inferred from synthetic noisy observation. Three cases are here investigated, varying the time window between successive assimilations. The estimator can efficiently represent the evolution in time of the parameter, as well as to provide an accurate state estimation. However, the global prediction is significantly degraded if the update frequency is lower than a threshold value, which is arguably connected to the physical features of the flow.

- In Sec. 5.5, a DA-experiment was performed on the two-dimensional spatially evolving mixing layer at $\mathrm{Re} = 100$. As in the test case performed in Sec. 5.4, a time-varying parameter concerning the inlet condition is inferred from synthetic noisy observation. The parameter estimation obtained with the algorithm is not perfectly accurate; however, the general evolution trend is captured. Moreover, the algorithm was successful in accurately updating the state of the main simulation on the fine grid. This is proven when evolution over time of the root-mean-square error (RMSE) is analysed; the error is more significantly reduced when both state-estimation on the fine grid and parameter estimation are performed than when only the parameter estimation procedure is considered.

- In Chapter 6, the MGEnKF was tested and validated against the international benchmark referred to as BARC. Despite the relatively simple geometry, the flow around the cylinder presents complex features. The simulation of the flow around such geometries with LES techniques is extremely sensitive to the discretization strategy used, to explicit/implicit filters that one might apply, and to small changes in the actual geometry. Taking the results obtained from a highly-resolved LES of the flow around the body as reference data, we aim to improve the accuracy of a lower-fidelity simulation run on a grid of intermediate refinement level with the MGEnKF. The synthetic observation used was generated from the reference simulation. In addition to the state estimation performed with the MGEnKF, we also tackled the optimization of a parameter w that regulates the numerical dissipation introduced by an explicit modal filter.

Pressure distribution on the surface of the body as well as the streamlines of the time-averaged velocity show that, when compared to a baseline case, the data-augmented LES (LES improved in streaming with the MGEnKF) is in better agreement with the reference data. The analysis of the results obtained for this DA experiment is still in progress. An in-depth investigation of the dynamical features of the flows obtained for each different test case performed would allow comparing the data-augmented LES with the optimized LES. In this test case, the parameter w obtained from the data-augmented LES is imposed throughout the entire simulation. This comparison would allow identifying more precisely the effects of state-estimation on the data-augmented LES.

The models used in Sec. 5.4, Sec. 5.5, introduce additional layers of complexity to the DA experiments performed:

- In Sec. 5.4, the model used is the compressible Euler system of equations. Apart from the non-linear convective transport in the momentum equation, which is already present in Burgers' equation (Sec. 5.3), in the Euler equations the velocity of the flow is coupled with the thermodynamic state of the flow (density and pressure).

- In Sec. 5.5, the flow model used is the fully compressible 2D Navier-Stokes equations, introducing 2D non-linear convective transport and viscosity effects.

The determination of the form of the model correction term $\mathcal{C}$ in this situation is much more challenging, and is currently under research. For this two cases, two simplifications have been performed: the inner loop is deactivated, i.e., $\mathcal{C} = 0$ and the observation zone is placed close to the inlet of the computational domain. This allows to reduce the loss of prediction accuracy in the ensemble. With the *inner loop* deactivated, the tests performed on the last two sections, mainly concerning state-estimation and time-varying parameter estimation, allow assessing the performance of the *outer loop*. The analysis performed represent the minimum estimation accuracy that one can expect from to achieve from the MGEnKF, and the *inner loop* should

only improve the global accuracy of the estimator.

$\hookrightarrow$***Perspectives***

The MGEnKF opens interesting perspectives for potential application to active flow control and industrial design:

- **Flow Control**. Linear model-feedback control problems based on limited and noisy measurements involve two independent sub-problems: (i) the state feedback control problem; (ii) the state estimation problem. The estimator developed can be used as state-estimation tool, since its capabilities to accurately reproduce unsteady flow phenomena have been proven.

- **Design and analysis**. In the case of aerospace design, for example, measurements and CFD have rarely been used together. CFD is at a relatively matured state in the design process of certain parts of the aeroplane, with the high-speed wing being a particular success-story. There remain, however, areas where CFD faces important limitations. The MGEnKF can be used in this scenario to improve the overall accuracy of the simulations carried out for analysis by integrating experimental data.

The MGEnKF estimator has been tested on simplified flow models, with the exception of the data assimilation experiment performed on the 3D BARC case. In this test case, we have shown that the MGEnKF can be used with a matured solver, namely Nek5000. However, the inner loop of the algorithm was deactivated, since it has been validated only on simple 1D test cases. Further research work is needed for this specific part of the algorithm, but the 1D experiments reported in Sec. 5.2 and Sec. 5.3 set the foundations for the application of the inner loop to 2D and 3D cases. More specifically, the proven capability of the inner loop to improve the accuracy of the low-fidelity low-cost ensemble simulation opens exciting perspectives of application to grid-dependent reduced-order models extensively used in fluid mechanics applications for complex flows, such as Large Eddy Simulation (LES). An

accurate model reconstruction via the inner loop may alleviate or even prevent one of the major issues associated with multilevel applications in fluid mechanics, that is, the sensitivity of the parametric description of the model (in the form of the set of parameters θ) to different mesh resolution. In this scenario, one may just tune the reduced-order model for the most refined grid resolution using the *outer* loop, and compensate the emerging differences from progressively coarser grids using the *inner* loop to optimize the additional model. Of course, the additional model provided for the coarse grids must be suitable for this task, as seen for the one-dimensional Burgers' equation, and difficulties are expected for applications with scale-resolved turbulent flows. Combined applications of EnKF with machine learning tools, which have been recently explored for simplified test cases, may provide success in deriving precise model structures when included in the formalism of the MGEnKF model for the *inner loop*.

Appendix A

Data Assimilation algorithms

A.1 Kalman filter algorithm

The Kalman filter algorithm given in Sec. 3.5.5.1 corresponds to Fig. A-1.

