Ensemble-based Data Assimilation For High-uncertainty systems: a case of study with Particulate Matter in the Aburra Valley

Lopez Restrepo, S.

DOI
10.4233/uuid:1724a16e-1f6a-423d-9a16-34e1afed3067

Publication date
2021

Document Version
Final published version

Citation (APA)
Lopez Restrepo, S. (2021). *Ensemble-based Data Assimilation For High-uncertainty systems: a case of study with Particulate Matter in the Aburra Valley*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:1724a16e-1f6a-423d-9a16-34e1afed3067

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Ensemble-based Data Assimilation For High-uncertainty systems: a case of study with Particulate Matter in the Aburrá Valley

Ensemble-based Data Assimilation For High-uncertainty systems: a case of study with Particulate Matter in the Aburrá Valley

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 9 September 2021 om 10.00 uur

door

Santiago LOPEZ RESTREPO

Master of Engineering in Chemical Engineering„
Universidad Nacional de Colombia, Colombia
geboren te Bello, Colombia.

Dit proefschrift is goedgekeurd door de promotoren.

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof.dr.ir. A.W. Heemink	Technische Universiteit Delft, promotor
Prof.dr. O.L. Quintero Montoya	Universidad EAFIT, promotor
Onafhankelijke leden:	
Prof.dr.ir. H.W.J. Russchenberg	Technische Universiteit Delft
Prof.dr.ir. M. Verlaan	Technische Universiteit Delft/Deltares
Prof.dr. M.C. Krol	Wageningen University & Research
Dr.ir. G.J.M. Velders	Universiteit Utrecht
Overige lid:	
Dr.ir. A.J. Segers	TNO

Het promotieonderzoek is uitgevoerd in het kader van een overeenkomst over gezamenlijke promotiebegeleiding tussen Universidad EAFIT, Colombia en Technische Universiteit Delft, Nederland.

Keywords: Data Assimilation, Chemical Transport Model, Ensemble-based methods, Covariance Estimation

Author email: s.lopezrestrepo@tudelft.nl, slopezr2@eafit.edu.co

ISBN 000-00-0000-000-0

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

All models are wrong,
but some are useful

George E.P. Box

Contents

Summary

In order to avoid the adverse effects of air pollution, efforts have been made to monitor when air pollution reaches dangerous levels. A Chemical Transport Model (CTM) can simulate trace gases and particles concentration in specific areas. These models are not entirely reliable, owing to incomplete knowledge about emissions and meteorological conditions. Explaining and predicting variability in air quality models remains a challenge. In this thesis we want to demonstrate that data assimilation (DA) can reduce uncertainty in the model process. DA is a mathematical family of techniques in which observed values are combined with a dynamic model to improve the accuracy of the model.

Standard DA methods have limitations when there is not a complete characterization of the uncertainties. In air quality applications, emission inventories' accuracy is often low, and weather models often do not predict events very well. The problem is worse in developing countries where the knowledge available is sparse and of relatively low quality. The thesis's main contribution is the development of a DA systems for improving the behavior of complex models in the presence of high uncertainty. The proposed methods and developments have been tested in the framework of the LOTOS-EUROS CTM with applications to forecast particular matter in the Aburrá Valley in Colombia. The use of a less expensive monitoring network is also discussed. The Aburrá valley represents a good testing scenario because of its current air quality issues, the difficulty of its terrain, the lack of a detailed emission inventory, and the operational availability of a low-cost monitoring network.

Our first step was to apply the Ensemble Kalman Filter (EnKF) to assimilate the official air quality monitoring network. Evaluations of the system were performed by varying values of the covariance localization influence area. Moreover, various inheritance strategies were evaluated to optimize the assimilation window's estimated information into the forecast window. Although the model's performance could be improved with application of DA, there were still issues with the emission inventories, the low number of observations, and the model's difficulties in capturing essential transport dynamics within the valley.

Given the significant impact the Aburrá Valley emission inventory has on air quality modeling and perceived issues with the available inventory, we built a high-resolution emission inventory for the Aburrá Valley metropolitan area. We also assessed the ability of a low-cost network's available in the metropolitan area to track the dynamics of $PM_{2.5}$ correctly and use it as observations in the DA process. With recent developments in the production of low-cost sensors, it is possible to use these devices for DA. The DA system is composed by the EnKF, LOTOS-EUROS, the latest emission inventory, and the low-cost monitoring network. The high measurement density of this type of network is an advantage in the DA process, and it

can be used in places that cannot afford a standard monitoring network. Finally, the city's air quality was improved through the revised emission inventory.

Combined with a new emission inventory and a denser observation network, we have proposed two ensemble-based DA methods to deal with the high uncertainties in the model. The first is a variant of the EnKF using a covariance-based estimator called Ensemble Kalman Filter Knowledge-Aided (EnKF-KA). The method's novelty is that it allows for incorporating prior knowledge of the system directly in the assimilation process through a target covariance matrix. The second method, the Ensemble Time Local H_∞ Filter Knowledge-Aided (EnTLHF-KA) is a robust version of the EnKF-KA that incorporates an adaptive covariance inflation factor to reduce the impact of uncertainties. Both approaches were first analyzed using simple models to isolate the proposed technique's advantages and drawbacks and to compare the results of this new method with traditional algorithms. The formulation of both new methods is sufficiently general to be applicable in other contexts.

Finally, we implemented the proposed methods with the LE model and the low-cost monitoring network in the Aburrá valley. We used the target matrix to limit the influence of the observations, following the complex topography of the valley. This reduced the impact caused by a low resolution of the dynamics within the valley of the meteorological input. The results of the proposed methods were compared with the results of the Local Ensemble Transform Kalman Filter (LETKF) algorithm. Both new methods outperformed the LETKF and resulted in a more accurate spatial representation of the PM concentrations. Thus, by applying the DA method to the Aburrá Valley, the modeling and forecasting of air quality improved tremendously when compared with the observations.

Samenvatting

Om nadelige effecten van luchtverontreiniging te verminderen, zijn er veel onderzoeken gedaan naar methoden om te kunnen voorkomen dat de luchtverontreiniging een gevaarlijk niveau bereikt. Een Chemical Transport Model (CTM) kan de concentratie van gassen en deeltjes in de atmosfeer in specifieke gebieden simuleren. Deze modellen zijn helaas niet altijd betrouwbaar vanwege onnauwkeurige emissies en meteorologische invoergegevens. De variabiliteit in de luchtkwaliteitsmodellen verklaren en voorspellen blijft een uitdaging. Data assimilatie (DA) kan de onzekerheid in het model verminderen. Data assimilatie is een wiskundige techniek waarbij waargenomen waarden worden gecombineerd met een dynamisch model om de nauwkeurigheid van het model te verhogen.