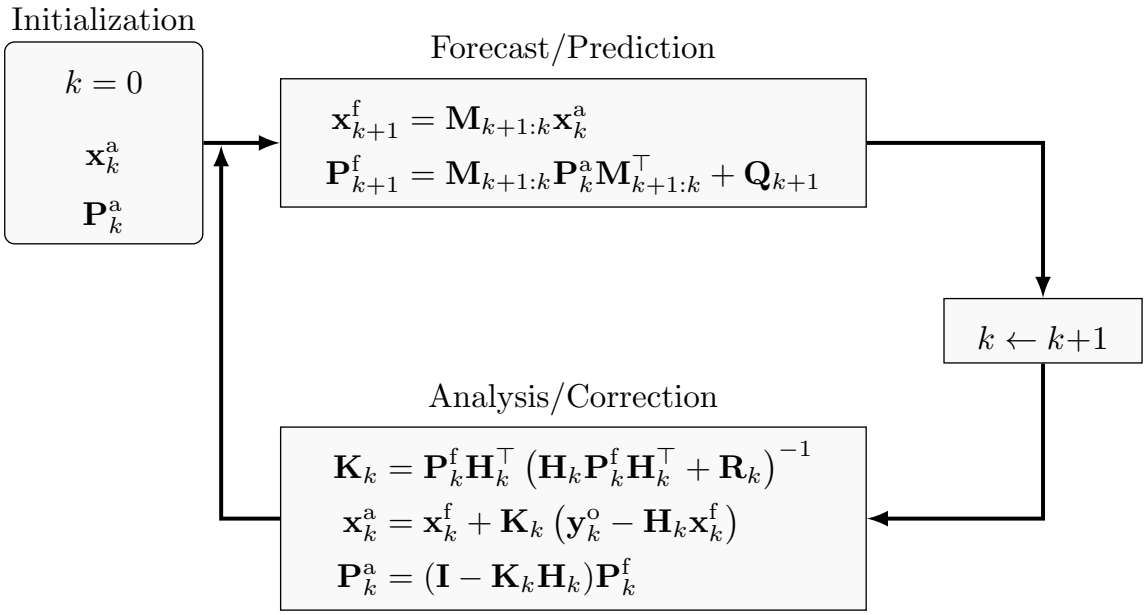


Figure A-1: Kalman Filter algorithm. The initialization is made with the analysed state.

A.2 Ensemble Kalman filter algorithm

An efficient implementation of the EnKF relying on anomaly matrices is given in Algo. 1. We have used the secant method described in Asch et al., 2016 to change the definition of the variable $\mathbf{Y}_k^{\mathrm{f}}$.

Algorithm 1: Stochastic Ensemble Kalman Filter (slightly adapted from Asch et al., 2016). Use of anomaly matrices with $\mathbf{Y}_k^{\mathrm{f}} = \mathbf{H}_k \mathbf{X}_k^{\mathrm{f}}$.

Input: For $k = 0, \ldots, K$: the forward models $\mathcal{M}_{k:k-1}$, the observation models $\mathcal{H}_k$, the observation error covariance matrices $\mathbf{R}_k$

Output: $\{\mathbf{x}_k^{\mathrm{a},(i)}\}\,;\ k = 0, \cdots, K\,;\ i = 1, \cdots, N_{\mathrm{e}}$

begin

Initialize the ensemble of forecasts $\{\mathbf{x}_0^{\mathrm{f},(i)}\}\,;\ i = 1, \cdots, N_{\mathrm{e}}$

for $k = 0, \ldots, K$ **do**

Draw a statistically consistent observation set$\,;\ i = 1, \cdots, N_{\mathrm{e}}$

$$\mathbf{y}_k^{\mathrm{o},(i)} = \mathbf{y}_k^{\mathrm{o}} + \epsilon_k^{\mathrm{o},(i)} \quad \text{with} \quad \epsilon_k^{\mathrm{o},(i)} \sim \mathcal{N}(0, \mathbf{R}_k)$$

Compute the model counterparts of the observation set$\,;\ i = 1, \cdots, N_{\mathrm{e}}$

$$\mathbf{y}_k^{\mathrm{f},(i)} = \mathcal{H}_k\left(\mathbf{x}_k^{\mathrm{f},(i)}\right)$$

Compute the ensemble means

$$\overline{\mathbf{x}_k^{\mathrm{f}}} = \frac{1}{N_{\mathrm{e}}} \sum_{i=1}^{N_{\mathrm{e}}} \mathbf{x}_k^{\mathrm{f},(i)}\,;\ \overline{\mathbf{y}_k^{\mathrm{f}}} = \frac{1}{N_{\mathrm{e}}} \sum_{i=1}^{N_{\mathrm{e}}} \mathbf{y}_k^{\mathrm{f},(i)}\,;\ \overline{\epsilon_k^{\mathrm{o}}} = \frac{1}{N_{\mathrm{e}}} \sum_{i=1}^{N_{\mathrm{e}}} \epsilon_k^{\mathrm{o},(i)}$$

Compute the normalized anomalies$\,;\ i = 1, \cdots, N_{\mathrm{e}}$

$$\left[\mathbf{X}_k^{\mathrm{f}}\right]_{:,i} = \frac{\mathbf{x}_k^{\mathrm{f},(i)} - \overline{\mathbf{x}_k^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}}\,;\ \left[\mathbf{Y}_k^{\mathrm{f}}\right]_{:,i} = \frac{\mathbf{y}_k^{\mathrm{f},(i)} - \overline{\mathbf{y}_k^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}}\,;\ \left[\mathbf{E}_k^{\mathrm{o}}\right]_{:,i} = \frac{\epsilon_k^{\mathrm{o},(i)} - \overline{\epsilon_k^{\mathrm{o}}}}{\sqrt{N_{\mathrm{e}} - 1}}$$

Compute the Kalman gain

$$\mathbf{K}_k^{\mathrm{e}} = \mathbf{X}_k^{\mathrm{f}} \left(\mathbf{Y}_k^{\mathrm{f}}\right)^\top \left(\mathbf{Y}_k^{\mathrm{f}} \left(\mathbf{Y}_k^{\mathrm{f}}\right)^\top + \mathbf{E}_k^{\mathrm{o}} \left(\mathbf{E}_k^{\mathrm{o}}\right)^\top\right)^{-1}$$

Update the ensemble$\,;\ i = 1, \cdots, N_{\mathrm{e}}$

$$\mathbf{x}_k^{\mathrm{a},(i)} = \mathbf{x}_k^{\mathrm{f},(i)} + \mathbf{K}_k^{\mathrm{e}}\left(\mathbf{y}_k^{\mathrm{o},(i)} - \mathbf{y}_k^{\mathrm{f},(i)}\right)$$

Compute the ensemble forecast$\,;\ i = 1, \cdots, N_{\mathrm{e}}$

$$\mathbf{x}_{k+1}^{\mathrm{f},(i)} = \mathcal{M}_{k+1:k}(\mathbf{x}_k^{\mathrm{a},(i)})$$

A.3 Dual Ensemble Kalman filter algorithm

An efficient implementation of the Dual EnKF relying on anomaly matrices is given in Algo. 2. We have slightly adapted this algorithm from Moradkhani et al., 2005.

Algorithm 2: Dual Ensemble Kalman Filter (slightly adapted from Moradkhani et al., 2005). Use of anomaly matrices with $\mathbf{Y}_k^{\mathrm{f}} = \mathbf{H}_k \mathbf{X}_k^{\mathrm{f}}$. We have $i = 1, \cdots, N_{\mathrm{e}}$.