Standaard methoden voor data assimilatie hebben beperkingen wanneer er geen volledige karakterisering van de onzekerheden beschikbaar is. Bij luchtkwaliteitsberekeningen is de nauwkeurigheid van emissies vaak laag, terwijl ook weersinvloeden vaak niet erg nauwkeurig bekend zijn. Dit probleem is nog groter in ontwikkelingslanden waar de beschikbare kennis over de luchtkwaliteit schaars en van relatief lage kwaliteit is. De belangrijkste bijdrage van dit proefschrift is de ontwikkeling van data assimilatie methoden die het mogelijk maken het gedrag van complexe luchtkwaliteitsmodellen te verbeteren, ook in de aanwezigheid van grote onzekerheden. De voorgestelde methoden zijn getest in realistische situaties met het LOTOS-EUROS CTM dat is toegepast op de Aburrá-vallei in Colombia om de voorspelling van de fijn stof concentratie te verbeteren. Ook het gebruik van een groot aantal goedkopere sensoren wordt onderzocht. De Aburrá-vallei vormt een goede testcase vanwege de problemen met de luchtkwaliteit, de moeilijkheidsgraad van het terrein, het ontbreken van gedetailleerde emissies en de operationele beschikbaarheid van een goedkoop meetnet.

De eerste stap was het toepassen van het Ensemble Kalman filter (EnKF) om data van het officiële meetnet voor luchtkwaliteit te assimileren. De aanpak is geëvalueerd door het gebruikte covariantie lokalisatie schema te variëren. Bovendien zijn verschillende strategieën geëvalueerd om informatie uit de assimilatieperiode te gebruiken bij het berekenen van de voorspellingen. Hoewel de prestaties van het model konden worden verbeterd met de toepassing van DA, waren er nog steeds problemen met de emissies, het lage aantal waarnemingen, en de problemen met het model bij het beschrijven van essentiële transportdynamiek in de vallei.

Gezien de aanzienlijke impact die de Aburrá Valley-emissies hebben op de luchtkwaliteitsmodellering, hebben we een emissie inventarisatie met hoge resolutie opgesteld voor het grootstedelijk gebied van Aburr á Valley. We hebben ook het gebruikt van het goedkope netwerk in het grootstedelijk gebied geëvalueerd met betrekking tot het beschrijven van de dynamiek van $PM_{2.5}$ en de waarnemingen ook in het DA proces gebruikt. We pasten hierbij het EnKF en LOTOS-EUROS toe met

de laatste emissie inventarisatie. Met de recente ontwikkelingen in de productie van goedkope sensoren, is het heel goed mogelijk om deze sensoren zinvol voor DA te gebruiken. De hoge dichtheid van dit type netwerk is een voordeel bij het data assimilatie proces en kan worden gebruikt op plaatsen waar een standaard monitoringnetwerk te duur zou zijn. Ten slotte werd de luchtkwaliteit van de stad verbeterd door de aangepaste emissie inventarisatie te gebruiken in het model.

In combinatie met een nieuwe emissieinventarisatie en een dichter observatienetwerk, hebben we twee ensemble-gebaseerde data assimilatie methoden ontwikkeld om de grote onzekerheid in de modelvoorspellingen aan te kunnen pakken. De eerste is een variant van het EnKF met behulp van een covariantie gebaseerde schatter, genaamd het EnKF-KA. De originaliteit van deze methode is dat het nu mogelijk wordt om voorkennis van het systeem rechtstreeks in het assimilatieproces op te nemen via een doelcovariantiematrix. De tweede methode (EnTLHF-KA) is een robuuste versie van het EnKF-KA die een adaptieve covariantie inflatiefactor bevat om de impact van niet gemodelleerde onzekerheden te verminderen. Beide benaderingen zijn geanalyseerd met behulp van eenvoudige modellen om de voor- en nadelen van de voorgestelde techniek te onderzoeken. De formulering van beide methoden is heel algemeen en ook in veel andere model problemen toepasbaar.

Tenslotte hebben we de voorgestelde nieuwe methoden geïmplementeerd in het LOTOS-EUROS model en toegepast door gebruik gemaakt van het goedkope meetnetwerk in de Aburrá-vallei. We gebruikte hierbij de doelmatrix om de invloed van de waarnemingen te beperken tot een aantal modelcomponenten, gebaseerd op de complexe topografie van de vallei. Dit hielp ons om de effecten die worden veroorzaakt door de lage resolutie van de meteorologische dynamiek in de vallei te verminderen. De voorgestelde methoden zijn vergeleken met het bekende LETKF algoritme. Beide nieuwe methoden presteerden beter dan het LETKF en produceerde nauwkeuriger ruimtelijke representaties van de PM-concentraties. Door de data assimilatiemethode toe te passen op de Aburrá Valley, zijn de modellering en de voorspelling van de luchtkwaliteit enorm verbeterd.

1

Introduction

Particulate matter (PM) is one of the most problematic pollutants in urban air. PM's effects on human health, associated with PM of $\leq 2.5\mu m$ in diameter, include asthma, lung cancer, and cardiovascular disease. Consequently, major urban centers commonly monitor $PM_{2.5}$ as part of their air quality management strategies.

Monitoring could be done using a static network of high-quality but expensive measurement devices. The use of low-cost air quality networks has been increasing in recent years to study urban pollution dynamics with more spatial detail. In addition to monitoring, Chemical Transport Models (CTM's) allow for permanent simulation and evaluation of pollutant behavior for all locations in a region of interest. Validated with observations should ensure their quality.

Part of this chapter is under review in: Data assimilation as a tool to improve Chemical Transport Models performance in developing countries (Chapter Book), Air Quality 2021, **IntechOpen**

1

1.1. Air quality in the Aburrá Valley

Air pollution is defined as the presence of solid, liquid, or gaseous components in the atmosphere that can cause risk and trouble for living beings or goods in general. Air pollution is one of the major environmental problems in modern human history (Green and Sánchez, 2012). Environmental pollution can be produced by natural or human actions. Natural sources include forest fires, volcanic emissions, dust, sand, vegetation (as pollen), and wetlands (as methane). The main human sources of air pollution are industry, power generation, transportation, deforestation, and cattle raising (Borrego *et al.*, 2015).

The current exponential growth in world population heightens the importance of public health issues related to air quality (Akimoto, 2003; Gurjar *et al.*, 2008). In developing countries, decision makers must cope with the environmental demands of expanding and overpopulated urban centers. Short term air quality forecasts and long term mitigation strategies for these centers are usually based on specialized assessments of particulate matter dynamics (Bell *et al.*, 2011; Sallis *et al.*, 2016). The Aburrá Valley houses the city of Medellín and neighboring municipalities. It is the second most populous urban agglomeration in Colombia, and the third densest in the world. The valley traces the course of the Medellín River along 60 km of a deep mountain canyon that ranges in width between 3 and 10 km, and with a height difference of up to 1800 m. Air quality conditions deteriorate severely within the valley twice a year around the time of the arrival of the Intertropical Convergence Zone (March-April, and with lower intensity in October-November), when the atmospheric inversion layer persists throughout the day below the rim of the canyon, thus trapping all of the urban atmospheric contaminants within the lower atmosphere (Jiménez, 2016). During these periods, the concentrations of particulate matter below 10 µm (PM_{10}) and 2.5 µm ($PM_{2.5}$) remain at levels considered hazardous for vulnerable populations and even for the general population.

Figure 1.1: Perspective of the air quality in the city of Medellín. (August 26, 2016, www.elmundo.com)

1.2. Using Chemical Transport Models for air quality modeling

Due to the large stress on human health induced by this air pollution, efforts have been made to monitor, reduce, and prevent episodes in which concentrations of pollutants reach hazardous levels. Before measures for reducing air pollution can be implemented, it is important to know the current concentration levels and how these evolve over time within the area of interest. This could be done using a Chemical Transport Model (CTM) to simulate concentrations of trace gasses and particulate matter (Thunis *et al.*, 2016; Lateb *et al.*, 2016). In the last 20 years, CTMs have seen huge growth and development; in consequence, a diversity of models exists, differing in their complexity, size of the region of study, and methods used for their development. CTMs can be broken down into four categories according to their dynamic behavior: i) Gaussian, ii) statistic, iii) Lagrangian and iv) Eulerian (Thunis *et al.*, 2016). Eulerian models are the most widely used and reported for monitoring and predicting the pollution behavior and define the air quality in bigger areas (Lateb *et al.*, 2016). These are frequently used in areas as large as countries or continents, and have been less used in areas such as cities. There have not been many applications of CTM's to study the air quality in Colombia yet. Most of the efforts have focused on the development of emission inventories and pollutant characterization (Toro *et al.*, 2005, 2006; Zarate *et al.*, 2007; Nedbor-Gross *et al.*, 2018; Pachón *et al.*, 2018). Applications of a CTM have however been reported too. An early study on atmospheric pollution in Colombia used the WRF-CHEM model (Weather Research and Forecasting with Chemistry) to simulate the concentrations of PM_{10} over the Bogotá metropolitan area (Kumar *et al.*, 2016). The Emissions Database for Global Atmospheric Research (EDGAR) global emission inventory was used as input. The simulations underestimated the PM_{10} concentrations by an order of magnitude compared to observations. The WRF-CHEM model has also been applied to study the behavior of O_3 over the medium-size, mountainous city of Manizales (Gonzalez *et al.*, 2018). By using high-resolution simulations (1 km x 1 km), the study compared the performance of the model when using either the EDGAR emission inventory or a high-resolution emission inventory previously developed (Gonzalez *et al.*, 2017). In Henao *et al.* (2020) the WRF-Chem model in a sub-kilometer configuration was used to reproduce the CO dynamics in the valley. The emission inventory was spatially dissagregated from the AMVA Official Emission Inventory (UPB and AMVA, 2017). Although the meteorological fields showed a high similarity with observations, the model underestimated the CO concentrations. The underestimation is attributed to mismatches in the official emission inventory and uncertainties generated by the simplifications of disaggregation methodologies.

This thesis uses simulations of the LOTOS-EUROS (LE) CTM for studying the atmospheric contaminant dynamics within the Aburrá valley . The model setup for the region of interest is described in Section 2.1.1. As a novelty for the region, this study not only uses a CTM, but also applies data assimilation to improve the forecast skills of the model. LE is equipped with several Ensemble-based data assimilation applications focused on the reanalysis and forecasting of gasses over

Europe (Manders *et al.*, 2017).

1.3. Data assimilation as a tool to improve model performance

Data assimilation (DA) is a mathematical process that provides integration between measured values (observations) and a dynamic model to improve the operation of the model. With DA, the output value provided by the model has a smaller error than the output value provided by the model without observations. DA has two key objectives: to improve the operation in predictions of model states; and estimate unknown parameters of the model (Berardi *et al.*, 2016). DA is used in different scientific fields such as oceanography, meteorology, air quality, greenhouse gas studies, and reservoir characterization (Van Loon *et al.*, 2000). DA allows integrating models and observations with different scales of size and temporal sampling (Lahoz and Schneider, 2014). When two sources of information are combined, DA assumes that both are subject to errors. These errors are intrinsically unknown and need to be specified in probabilistic terms. DA aims to reduce the model error in space or time with observations, but its mission is also to digest the observation based on the laws given by the model and to determine the dynamic evolution of the model state that represents better measurements (Bocquet *et al.*, 2015).

Large-scale model uncertainty is difficult to characterize, and even more difficult to reduce. Increasing the accuracy of initial conditions, such as accurate land cover representations or updated emissions inventories, or using observations and DA, may reduce uncertainty. DA an alternative that is dynamically driven to reduce the lack of knowledge about the behavior of air pollution. The addition of surface, satellite, in situ, and laser-based remote sensing data to a model will enhancethe simulation skill, and with that improve the thrust in proper scenario simulation and online decision-making. A further promise lies in the incorporation of the DA, not only for its contribution to the reduction of uncertainty, but also for opening the door to more accurate air quality forecasting in atmospheric pollution modeling (Quintero Montoya *et al.*, 2020). CTM based forecasting presents us with interesting and complex challenges associated with the uncertainty of weather forecasts, the lack of precise inventory of emissions, and the scarcity and sparsity of monitoring networks for air quality. Such challenges require creative solutions; these challenges are opportunities for knowledge advancement. Due to the scarcity of data and high uncertainty in the model inputs, a mathematical, analytical, and computational effort is needed to push the frontiers of knowledge in the field.