Input: For $k = 1, \ldots, K$: the forward models $\mathcal{M}_{k:k-1}$, the observation models $\mathcal{H}_k$, the observation error covariance matrices $\mathbf{R}_k$

Output: $\{\theta_k^{\mathrm{a},(i)}\}$ and $\{\mathbf{x}_k^{\mathrm{a},(i)}\}$; $k = 0, \cdots, K$

begin

1: Initialize $\{\theta_0^{\mathrm{a},(i)}\}$ and $\{\mathbf{x}_0^{\mathrm{a},(i)}\}$

for $k = 1, \ldots, K$ **do**

2: Observation ensemble:

$$\mathbf{y}_k^{\mathrm{o},(i)} = \mathbf{y}_k^{\mathrm{o}} + \epsilon_k^{\mathrm{o},(i)} \quad \text{with} \quad \epsilon_k^{\mathrm{o},(i)} \sim \mathcal{N}(0, \mathbf{R}_k)$$

$$\mathbf{R}_k^{\mathrm{e}} = \frac{1}{N_{\mathrm{e}} - 1} \sum_{i=1}^{N_{\mathrm{e}}} \epsilon_k^{\mathrm{o},(i)} \left(\epsilon_k^{\mathrm{o},(i)}\right)^{\top}$$

3: Parameter forecast:

$$\theta_k^{\mathrm{f},(i)} = \theta_{k-1}^{\mathrm{a},(i)} + \tau_k^{(i)} \quad \text{with} \quad \tau_k^{(i)} \sim \mathcal{N}(0, \mathbf{\Sigma}_k^{\theta})$$

$$\mathbf{x}_k^{\mathrm{f},(i)} = \mathcal{M}_{k:k-1}(\mathbf{x}_{k-1}^{\mathrm{a},(i)}, \theta_k^{\mathrm{f},(i)})$$

$$\mathbf{y}_k^{\mathrm{f},(i)} = \mathcal{H}_k\left(\mathbf{x}_k^{\mathrm{f},(i)}\right)$$

4: Compute the normalized anomalies

$$\left[\mathbf{\Theta}_k^{\mathrm{f}}\right]_{:,i} = \frac{\theta_k^{\mathrm{f},(i)} - \overline{\theta_k^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}} \; ; \; \left[\mathbf{Y}_k^{\mathrm{f}}\right]_{:,i} = \frac{\mathbf{y}_k^{\mathrm{f},(i)} - \overline{\mathbf{y}_k^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}} \; ; \; \left[\mathbf{E}_k^{\mathrm{o}}\right]_{:,i} = \frac{\epsilon_k^{\mathrm{o},(i)} - \overline{\epsilon_k^{\mathrm{o}}}}{\sqrt{N_{\mathrm{e}} - 1}}$$

5: Parameter update:

$$\mathbf{K}_k^{\theta,\mathrm{e}} = \mathbf{\Theta}_k^{\mathrm{f}} \left(\mathbf{Y}_k^{\mathrm{f}}\right)^{\top} \left(\mathbf{Y}_k^{\mathrm{f}} \left(\mathbf{Y}_k^{\mathrm{f}}\right)^{\top} + \mathbf{E}_k^{\mathrm{o}} \left(\mathbf{E}_k^{\mathrm{o}}\right)^{\top}\right)^{-1}$$

$$\theta_k^{\mathrm{a},(i)} = \theta_k^{\mathrm{f},(i)} + \mathbf{K}_k^{\theta,\mathrm{e}} \left(\mathbf{y}_k^{\mathrm{o},(i)} - \mathbf{y}_k^{\mathrm{f},(i)}\right)$$

6: State forecast:

$$\mathbf{x}_k^{\mathrm{f},(i)} = \mathcal{M}_{k:k-1}(\mathbf{x}_{k-1}^{\mathrm{a},(i)}, \theta_k^{\mathrm{a},(i)})$$

$$\mathbf{y}_k^{\mathrm{f},(i)} = \mathcal{H}_k\left(\mathbf{x}_k^{\mathrm{f},(i)}\right)$$

7: Compute the normalized anomalies

$$\left[\mathbf{X}_k^{\mathrm{f}}\right]_{:,i} = \frac{\mathbf{x}_k^{\mathrm{f},(i)} - \overline{\mathbf{x}_k^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}} \; ; \; \left[\mathbf{Y}_k^{\mathrm{f}}\right]_{:,i} = \frac{\mathbf{y}_k^{\mathrm{f},(i)} - \overline{\mathbf{y}_k^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}} \; ; \; \left[\mathbf{E}_k^{\mathrm{o}}\right]_{:,i} = \frac{\epsilon_k^{\mathrm{o},(i)} - \overline{\epsilon_k^{\mathrm{o}}}}{\sqrt{N_{\mathrm{e}} - 1}}$$

8: State update:

$$\mathbf{K}_k^{x,\mathrm{e}} = \mathbf{X}_k^{\mathrm{f}} \left(\mathbf{Y}_k^{\mathrm{f}}\right)^{\top} \left(\mathbf{Y}_k^{\mathrm{f}} \left(\mathbf{Y}_k^{\mathrm{f}}\right)^{\top} + \mathbf{E}_k^{\mathrm{o}} \left(\mathbf{E}_k^{\mathrm{o}}\right)^{\top}\right)^{-1}$$

$$\mathbf{x}_k^{\mathrm{a},(i)} = \mathbf{x}_k^{\mathrm{f},(i)} + \mathbf{K}_k^{x,\mathrm{e}} \left(\mathbf{y}_k^{\mathrm{o},(i)} - \mathbf{y}_k^{\mathrm{f},(i)}\right)$$

Algorithm 3: Multigrid EnKF algorithm. We have $i = 1, \cdots, N_{\mathrm{e}}$.

begin

Initialize $\left\{ \left(\mathbf{x}_0^{\mathrm{F}}\right)^{\mathrm{a}}, \overline{\boldsymbol{\theta}_0^{\mathrm{a}}}, \boldsymbol{\theta}_0^{\mathrm{a},(i)}, \overline{\boldsymbol{\psi}_0^{\mathrm{a}}}, \boldsymbol{\psi}_0^{\mathrm{a},(i)} \left(\mathbf{x}_0^{\mathrm{C}}\right)^{\mathrm{a},(i)} \right\}$

for $k = 1, \ldots, K$ **do**

Fine grid forecast:

$$\left(\mathbf{x}_k^{\mathrm{F}}\right)^{\mathrm{f}} = \mathcal{M}_{k:k-1}^{\mathrm{F}} \left(\left(\mathbf{x}_{k-1}^{\mathrm{F}}\right)^{\mathrm{a}}, \overline{\boldsymbol{\theta}_k^{\mathrm{a}}} \right)$$

Apply Inner/Outer loop on the coarse grid(Algo. 4)

if *Observation available* **then**

Projection on the coarse grid

$$\left(\mathbf{x}_k^{\mathrm{C}}\right)^* = \Pi_{\mathrm{C}} \left(\left(\mathbf{x}_k^{\mathrm{F}}\right)^{\mathrm{f}} \right)$$

Fine grid state correction using the ensemble statistics:

$$\left(\mathbf{x}_k^{\mathrm{C}}\right)' = \left(\mathbf{x}_k^{\mathrm{C}}\right)^* + \left(\mathbf{K}_k^{\mathrm{C}}\right)^{x,\mathrm{e}} \left[\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{o}} - \mathcal{H}_k^{\mathrm{C}} \left(\left(\mathbf{x}_k^{\mathrm{C}}\right)^* \right) \right]$$
$$\left(\mathbf{x}_k^{\mathrm{F}}\right)' = \left(\mathbf{x}_k^{\mathrm{F}}\right)^{\mathrm{f}} + \Pi_{\mathrm{F}} \left(\left(\mathbf{x}_k^{\mathrm{C}}\right)' - \left(\mathbf{x}_k^{\mathrm{C}}\right)^* \right)$$

$\left(\mathbf{x}_k^{\mathrm{F}}\right)^{\mathrm{a}}$ is obtained through a matrix-splitting iterative procedure starting from $\left(\mathbf{x}_k^{\mathrm{F}}\right)'$.