1.4. A Low-cost alternative for Particulate Matter monitoring

Public air quality monitoring networks often consist of fixed measuring stations equipped with expensive sensors and maintained under rigorous operational and calibration regimes in order to provide high quality data. The high costs associated with establishing and maintaining such stations means that not all cities in develop-

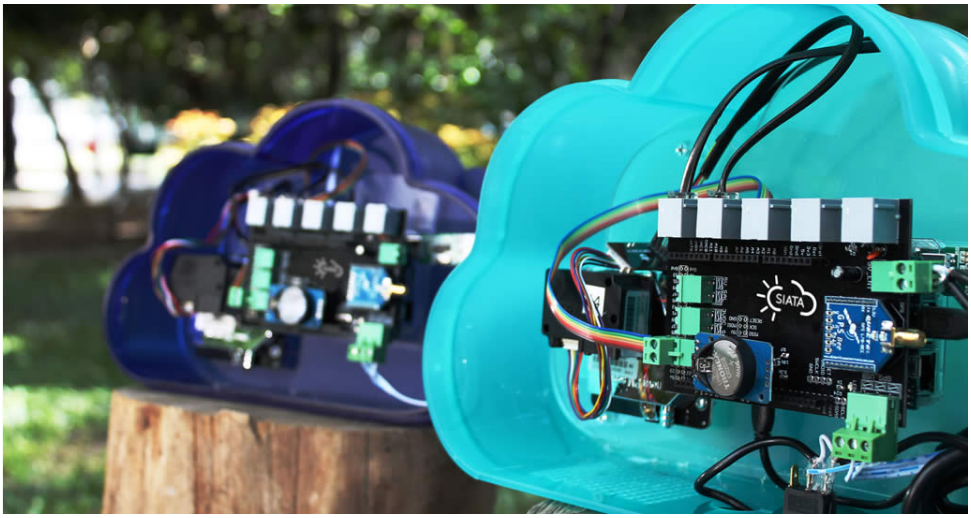


Figure 1.2: Ciudadanos cientificos (Scientific Citizen) measurement device used in the low-cost sensor network deployed on the Aburrá Valley, Colombia. Taken from https://www.metropol.gov.co/ambiental/siata/Paginas/ciudadanos-cientificos.aspx.

ing countries can afford monitoring networks of sufficient spatial coverage (Kumar and Gurjar, 2019). Even in large cities in developed countries, the official air quality monitoring networks do not always provide information at the spatial and temporal resolution required to assess the impact of pollution sources on health (Ahangar *et al.*, 2019), as the cost of the equipment makes the necessary density prohibitive. In the city of Medellín (Colombia) and its conurban municipalities for example, there are 21 main $PM_{2.5}$ monitoring stations, at an average density of 8.25 km^2 over the entire area of the 10 municipalities. This has motivated the expansion and improvement of low-cost systems and programs to measure PM (Kumar *et al.*, 2015). The limited number of studies that have evaluated newer generations of low-cost $PM_{2.5}$ sensors have shown that the most widely used low-cost sensors attain high accuracy when compared to standard monitoring stations (R^2 value ranging from 0.93 to 0.95) (Liu *et al.*, 2019). The data provided by these sensors can complement those generated by conventional systems, increasing the data resolution and allowing for studies of exposure at the human level (Ahangar *et al.*, 2019; Schneider *et al.*, 2017). By DA, the incorporation of air pollution data into CTM increases the ability to grasp local and regional patterns and fill spatial coverage gaps. Additionally, the combination of different sources of information and knowledge (data and model) increases the robustness and reliability of low-cost observations (Lahoz and Schneider, 2014; Castell *et al.*, 2017).

1.5. Aim and Research Questions

As described in the previous sections, the uncertainties in the inputs and the scarcity of observations impose a challenge on the assimilation of data in a CTM, especially

in regions with complex conditions and developing countries. Therefore, this thesis aims to develop a DA scheme for high-uncertainty systems toward the improvement of forecast capabilities, using alternative and more accessible sources of measurements. In this thesis, we use the LE model and the Aburrá Valley air quality case to evaluate the proposed methods and techniques. Several research questions should be addressed to achieve the main aim of this thesis. The proposed research question and the methodological approaches are:

RQ1: How to integrate all the possible information captured in the Ensemble-based DA process (state value, parameters value and dynamic, etc.) into the forecast simulation of systems with high uncertainty?
To answer this question, we implement the standard ensemble-based DA technique EnKF, with the LE model over the Aburrá valley using the configuration shown in Chapter 5.2. We evaluate the improvements in the particulate matter's representation after assimilation, and compare different information inheritance schemes between the assimilation window and the forecast window (Chapter 3). Additionally, we study different sources of error present in the assimilation system to formulate new assimilation algorithms.

RQ2: Can low-cost monitoring networks assimilated into a CTM be a more accessible alternative to standard air quality monitoring systems?
The metropolitan area of the Aburrá Valley has a low-cost operational network to monitor $PM_{2.5}$. We evaluate the network's capacity to correctly represent the dynamics of $PM_{2.5}$ by comparing it with the city's official monitoring network. After evaluation, we use the low-cost network measurements as observations in the standard DA system. The performance of the system using different configurations of the low-cost network and the official network is compared. The results and conclusions are shown in Chapter 4.

RQ3: How can a covariance localization scheme that uses direct knowledge of the system, for instance, a very complex topography, improve the performance of an Ensemble-based Data Assimilation method?
The results shown in Chapters 3 and 4 suggest a poor representation of the valley's pollutant transport dynamics, caused by the low resolution and high uncertainty in the meteorological fields. In Chapter 5, we introduce alternatives in estimating the state covariance that allow us to introduce prior knowledge of the system directly into the assimilation, e.g., spatial localization based on distance and orography. An efficient implementation of the EnKF based on shrinkage-estimator is formulated and evaluated using toy models to understand the new technique's advantages and possibilities.