Algorithm 4: Outer/Inner loop of the MGEnKF applied on the coarse grid. Inspired from Algo. 2. We have $i = 1, \cdots, N_{\mathrm{e}}$.

Input: For $k = 1, \ldots, K$: $\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a},(i)}$, $\boldsymbol{\theta}_{k-1}^{\mathrm{a},(i)}$, $\boldsymbol{\psi}_{k-1}^{\mathrm{a},(i)}$, $\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{o}}$, $\left(\mathbf{x}_k^{\mathrm{C}}\right)^{*}$, $\mathcal{M}_{k:k-1}^{\mathrm{C}}$, $\mathcal{C}_{k:k-1}$, $\left(\mathcal{H}_k^{\mathrm{C}}\right)^{\mathrm{O}}$, $\left(\mathcal{H}_k^{\mathrm{C}}\right)^{\mathrm{SO}}$, $\left(\mathbf{R}_k^{\mathrm{C}}\right)^{\mathrm{O}}$, $\left(\mathbf{R}_k^{\mathrm{C}}\right)^{\mathrm{SO}}$

Output: $\boldsymbol{\theta}_k^{\mathrm{a},(i)}$, $\boldsymbol{\psi}_k^{\mathrm{a},(i)}$, $\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{a},(i)}$; $k = 0, \cdots, K$

begin

1: Forecast:

$$\boldsymbol{\theta}_k^{\mathrm{f},(i)} = \boldsymbol{\theta}_{k-1}^{\mathrm{a},(i)} + \boldsymbol{\tau}_k^{\theta,(i)} \quad \text{with} \quad \boldsymbol{\tau}_k^{\theta,(i)} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_k^{\theta})$$

$$\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)} = \mathcal{M}_{k:k-1}^{\mathrm{C}}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a},(i)}, \boldsymbol{\theta}_k^{\mathrm{f},(i)}\right) + \mathcal{C}_{k:k-1}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a},(i)}, \boldsymbol{\psi}_{k-1}^{\mathrm{a},(i)}\right)$$

Apply Inner loop (Algo. 5)

if *Observation available* **then**

$$\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)} = \left(\mathcal{H}_k^{\mathrm{C}}\right)^{\mathrm{O}}\left(\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)}\right)$$

2: Observation ensemble:

$$\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{O},(i)} = \left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{O}} + \left(\boldsymbol{\epsilon}_k^{\mathrm{C}}\right)^{\mathrm{O},(i)} \quad \text{with} \quad \left(\boldsymbol{\epsilon}_k^{\mathrm{C}}\right)^{\mathrm{O},(i)} \sim \mathcal{N}(0, \left(\mathbf{R}_k^{\mathrm{C}}\right)^{\mathrm{O}})$$

3: Compute the normalized anomalies

$$\left[\boldsymbol{\Theta}_k^{\mathrm{f}}\right]_{:,i} = \frac{\boldsymbol{\theta}_k^{\mathrm{f},(i)} - \overline{\boldsymbol{\theta}_k^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}} \; ; \; \left[\mathbf{Y}_k^{\mathrm{f}}\right]_{:,i} = \frac{\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)} - \overline{\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}}$$

4: Parameter update:

$$\left(\mathbf{K}_k^{\mathrm{C}}\right)^{\theta,\mathrm{e}} = \boldsymbol{\Theta}_k^{\mathrm{f}}\left(\mathbf{Y}_k^{\mathrm{f}}\right)^{\top}\left(\mathbf{Y}_k^{\mathrm{f}}\left(\mathbf{Y}_k^{\mathrm{f}}\right)^{\top} + \left(\mathbf{R}_k^{\mathrm{C}}\right)^{\mathrm{O}}\right)^{-1}$$

$$\boldsymbol{\theta}_k^{\mathrm{a},(i)} = \boldsymbol{\theta}_k^{\mathrm{f},(i)} + \left(\mathbf{K}_k^{\mathrm{C}}\right)^{\theta,\mathrm{e}}\left(\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{O},(i)} - \left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)}\right)$$

5: State re-forecast (only if pure dual formulation is retained, otherwise the state update can be evaluated with the first forecast $\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)}$) :

$$\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)} = \mathcal{M}_{k:k-1}^{\mathrm{C}}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a},(i)}, \boldsymbol{\theta}_k^{\mathrm{a},(i)}\right) + \mathcal{C}_{k:k-1}\left(\left(\mathbf{x}_{k-1}^{\mathrm{C}}\right)^{\mathrm{a},(i)}, \boldsymbol{\psi}_{k-1}^{\mathrm{a},(i)}\right)$$

$$\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)} = \left(\mathcal{H}_k^{\mathrm{C}}\right)^{\mathrm{O}}\left(\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)}\right)$$

6: Compute the normalized anomalies

$$\left[\mathbf{X}_k^{\mathrm{f}}\right]_{:,i} = \frac{\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)} - \overline{\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}} \; ; \; \left[\mathbf{Y}_k^{\mathrm{f}}\right]_{:,i} = \frac{\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)} - \overline{\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}}$$

7: State update:

$$\left(\mathbf{K}_k^{\mathrm{C}}\right)^{x,\mathrm{e}} = \mathbf{X}_k^{\mathrm{f}}\left(\mathbf{Y}_k^{\mathrm{f}}\right)^{\top}\left(\mathbf{Y}_k^{\mathrm{f}}\left(\mathbf{Y}_k^{\mathrm{f}}\right)^{\top} + \mathbf{E}_k^{\mathrm{o}}\left(\mathbf{E}_k^{\mathrm{o}}\right)^{\top}\right)^{-1}$$

$$\left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{a},(i)} = \left(\mathbf{x}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)} + \left(\mathbf{K}_k^{\mathrm{C}}\right)^{x,\mathrm{e}}\left(\left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{O},(i)} - \left(\mathbf{y}_k^{\mathrm{C}}\right)^{\mathrm{f},(i)}\right)$$

Algorithm 5: Inner loop of the MGEnKF algorithm applied on the coarse mesh. We have $i = 1, \cdots, N_{\mathrm{e}}$.