RQ4: How does the performance of robust estimators compared to the EnKF under a scenario of high uncertainty sources like emissions, meteorology and observations?
This question is addressed in two chapters. In Chapter 6, we propose a robust

version of the algorithm proposed in Chapter 5. The idea of a robust method is approached given its capabilities to exploit the information from observations in high uncertainty scenarios. On the other hand, in Chapter 7, the implementation, evaluation, and comparison of the proposed techniques (robust and not robust) are carried out against standard assimilation techniques, using the $PM_{2.5}$ low-cost network.

1.6. Organization of this thesis

The thesis is organized as follows: Chapter 2 describes the LE model and the experimental setup used for all the PM simulation, the developed emission inventory, and introduces the preliminaries concepts of ensemble-based DA used among the thesis. In Chapter 3, the LE model coupled with the EnKF using covariance localization is implemented for PM concentration and emission estimations. Chapter 4 shows the evaluation and utilization of an alternative low-cost sensor network for DA proposes. In Chapter 5, we propose a new implementation of the EnKF using shrinkage-based covariance estimation that allows the incorporation of *prior* knowledge. Chapter 6 presents a robust and non-gaussian version of the EnKF implementation introduced in Chapter 5. In Chapter 7, we implement and evaluate the proposed robust and non-gaussian algorithm using the LE model over the Aburrá Valley. Finally, Chapter 8 summarizes the conclusions of this thesis, and the recommendations for further study.

References

J. Green and S. Sánchez, *Clean air Institute*, Tech. Rep. (Clean air Institute, Washington D.C., USA, 2012).

C. Borrego, M. Coutinho, a. M. Costa, J. Ginja, C. Ribeiro, a. Monteiro, I. Ribeiro, J. Valente, J. H. Amorim, H. Martins, D. Lopes, and a. I. Miranda, *Challenges for a New Air Quality Directive: The role of monitoring and modelling techniques,* Urban Climate **14**, 328 (2015).

H. Akimoto, *Global air quality and pollution,* Science **302**, 1716 (2003).

B. Gurjar, T. Butler, M. Lawrence, and J. Lelieveld, *Evaluation of emissions and air quality in megacities,* Atmospheric Environment **42**, 1593 (2008).

M. L. Bell, L. A. Cifuentes, D. L. Davis, E. Cushing, A. G. Telles, and N. Gouveia, *Environmental health indicators and a case study of air pollution in latin american cities,* Environmental Research **111**, 57 (2011).

J. F. Sallis, F. Bull, R. Burdett, L. D. Frank, P. Griffiths, B. Giles-Corti, and M. Stevenson, *Use of science to guide city planning policy and practice: how to achieve healthy and sustainable future cities,* The Lancet **388**, 2936 (2016).

J. F. Jiménez, *Altura de la Capa de Mezcla en un area urbana montanosa y tropical. Caso de estudio: Valle de Aburra*, Doctoral thesis, Universidad de Antioquia, Medellin (2016).

P. Thunis, A. Miranda, J. M. Baldasano, N. Blond, J. Douros, A. Graff, S. Janssen, K. Juda-Rezler, N. Karvosenoja, G. Maffeis, A. Martilli, M. Rasoloharimahefa, E. Real, P. Viaene, M. Volta, and L. White, *Overview of current regional and local scale air quality modelling practices: Assessment and planning tools in the EU,* Environmental Science & Policy **65**, 13 (2016).

M. Lateb, R. Meroney, M. Yataghene, H. Fellouah, F. Saleh, and M. Boufadel, *On the use of numerical modelling for near-field pollutant dispersion in urban environments: A review,* Environmental Pollution **208**, 271 (2016).

M. V. Toro, L. Cremades, and J. Ramírez-Echeverry, *Inventario de emisiones biogenicas en el valle de aburrá,* Revista ingeniería y gestión ambiental **1**, 31 (2005).

M. V. Toro, L. V. Cremades, and J. Calbó, *Relationship between VOC and NOx-emissions and chemical production of tropospheric ozone in the Aburrá Valley (Colombia),* Chemosphere **65**, 881 (2006).

E. Zarate, L. Carlos Belalcazar, A. Clappier, V. Manzi, and H. Van den Bergh, *Air quality modelling over Bogota, Colombia: Combined techniques to estimate and evaluate emission inventories,* Atmospheric Environment **41**, 6302 (2007).

R. Nedbor-Gross, B. H. Henderson, M. P. Pérez-Peña, and J. E. Pachón, *Air quality modeling in Bogotá Colombia using local emissions and natural mitigation factor adjustment for re-suspended particulate matter,* Atmospheric Pollution Research **9**, 95 (2018).

J. E. Pachón, B. Galvis, O. Lombana, L. G. Carmona, S. Fajardo, A. Rincón, S. Meneses, R. Chaparro, R. Nedbor-Gross, and B. Henderson, *Development and evaluation of a comprehensive atmospheric emission inventory for air quality modeling in the megacity of Bogotá,* Atmosphere **9**, 1 (2018).

A. Kumar, R. Jimenez, L. C. Belalcazar, and N. Y. Rojas, *Application of WRF-Chem Model to Simulate PM10 Concentration over Bogota,* Aerosol and Air Quality Research **16**, 1206 (2016).

C. M. Gonzalez, R. Y. Ynoue, A. Vara-Vela, N. Y. Rojas, and B. H. Aristizabal, *High-resolution air quality modeling in a medium-sized city in the tropical Andes: Assessment of local and global emissions in understanding ozone and PM10 dynamics,* Atmospheric Pollution Research , 1 (2018).

C. M. Gonzalez, C. D. Gomez, N. Y. Rojas, H. Acevedo, and B. H. Aristizabal, *Relative impact of on-road vehicular and point-source industrial emissions of air pollutants in a medium-sized Andean city,* Atmospheric Environment **152**, 279 (2017).

J. J. Henao, J. F. Mejía, A. M. Rendón, and J. F. Salazar, *Sub-kilometer dispersion simulation of a CO tracer for an inter-Andean urban valley,* Atmospheric Pollution Research **11**, 0 (2020).

UPB and AMVA, *Inventario de Emisiones Atmosféricas del Valle de Aburrá - actualización 2015*, Tech. Rep. (Universidad Pontificia Bolivariana - Grupo de Investigaciones Ambientales, Area Metropolitana del Valle de Aburra, Medellín, 2017).