Input: For $k = 1, \ldots, K$: the ensemble of parameters related to the model correction term $\boldsymbol{\psi}_{k-1}^{\mathrm{a},(\mathrm{i})}$, the ensemble coarse grid forecast $(\mathbf{x}_k^{\mathrm{C}})^{\mathrm{f},(i)}$, the surrogate observation $(\mathbf{x}_k^{\mathrm{C}})^{*}$, the surrogate observation operator $(\mathcal{H}_k^{\mathrm{C}})^{\mathrm{SO}}$, the surrogate observation error covariance matrix $(\mathbf{R}_k^{\mathrm{C}})^{\mathrm{SO}}$

Output: $\boldsymbol{\psi}_k^{\mathrm{a},(\mathrm{i})}$; $k = 0, \cdots, K$

begin

Parameter forecast:

$$\boldsymbol{\psi}_k^{\mathrm{f},(i)} = \boldsymbol{\psi}_{k-1}^{\mathrm{a},(i)} + \tau_k^{\psi,(i)} \quad \text{with} \quad \tau_k^{\psi,(i)} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_k^{\psi})$$

if *Surrogate observation available* **then**

$$(\mathbf{y}_k^{\mathrm{C}})^{\mathrm{f},(i)} = (\mathcal{H}_k^{\mathrm{C}})^{\mathrm{SO}} \left((\mathbf{x}_k^{\mathrm{C}})^{\mathrm{f},(i)} \right)$$

Observation ensemble:

$$(\mathbf{y}_k^{\mathrm{C}})^{\mathrm{SO},(i)} = (\mathbf{y}_k^{\mathrm{C}})^{\mathrm{SO}} + (\epsilon_k^{\mathrm{C}})^{\mathrm{SO},(i)} \quad \text{with} \quad (\epsilon_k^{\mathrm{C}})^{\mathrm{SO},(i)} \sim \mathcal{N}(0, (\mathbf{R}_k^{\mathrm{C}})^{\mathrm{SO}})$$

Compute the normalized anomalies

$$\left[\boldsymbol{\Psi}_k^{\mathrm{f}}\right]_{:,i} = \frac{\boldsymbol{\psi}_k^{\mathrm{f},(i)} - \overline{\boldsymbol{\psi}_k^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}} \; ; \; \left[\mathbf{Y}_k^{\mathrm{f}}\right]_{:,i} = \frac{(\mathbf{y}_k^{\mathrm{C}})^{\mathrm{f},(i)} - \overline{(\mathbf{y}_k^{\mathrm{C}})^{\mathrm{f}}}}{\sqrt{N_{\mathrm{e}} - 1}}$$

Parameter update:

$$(\mathbf{K}_k^{\mathrm{C}})^{\psi,\mathrm{e}} = \boldsymbol{\Psi}_k^{\mathrm{f}} \left(\mathbf{Y}_k^{\mathrm{f}}\right)^{\top} \left(\mathbf{Y}_k^{\mathrm{f}} \left(\mathbf{Y}_k^{\mathrm{f}}\right)^{\top} + (\mathbf{R}_k^{\mathrm{C}})^{\mathrm{SO}} \right)^{-1}$$

$$\boldsymbol{\psi}_k^{\mathrm{a},(i)} = \boldsymbol{\psi}_k^{\mathrm{f},(i)} + (\mathbf{K}_k^{\mathrm{C}})^{\psi,\mathrm{e}} \left((\mathbf{y}_k^{\mathrm{C}})^{\mathrm{SO},(i)} - (\mathbf{y}_k^{\mathrm{C}})^{\mathrm{f},(i)} \right)$$

A.4 Multigrid Ensemble Kalman filter algorithm

The algorithm 3 represents a simplified, ready-to-use application of the conceptual methodology presented in Chapter 4.

First, when observation is not available, the two main forecast operations (fine grid forecast and ensemble coarse forecast) are performed using explicit time advancement schemes. This choice allows reducing the computational costs. However, when observation is available, the following strategies are employed:

1. The two forecast operations (main simulation and ensemble members) are performed using an implicit matrix-splitting iterative procedure, using a single iteration.

2. The number of iterative solutions for the main simulation on the coarse-grid level is equal to zero. That is, the solution from the first forecast is projected on the coarse grid, and the difference between the KF state estimation and this forecast is re-projected over the fine grid.

3. In the final iteration on the fine grid, an implicit matrix-splitting iterative procedure is employed, using a single iteration and a relaxation coefficient $\alpha = 0.5$. This is a general recommendation, but the value of the relaxation coefficient will depend on the test case and on the stability of the simulation.

Bibliography

Anderson, B. D. O., & Moore, J. B. (1979). *Optimal filtering*. Prentice-Hall.

Aris, R. (1990). *Vectors, tensors and the basic equations of fluid mechanics*. Dover Publications.

Artana, G., Cammilleri, A., Carlier, J., & Mémin, E. (2012). Strong and weak constraint variational assimilations for reduced order fluid flow modeling. *Journal of Computational Physics*, *231* (8), 3264–3288.

Asch, M., Bocquet, M., & N., M. (2016). *Data Assimilation: Methods, Algorithms, and Applications*. Society for Industrial; Applied Mathematics.

Bain, A., & Crisan, D. (2009). *Fundamentals of stochastic filtering*. Springer New York.

Balachandarb, S., Mittal, R., & Najjar, F. M. (1997). Properties of the mean recirculation region in the wakes of two-dimensional bluff bodies. *Journal of Fluid Mechanics*, *351*, 167–199.

Barrett, R., Berry, M., Chan, T. F., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C., & van der Vorst, H. (1994). *Templates for the solution of linear systems: Building blocks for iterative methods*. Society for Industrial; Applied Mathematics.

Barros, D., Borée, J., Noack, B. R., Spohn, A., & Ruiz, T. (2016). Bluff body drag manipulation using pulsed jets and coanda effect. *Journal of Fluid Mechanics*, *805*, 422–459.

Bewley, T. R., & Liu, S. (1998). Optimal and robust control and estimation of linear paths to transition. *Journal of Fluid Mechanics*, *365*, 305–349.

Bewley, T. R., Moin, P., & Temam, R. (2001). DNS-based predictive control of turbulence: An optimal benchmark for feedback algorithms. *Journal of Fluid Mechanics*, *447*, 179–225.

Bogey, C., & Bailly, C. (2004). A family of low dispersive and low dissipative explicit schemes for flow and noise computations. *Journal of Computational Physics*, *194*, 194–214.

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2021). Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200086.

Brandt, A. (1977). Multi-level adaptive solutions to boundary-value problems. *Mathematics of Computation*, *31*, 333–390.

Bruno, L., Coste, N., & Fransos, D. (2012). Simulated flow around a rectangular 5:1 cylinder: Spanwise discretisation effects and emerging flow features. *J. Wind Eng. Ind. Aerodyn.*, *104-106*, 203–215.

Bruno, L., Salvetti, M. V., & Ricciardelli, F. (2014). Benchmark on the Aerodynamics of a Rectangular 5:1 Cylinder: An overview after the first four years of activity. *J. Wind Eng. Ind. Aerodyn.*, *126*, 87–106.

Brunton, S. L., & Noack, B. R. (2015a). Closed-Loop Turbulence Control: Progress and Challenges. *Applied Mechanics Review*, *67*, 050801.

Brunton, S. L., & Noack, B. R. (2015b). Closed-Loop Turbulence Control: Progress and Challenges [050801]. *Applied Mechanics Reviews*, *67*(5).

Buis, S., Piacentini, A., & Déclat, D. (2006). Palm: A computational framework for assembling high-performance computing applications. *Concurrency and Computation: Practice and Experience*, *18*(2), 231–245.

Burgers, G., Van Leeuwen, P. J., & Evensen, G. (1998). On the analysis scheme in the ensemble kalman filter. *Monthly Weather Review*, *126*.

Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, *9*(5).

Chandramouli, P., Memin, E., & Heitz, D. (2020). 4D large scale variational data assimilation of a turbulent flow with a dynamics error model. *Journal of Computational Physics*, *412*, 109446.

Chevalier, M., Hœpffner, J., Bewley, T. R., & Henningson, D. S. (2006). State estimation in wall-bounded flow systems. part 2. turbulent flows. *Journal of Fluid Mechanics*, *552*(-1), 167.

Choi, H., Jeon, W.-P., & Kim, J. (2008). Control of flow over a bluff body. *Annual Review of Fluid Mechanics*, *40*(1), 113–139.

Choi, H., Lee, J., & Park, H. (2014). Aerodynamics of heavy vehicles. *Annual Review of Fluid Mechanics*, *46*(1), 441–468.

Colburn, C. H., Cessna, J. B., & Bewley, T. R. (2011). State estimation in wall-bounded flow systems. part 3. the ensemble kalman filter. *Journal of Fluid Mechanics*, *682*, 289–303.

Courant, R., & Hilbert, D. (1989). *Methods of mathematical physics*. Wiley.

D'adamo, J., Papadakis, N., Mémin, E., & Artana, G. (2007). Variational assimilation of POD low-order dynamical systems. *Journal of Turbulence*, *8*, N9.

Deardorff, J. W. (1970). A numerical study of three-dimensional turbulent channel flow at large reynolds numbers. *Journal of Fluid Mechanics*, *41*(2), 453–480.

Domarazdki, J. A. (2010). Large eddy simulations without explicit eddy viscosity models. *Int. J. Comput. Fluid Dyn.*, *24(10)*, 435–447.

Doucet, A., & Johansen, A. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In *Oxford handbook of nonlinear filtering,* (pp. 656–704). Oxford University Press.

Duchaine, F., Jauré, S., Poitou, D., Quémerais, E., Staffelbach, G., Morel, T., & Gicquel, L. (2015). Analysis of high performance conjugate heat transfer with the OpenPALM coupler. *Computational Science & Discovery*, *8*(1), 015003.

Eswaran, V., & Pope, S. B. (1988). Direct numerical simulations of the turbulent mixing of a passive scalar. *The Physics of Fluids*, *31*(3), 506–520.

Eswaran, V., & Pope, S. (1988). An examination of forcing in direct numerical simulations of turbulence. *Computers & Fluids*, *16*(3), 257–278.

Eurostat. (2022). Simplified energy statistics.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using monte-carlo methods to forecast error statistics. *Journal of Geophysical Research*, *99*, 10143–10162.

Evensen, G. (2009). *Data Assimilation*. Springer Berlin Heidelberg.

Falgout, R. (2006). An introduction to algebraic multigrid. *Computing in Science Engineering*, *8*(6), 24–33.

Fedorov, A. (2011). Transition and stability of high-speed boundary layers. *Annual Review of Fluid Mechanics*, *43*(1), 79–95.

Ferziger, J., & Peric, M. (1996). *Computational methods in fluid dynamics*. New-York : Springer-Verlag.

Fischer, P. F., Lottes, J. W., & Kerkemeier, S. (2008). nek5000 Web page.

Fossum, K., Mannseth, T., & Stordal, A. S. (2020). Assessment of multilevel ensemble-based data assimilation for reservoir history matching. *Computational Geosciences*, *24*, 217–239.

Foures, D. P. G., Dovetta, N., Sipp, D., & Schmid, P. J. (2014). A data-assimilation method for reynolds-averaged navier–stokes-driven mean flow reconstruction. *Journal of Fluid Mechanics*, *759*, 404–431.

Garthwaite, P., Jolliffe, I., Jolliffe, I., & Jones, B. (2002). *Statistical Inference*. Oxford University Press.

Gaspari, G., & Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, *125*(554), 723–757.

Gorodetsky, A. A., Geraci, G., Eldred, M. S., & Jakeman, J. D. (2020). A generalized approximate control variate framework for multifidelity uncertainty quantification. *Journal of Computational Physics*, *408*, 109257.

Grandemange, M., Gohlke, M., & Cadot, O. (2013). Turbulent wake past a three-dimensional blunt body. part 1. global modes and bi-stability. *Journal of Fluid Mechanics*, *722*, 51–84.

Gronskis, A., Heitz, D., & Mémin, E. (2013). Inflow and initial conditions for direct numerical simulation based on adjoint data assimilation. *Journal of Computational Physics*, *242*, 480–497.

Hackbusch, W. (1985). *Multi-grid methods and applications*. Springer Berlin.

Haffner, Y., Borée, J., Spohn, A., & Castelain, T. (2020). Mechanics of bluff body drag reduction during transient near-wake reversals. *Journal of Fluid Mechanics*, *894*, A14.

Hirsch, C. (2007). *Numerical computation of internal and external flows (second edition)* (Second Edition). Butterworth-Heinemann.

Hoel, H., Law, K. J. H., & Tempone, R. (2016). Multilevel ensemble kalman filtering. *SIAM J. Numer. Anal.*, *54(3)*, 1813–1839.

Hœpffner, J., Chevalier, M., Bewley, T. R., & Henningson, D. S. (2005). State estimation in wall-bounded flow systems. part 1. perturbed laminar flows. *Journal of Fluid Mechanics*, *534*, 263–294.

Högberg, M., Bewley, T. R., & Henningson, D. S. (2003). Linear feedback control and estimation of transition in plane channel flow. *Journal of Fluid Mechanics*, *481*, 149–175.

Holzmann, T. (2017). Openfoam guide/the pimple algorithm in openfoam [https://openfoamwiki.net/index.php/OpenFOAM_guide].

Houtekamer, P. L., & Zhang, F. (2016). Review of the ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, *144*(12), 4489–4532.

Ide, K., Ghil, M., & Lorenc, A. (1999). Unified notation for data assimilation: Operational, sequential and variational. *Journal of the Meteorological Society of Japan*, *75*.

Jasak, H. (1996). *Error analysis and estimation for the finite volume method with applications to fluid flows* (Doctoral dissertation). Imperial College London (University of London
PhD Thesis.

Jazwinski, A. H. (1970). *Stochastic processes and filtering theory*. Academic Press.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35–45.

Kato, H., & Obayashi, S. (2013). Approach for uncertainty of turbulence modeling based on data assimilation technique. *Computers & Fluids*, *85*, 2–7.

Kato, H., Yoshizawa, A., Ueno, G., & Obayashi, S. (2015). A data assimilation methodology for reconstructing turbulent flows around aircraft. *Journal of Computational Physics*, *283*, 559–581.

Ko, J., Lucor, D., & Sagaut, P. (2008). Sensitivity of two-dimensional spatially developing mixing layers with respect to uncertain inflow conditions. *Physics Fluids*, *20*, 077102.

Kreyszig, E. (1978). *Introductory functional analysis with applications*. Wiley.