A. M. M. Manders, P. J. H. Builtjes, L. Curier, H. A. C. Denier Van Der Gon, C. Hendriks, S. Jonkers, R. Kranenburg, J. J. P. Kuenen, A. J. Segers, R. M. A. Timmermans, A. J. H. Visschedijk, R. J. W. Kruit, W. Addo, J. Van Pul, F. J. Sauter, E. Van Der Swaluw, D. P. J. Swart, J. Douros, H. Eskes, E. Van Meijgaard, B. Van Ulft, P. Van Velthoven, S. Banzhaf, A. C. Mues, R. Stern, G. Fu, S. Lu, A. Heemink, N. Van Velzen, and M. Schaap, *Curriculum vitae of the LOTOS–EUROS (v2.0) chemistry transport model,* Geosci. Model Dev **10**, 4145 (2017).

M. Berardi, A. Andrisani, L. Lopez, and M. Vurro, *A new data assimilation technique based on ensemble Kalman filter and Brownian bridges: An application to Richards' equation,* Computer Physics Communications **208**, 43 (2016).

M. Van Loon, P. J. H. Builtjes, and a. J. Segers, *Data assimilation of ozone in the atmospheric transport chemistry model LOTOS,* Environmental Modelling and Software **15**, 603 (2000).

W. a. Lahoz and P. Schneider, *Data assimilation: making sense of Earth Observation,* Frontiers in Environmental Science **2**, 1 (2014).

M. Bocquet, H. Elbern, H. Eskes, M. Hirtl, R. Aabkar, G. R. Carmichael, J. Flemming, a. Inness, M. Pagowski, J. L. Perez Camacho, P. E. Saide, R. San Jose, M. Sofiev, J. Vira, a. Baklanov, C. Carnevale, G. Grell, and C. Seigneur, *Data assimilation in atmospheric chemistry models: Current status and future prospects for coupled chemistry meteorology models,* Atmospheric Chemistry and Physics **15**, 5325 (2015).

O. L. Quintero Montoya, E. D. Niño-Ruiz, and N. Pinel, *On the mathematical modelling and data assimilation for air pollution assessment in the Tropical Andes,* Environmental Science and Pollution Research **27**, 35993 (2020).

A. Kumar and B. R. Gurjar, *Low-Cost Sensors for Air Quality Monitoring in Developing Countries - A Critical View,* Asian Journal of Water, Environment and Pollution **16**, 65 (2019).

F. E. Ahangar, F. R. Freedman, and A. Venkatram, *Using low-cost air quality sensor networks to improve the spatial and temporal resolution of concentration maps,* International Journal of Environmental Research and Public Health **16** (2019), 10.3390/ijerph16071252.

P. Kumar, L. Morawska, C. Martani, G. Biskos, M. Neophytou, S. Di Sabatino, M. Bell, L. Norford, and R. Britter, *The rise of low-cost sensing for managing air pollution in cities,* Environment International **75**, 199 (2015).

H. Y. Liu, P. Schneider, R. Haugen, and M. Vogt, *Performance assessment of a low-cost PM 2.5 sensor for a near four-month period in Oslo, Norway,* Atmosphere **10** (2019), 10.3390/atmos10020041.

P. Schneider, N. Castell, M. Vogt, F. R. Dauge, W. A. Lahoz, and A. Bartonova, *Mapping urban air quality in near real-time using observations from low-cost sensors and model information,* Environment International **106**, 234 (2017).

N. Castell, F. R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, and A. Bartonova, *Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?* Environment International **99**, 293 (2017).

2

Preliminaries

2.1. Particulate Matter modelling

2.1.1. LOTOS-EUROS model

The chemical transport model that is used to simulate atmospheric concentrations of pollutants is the LOTOS-EUROS model (Manders *et al.*, 2017). The model computes concentrations of trace gasses and aerosols in three dimensions for the lower parts of the atmosphere: the boundary layer and (part of) the free troposphere. The simulated trace gasses include ozone, nitrogen and sulphuric oxides, and hydrocarbons; aerosols include primary matter, secondary inorganic aerosol, elemental and organic carbon, sea-salt, and dust. There is the possibility to calculate secondary organic aerosol with a 1-D VBS scheme (Manders *et al.*, 2017; Sauter *et al.*, 2012). The LOTOS-EUROS model has been used for air quality studies in different projects around the world (Manders *et al.*, 2017), demonstrating the adaptability of the model for different regions.

In the following sections, the dynamic time step of the LOTOS-EUROS model will be denoted by:

$$\mathbf{c}_k = \mathbf{M}_{LE}(\mathbf{c}_{k-1}, \mathbf{c}_0, \mathbf{e}_{k-1}) \tag{2.1}$$

In here, the state vector $\mathbf{c}_k$ contains the concentrations of all trace gases and aerosols in each cell of the three dimensional grid valid for time t_k, $\mathbf{e}_k$ is the nominal emission from the emission inventory. The model operator $\mathbf{M}_{LE}$ computes the state at time t_k from the concentrations at t_{k-1}, and using the model input which is yet not further specified; note that in following equations some arguments of $\mathbf{M}_{LE}$ might be omitted to simplify notations. The processes included in the model operator include three dimensional transport by wind, vertical diffusion due to turbulence, entrainment and detrainment by changing boundary layer heights, emissions from anthropogenic and biogenic sources, chemical reactions, aerosol physics, and dry and wet deposition. The gas-phase chemistry is a condensed version of CBM-IV proposed in (Manders-Groot *et al.*, 2016) and for secondary inorganic chemistry Isorropia II (Fountoukis and Nenes, 2007) is used. The default meteorology of the model is 3-hourly ECMWF short-term forecast, but the models has also been run with meteorological input from WRF and COSMO, and has been coupled semi-online to the regional climate model RACMO2 (Manders *et al.*, 2017).

2.1.2. Domain and experimental setup

Simulations were conducted with the LE model, adopting a nested domain configuration as depicted in Figure 2.1 and detailed in Table 2.1. Four nested domains were used to have a smooth transition on the dynamics from the regional scales (Caribbean and Northern part of South America) to the local conditions of the Aburrá Valley. The first Domain (D1) spans from the coast of Nicaragua in the West, to the Caribbean Dutch Islands and Venezuela in the East; model resolution was set to 0.27° (about 28 km). For this domain, meteorological data from ECMWF was used at a resolution of 0.14°; also the orography was obtained from this data set. The inner domain D2 is centered over the valley and includes the Northwest part of Colombia, encompassing most of the Colombian Andes; model resolution was set to

0.09°(about 9 km). For this and the following inner domains, meteorological data were obtained from ECMWF at 0.07°resolution, while for the elevation model was derived from the Global Multi Resolution Terrain Elevation Data set (GMTED2010) (Danielson and Gesch, 2011) at a resolution of 0.002°(approx. 220 m). The third inner domain D3 includes the department of Antioquia, at a model resolution of 0.03°(about 3 km). The innermost domain D4, the focus of the present study, includes primarily the region of the Aburrá Valley using model resolution of 0.01° (about 1 km).