Kumar, N. (2021). *Data-driven flow modelling using machine learning and data assimilation approaches* (Doctoral dissertation). École Doctorale Sciences et Ingénierie des Matériaux, Mécanique, Énergétique.

Labahn, J. W., Wu, H., Coriton, B., Frank, J. H., & 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*, 2259–2266.

Law, K., Hoel, H., Tempone, R., Nobile, F., & Chernov, A. (2020). Multilevel ensemble kalman filtering for spatio-temporal processes. *NUMERISCHE MATHEMATIK*.

Le Maître, O. P., & Knio, O. M. (2010). *Spectral methods for uncertainty quantification*. Springer Netherlands.

Lehmkuhl, O., Piomelli, U., & Houzeaux, G. (2019). On the extension of the integral length-scale approximation model to complex geometries. *International Journal of Heat and Fluid Flow*, *78*, 108422.

Li, R., Barros, D., Borée, J., Cadot, O., Noack, B. R., & Cordier, L. (2016). Feedback control of bimodal wake dynamics. *Experiments in Fluids*, *57*(10), 158.

Lunghi, G., Pasqualetto, E., Rocchio, B., Mariotti, A., & Salvetti, M. V. (2022). Impact of the lateral mean recirculation characteristics on the near wake and

bulk quantities of the barc configuration. *Wind and Structures*, *34(1)*, 115–125.

Maday, Y., Patera, A. T., & Rønquist, E. M. (1990). An operator-integration-factor splitting method for time-dependent problems: Application to incompressible fluid flow. *J. Sci. Comput.*, *5 (4)*, 263–292.

Mariotti, A., Siconolfi, L., & Salvetti, M. V. (2017). Stochastic sensitivity analysis of large-eddy simulation predictions of the flow around a 5:1 rectangular cylinder. *Eur. J. Mech. B-Fluids*, *62*, 149–165.

Mathew, J., Lechner, R., Foysi, H., Setsterhenn, J., & Friedrich, R. (2003). An explicit filtering method for large eddy simulation of compressible flows. *Phys. Fluids*, *15*, 2279–2289.

McMullan, W. A., Gao, S., & Coats, C. M. (2007). A comparative study of inflow conditions for two- and three-dimensional spatially developing mixing layers using large eddy simulation. *International Journal for Numerical Methods in Fluids*, *55*(6), 589–610.

Meldi, M., & Poux, A. (2017). A reduced order model based on Kalman filtering for sequential data assimilation of turbulent flows. *Journal of Computational Physics*, *347*, 207–234.

Meldi, M., Salvetti, M. V., & Sagaut, P. (2012). Quantification of errors in large-eddy simulations of a spatially evolving mixing layer using polynomial chaos. *Physics of Fluids*, *24*, 035101.

Meldi, M. (2018). Augmented Prediction of Turbulent Flows via Sequential Estimators: Sensitivity of State Estimation to Density of Time Sampling for Available Observation. *Flow, Turbulence and Combustion*, *101*(2), 389–412.

Meyers, J., & Sagaut, P. (2006). On the model coefficients for the standard and the variational multi-scale Smagorinsky model. *Journal of Fluid Mechanics*, *569*, 287–319.

Moin, P., & Kim, J. (1982). Numerical investigation of turbulent channel flow. *Journal of Fluid Mechanics*, *118*, 341–377.

Moldovan, G., Lehnasch, G., Cordier, L., & Meldi, M. (2021). A multigrid/ensemble Kalman Filter strategy for assimilation of unsteady flows. *Journal of Computational Physics*, *443*, 110481.

Mons, V., Chassaing, J.-C., Gomez, T., & Sagaut, P. (2014). Is isotropic turbulence decay governed by asymptotic behavior of large scales? An eddy-damped quasi-normal Markovian-based data assimilation study. *Physics of Fluids*, *26*(11), 115105.

Mons, V., Chassaing, J.-C., Gomez, T., & Sagaut, P. (2016). Reconstruction of unsteady viscous flows using data assimilation schemes. *Journal of Computational Physics*, *316*, 255–280.

Mons, V., Chassaing, J.-C., & Sagaut, P. (2017). Optimal sensor placement for variational data assimilation of unsteady flows past a rotationally oscillating cylinder. *Journal of Fluid Mechanics*, *823*, 230–277.

Mons, V., Du, Y., & Zaki, T. A. (2021). Ensemble-variational assimilation of statistical data in large-eddy simulation. *Physical Review Fluids*, *6*(10), 104607.

Mons, V., & Marquet, O. (2021). Linear and nonlinear sensor placement strategies for mean-flow reconstruction via data assimilation. *Journal of Fluid Mechanics*, *923*.

Mons, V., Wang, Q., & Zaki, T. A. (2019). Kriging-enhanced ensemble variational data assimilation for scalar-source identification in turbulent environments. *Journal of Computational Physics*, *398*, 108856.

Moradkhani, H., Sorooshian, S., Gupta, H. V., & Houser, P. R. (2005). Dual state–parameter estimation of hydrological models using ensemble kalman filter. *Advances in Water Resources*, *28*(2), 135–147.

Nachi, G. (2007). Kalman filtering in the presence of state space equality constraints. *IEEE Chinese Control Conference*, 107–133.

Nisugi, K., Hayase, T., & Shirai, A. (2004). Fundamental study of hybrid wind tunnel integrating numerical simulation and experiment in analysis of flow field. *JSME International Journal Series B*, *47*(3), 593–604.

Nocedal, J., & Wright, S. J. (2006). *Numerical optimization* (2nd ed) [OCLC: ocm68629100]. Springer.

Olson, D. A., Katz, A. W., Naguib, A. M., Koochesfahani, M. M., Rizzetta, D. P., & Visbal, M. R. (2013). On the challenges in experimental characterization of flow separation over airfoils at low Reynolds number. *Experiments in Fluids*, *54*(2), 1470.

Owhadi, H., Scovel, C., Sullivan, T. J., McKerns, M., & Ortiz, M. (2013). Optimal Uncertainty Quantification. *SIAM Review*, *55*(2), 271–345.

Patera, A. T. (1984). A spectral element method for fluid dynamics: Laminar flow in a channel expansion. *Journal of Computational Physics*, *54*(3), 468–488.

Pedrizzetti, G., & Novikov, E. A. (1994). On markov modelling of turbulence. *Journal of Fluid Mechanics*, *280*, 69–93.

Piomelli, U., Rouhi, A., & Geurts, B. J. (2015). A grid-independent length scale for large-eddy simulations. *Journal of Fluid Mechanics*, *766*, 499–527.

Pope, S. B. (2000). *Turbulent flows*. Cambridge University Press.

Popov, A. A., Mou, C., Sandu, A., & Iliescu, T. (2021). A multifidelity ensemble kalman filter with reduced order control variates. *SIAM Journal on Scientific Computing*, *43*, A1134–A1162.

Reed, M., & Simon, B. (1981). *I: Functional Analysis*. Elsevier Science.

Rizzi, A., & Luckring, J. M. (2021). Historical development and use of cfd for separated flow simulations relevant to military aircraft. *Aerospace Science and Technology*, *117*, 106940.