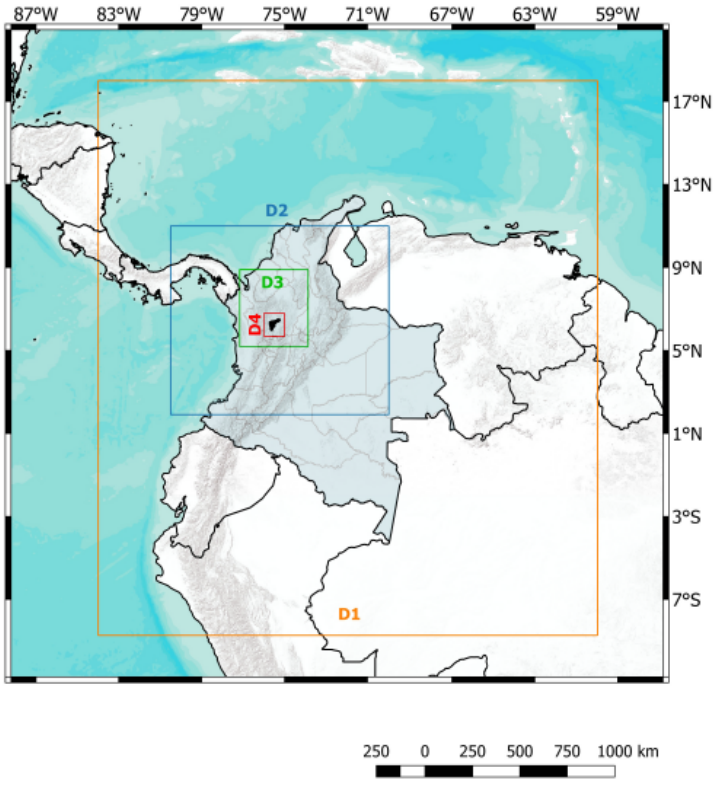


Figure 2.1: Four nested domains for Metropolitan Area of Aburra Valley assesment.

The data sets used in the model are summarized in Table 2.2.

2.1.3. Local Emissions Inventory

An anthropogenic urban emission inventory for 2016 specific to Medellín and the other nine municipalities of the Aburrá Valley was used for the simulations on the D4 domain. This inventory provides a complete set of emitted trace gases such as carbon monoxide (CO), nitrogen oxides (NO_x), sulphur oxides (SO_x), and volatile

Domain	Longitude	Latitude	Cell size
D1	84°W-60°W	8.5°S-18°N	0.27°× 0.27°
D2	80.5°W-70°W	2°N-11°N	0.09°× 0.09°
D3	77.2°W-73.9°W	5.2°N-8.9°N	0.03°× 0.03°
D4	76°W-75°W	5.7°N-6.8°N	0.01°× 0.01°

Table 2.1: Nested domain specifications

Period	31-March-2016 to 25-April-2016
Metereology	ECMWF; Temp.res: 3h; spat.res: $0.07^\circ \times 0.07^\circ$
Initial and boundary conditions	LOTOS-EUROS (D3). Temp.res: 1h. Spat.Res: $0.03^\circ \times 0.03^\circ$
Anthropogenic emissions	EDGAR v4.2. Spat.res:10 km × 10 km
Biogenic emissions	MEGAN Spat.res:10 km × 10 km
Fire emissions	MACC/CAMS GFAS Spat.res:10 km × 10 km
Landuse	GLC2000. Spat.res:1 km × 1 km
Orography	GMTED2010. Spat.res: 0.002°× 0.002°

Table 2.2: Data set used in the D4 domain.

organic compounds (VOC's), as well as particulate matter with diameter less than 2.5 μm ($PM_{2.5}$) or less than 10 μm (PM_{10}). The construction of the inventory followed a bottom-up methodology, combining activity data (traffic intensities, industrial production) with emission factors. Only traffic and industrial point sources were considered, without accounting for neither household nor commercial emissions (UPB and AMVA, 2017).

For integration into LOTOS-EUROS, the emission inventory was disaggregated over the Aburrá Valley (76°W-75°W and 5.7°N-6.8°N) at a resolution of 0.01°× 0.01°(approximately 1 km × 1 km), using a method based on road density as in Ossés de Eicker *et al.* (2008). The road network map was obtained from the OpenStreetMap database (Haklay and Weber, 2008), and simplified by removing segments classified as residential, as recommended in (Tuia *et al.*, 2007; Gómez *et al.*, 2018). The simplification of the road network can reduce errors in the spatial disaggregation since residential roads correspond to a high portion of the road network length but carry a low percentage of total vehicular traffic. For each grid cell j, the corresponding dissagregation factor DF was calculated as in (Ossés de Eicker *et al.*, 2008):

$$DF_j = \frac{\sum_{i=0}^{I} S_{i,j}}{\sum_{j=0}^{J}\sum_{i=0}^{I} S_{i,j}} \tag{2.2}$$

where $S_{i,j}$ is the length of road segment i in the grid cell j, I is the number of road segments in cell j, and J is the total number of grid cells. The point-source emissions were distributed on the grid using their known location, obtained from the official emissions inventory (UPB and AMVA, 2017). Figure 2.2 shows the resulting

emissions maps for $PM_{2.5}$ and PM_{10}.

(a) $PM_{2.5}$.

(b) PM_{10}.