Rocchio, B., Mariotti, A., & Salvetti, M. V. (2020). Flow around a 5:1 rectangular cylinder: Effects of upstream-edge rounding. *J. Wind Eng. Ind. Aerodyn.*, *204*, 104237.

Rochoux, M. C., Ricci, S., Lucor, D., Cuenot, B., & Trouve, A. (2015). 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*, 2951–2973.

Rochoux, M., Catherine. (2014). *Vers une meilleure prévision de la propagation d'incendies de forêt : Évaluation de modèles et assimilation de données* (Doctoral dissertation). Ecole centrale de Paris

Thèse de doctorat Mathématiques appliquées - combustion Châtenay-Malabry, Ecole centrale de Paris 2014.

Rogallo, R. S., & Moin, P. (1984). Numerical simulation of turbulent flows. *Annual Review of Fluid Mechanics*, *16*(1), 99–137.

Rogers, M. M., & Moin, P. (1987). The structure of the vorticity field in homogeneous turbulent flows. *Journal of Fluid Mechanics*, *176*, 33–66.

Ross, S. M. (2014). *Introduction to probability and statistics for engineers and scientists* (5th ed.). Academic Press.

Rouhi, A., Piomelli, U., & Geurts, B. J. (2016). Dynamic subfilter-scale stress model for large-eddy simulations. *Phys. Rev. Fluids*, *1*, 044401.

Ruhnau, P., Kohlberger, T., Schnrr, C., & Nobach, H. (2005). Variational optical flow estimation for particle image velocimetry. *Experiments in Fluids*, *38*(1), 21–32.

Ruhnau, P., & Schnörr, C. (2006). Optical Stokes flow estimation: An imaging-based control approach. *Experiments in Fluids*, *42*(1), 61–78.

Ruhnau, P., Stahl, A., & Schnörr, C. (2007). Variational estimation of experimental fluid flows with physics-based spatio-temporal regularization. *Measurement Science and Technology*, *18*(3), 755–763.

Saad, Y. (2003). *Iterative methods for sparse linear systems*. Society for Industrial; Applied Mathematics.

Sagaut, P. (2006). *Large eddy simulation for incompressible flows : An introduction*. Springer-Verlag.

Sakov, P., Olivier, D. S., & Bertino, L. (2011). An iterative enkf for strongly nonlinear systems. *Monthly Weather Review*, *140*, 1988–2004.

Simon, D., & Chia, T. L. (2002). Kalman filtering with state equality constraints. *IEEE Trans. Aerosp. Electron. Syst.*, *38*, 128–136.

Siripatana, A., Giraldi, L., Le Maître, O. P., Knio, O. M., & Hoteit, I. (2019). Combining ensemble kalman filter and multiresolution analysis for efficient assimilation into adaptive mesh models. *Computational Geosciences*, *23*, 1259–1276.

Soares, R. V., Maschio, C., & Schiozer, D. J. (2018). Applying a localization technique to kalman gain and assessing the influence on the variability of models in history matching. *Journal of Petroleum Science and Engineering*, *169*, 110–125.

Spalart, P. R., & Venkatakrishnan, V. (2016). On the role and challenges of cfd in the aerospace industry. *The Aeronautical Journal*, *120*(1223), 209–232.

Stonebridge, G. (2017). *Diagonal Approximations to the ObservationError Covariance Matrix in Sea Ice ThicknessData Assimilation* (Master of science). University of Waterloo.

Suzuki, T. (2012). Reduced-order kalman-filtered hybrid simulation combining particle tracking velocimetry and direct numerical simulation. *Journal of Fluid Mechanics*, *709*, 249–288.

Suzuki, T., Ji, H., & Yamamoto, F. (2009). Unsteady PTV velocity field past an airfoil solved with DNS: Part 1. Algorithm of hybrid simulation and hybrid velocity field at Re = 103. *Experiments in Fluids*, *47*(6), 957–976.

Suzuki, T., Ji, H., & Yamamoto, F. (2010). Instability waves in a low-reynolds-number planar jet investigated with hybrid simulation combining particle tracking velocimetry and direct numerical simulation. *Journal of Fluid Mechanics*, *655*, 344–379.

Suzuki, T., Sanse, A., Mizushima, T., & Yamamoto, F. (2009). Unsteady PTV velocity field past an airfoil solved with DNS: Part 2. Validation and application at Reynolds numbers up to Re $<$ 104. *Experiments in Fluids*, *47*(6), 977–994.

Symon, S., Dovetta, N., McKeon, B. J., Sipp, D., & Schmid, P. J. (2017). Data assimilation of mean velocity from 2D PIV measurements of flow over an idealized airfoil. *Experiments in Fluids*, *58*(5), 61.

Symon, S., Sipp, D., Schmid, P. J., & McKeon, B. J. (2020). Mean and Unsteady Flow Reconstruction Using Data-Assimilation and Resolvent Analysis. *AIAA Journal*, *58*(2), 575–588.

Talagrand, O. (2014). 4D-VAR: Four-dimensional variational assimilation. In *Advanced Data Assimilation for Geosciences*. Oxford University Press.

Talagrand, O. (1997). Assimilation of observations, an introduction (gtSpecial IssueltData assimilation in meteology and oceanography: Theory and practice). *Journal of the Meteorological Society of Japan. Ser. II*, *75*(1B), 191–209.

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M., & Zhen, Y. (2020). A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation. *Monthly Weather Review*, *148*(10), 3973–3994.

Tissot, G. (2014). *Réduction de modèle et contrôle d'écoulements* (Doctoral dissertation) [2014POIT2284].

Wesseling, P., & Oosterlee, C. W. (2001). Geometric multigrid with applications to computational fluid dynamics. *Journal of Computational and Applied Mathematics*, *128*, 311–334.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. The MIT Press.

Wilcox, D. (1994). *Turbulence Modeling for CFD*. DCW Industries, Incorporated.

Xiao, H., Wu, J.-L., Wang, J.-X., Sun, R., & 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.

Yang, Y., & Mémin, E. (2017). High-resolution data assimilation through stochastic subgrid tensor and parameter estimation from 4DEnVar. *Tellus A: Dynamic Meteorology and Oceanography*, *69*(1), 1308772.

Yang, Y., Robinson, C., Heitz, D., & Mémin, E. (2015). Enhanced ensemble-based 4DVar scheme for data assimilation. *Computers & Fluids*, *115*, 201–210.

Zhang, X., Michelin-Ströfer, C., & Xiao, H. (2020). Regularized ensemble kalman methods for inverse problems. *Journal of Computational Physics*, *416*, 109517.

Zhang, X.-L., Michelén-Ströfer, C., & Xiao, H. (2020). Regularized ensemble Kalman methods for inverse problems. *Journal of Computational Physics*, *416*, 109517.

Zhang, X.-L., Xiao, H., He, G.-W., & Wang, S.-Z. (2021). Assimilation of disparate data for enhanced reconstruction of turbulent mean flows. *Computers & Fluids*, *224*, 104962.