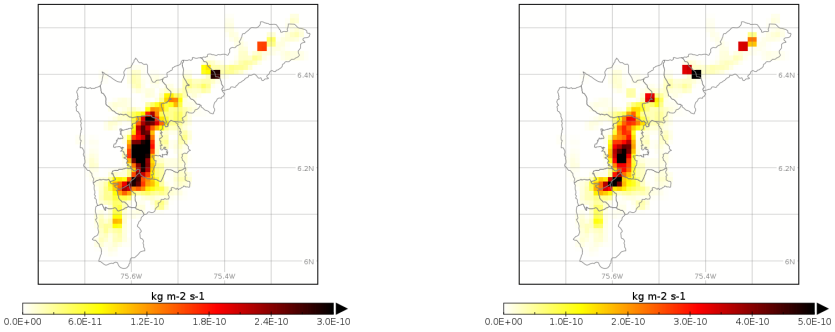


Figure 2.2: Local particulate matter emission inventories for the Aburrá Valley: (a) $PM_{2.5}$, and (b) PM_{10}. The values correspond with the estimated annual emissions.

2.2. Ensemble-based Data Assimilation

2.2.1. Stochastic and uncertainty representation

For implementation of the data assimilation algorithm a stochastic representation of the model uncertainty is needed. A major source of uncertainty are the emissions, which might in reality differ strongly from the inventory in both space and time. The emissions that are used in the model operator are therefore modelled as a stochastic process using a randomly varying deviation factor:

$$\hat{\mathbf{e}}_k \;=\; \mathbf{e}_k \cdot (\,1 \;+\; \delta\mathbf{e}_k\,) \tag{2.3}$$

The emission deviation is modelled as an autoregressive model of order one (AR-1), following the structure of a colored noise process (Jazwinski, 1970):

$$\delta\mathbf{e}_k \;=\; \alpha_k \cdot \delta\mathbf{e}_{k-1} + \sigma \cdot \sqrt{1-\alpha_k^2} \cdot \mathbf{w}_{k-1} \tag{2.4}$$

where $\mathbf{w}_k$ is a white noise process with zero mean and unity standard deviation:

$$\mathbf{w}_k \;\sim\; N(0,1) \tag{2.5}$$

Over an infinite number of samples, the stochastic factors are drawn out of a normal distribution with zero mean and standard deviation σ. The temporal correlation coefficient $\alpha_k \in [0,1]$ is used to describe the temporal variation, where the value should be set between two extremes: for $\alpha = 0$, the deviation is pure white noise with completely different values for every sample; for $\alpha = 1$ there is no temporal variation at all and the deviation factor is a single sample out of the normal

distribution. In this study the correlation parameter is described using a temporal length scale τ following (Barbu *et al.*, 2009):

$$\alpha_k = \exp\left(-|t_k - t_{k-1}|/\tau\right) \tag{2.6}$$

2

A stochastic model state is formed by augmenting the state vector (2.1) with the correction factor $\delta\mathbf{e}$:

$$\begin{bmatrix} \mathbf{c}_k \\ \delta\mathbf{e}_k \end{bmatrix} = \begin{bmatrix} \mathbf{M}_{LE}\left(\mathbf{c}_{k-1}, \mathbf{c}_0, \hat{\mathbf{e}}_{k-1}\right) \\ \alpha_k \cdot \delta\mathbf{e}_{k-1} \end{bmatrix} + \begin{bmatrix} 0 \\ \sigma \cdot \sqrt{1-\alpha^2} \end{bmatrix} \mathbf{w}_{k-1} \tag{2.7}$$

or simply:

$$\mathbf{x}_k = \mathbf{M}(\mathbf{x}_{k-1}) + \mathbf{G} \cdot \mathbf{w}_{k-1} \tag{2.8}$$

With the augmented vector (2.7), it is possible to apply a sequential data assimilation scheme to estimate both the state and the emission correction factor. The non-linear operator $\mathbf{M}$ propagates the augmented state vector $\mathbf{x}$ in time, while $\mathbf{G}$ distributes the stochastic forcing $\mathbf{w}_k$ over the elements of the state.

2.2.2. Ensemble Kalman Filter

The Ensemble-Based DA is a family of methods that uses an ensemble to model the statistics of the first guess (background). In each assimilation step, a forecast from the previous model simulation is used as a first guess, and using the available observation this forecast is then modified to bring it in better agreement with these observations. Due to its rather easy implementation (compared with other DA techniques), and its very general statistical formulation, it is one of the most widely used approaches for tackling assimilation problems (Fu *et al.*, 2017). The Ensemble Kalman filter (EnKF) is the most frequently used ensemble-based data assimilation method (Evensen, 2003). The EnKF is a Monte Carlo ensemble method, based on the representation of the probability density of the state estimates in an ensemble of N states:

$$\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)} \tag{2.9}$$

Each ensemble member is assumed to be a single sample out of a distribution of the true state (Fu, 2017).

The EnKF is initialized by generating a random ensemble $\mathbf{x}_0^{(i)}$ to represent the uncertainty in the initial condition $\mathbf{x}_0$. Then, the forecast step of the EnKF propagates each ensemble member in time using the state-space operator from Eq. (2.8) and a random forcing:

$$\mathbf{x}_k^{b(i)} = \mathbf{M}(\mathbf{x}_{k-1}^{a(i)}) + \mathbf{G} \cdot \mathbf{w}_{k-1}^{(i)} \tag{2.10}$$

where $\mathbf{x}_k^{b(i)}$ is the $i-th$ member of the forecast ensemble at time t_k. The forecast ensemble describes a stochastic distribution with mean and covariance respectively:

$$\overline{\mathbf{x}}_k^b = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_k^{b(i)} \tag{2.11}$$

$$\mathbf{P}_k^b = \frac{1}{N-1}\cdot\Delta\mathbf{X}_k\cdot\Delta\mathbf{X}_k^T \tag{2.12}$$

where the matrix $\Delta\mathbf{X}$ is formed by deviations of the ensemble members from the mean:

$$\Delta\mathbf{X} = \left[\mathbf{x}_k^{b(1)} - \overline{\mathbf{x}}_k^b, \ldots, \mathbf{x}_k^{b(N)} - \overline{\mathbf{x}}_k^b\right] \tag{2.13}$$

When observations are available, the EnKF uses them to update the forecast ensemble into an analysis ensemble, which has a smaller covariance since it incorporates observation information. The vector with observation values is described as a linear mapping from the state vector plus a random error:

$$\mathbf{y}_k = \mathbf{H}_k\cdot\mathbf{x}_k + \mathbf{v}_k, \quad \mathbf{v}_k \sim N(0,\mathbf{R}_k) \tag{2.14}$$

The observation operator $\mathbf{H}$ describes how the observations are sampled from the concentration fields in the state. The observation representation error $\mathbf{v}_k$ describes the difference between the observations and the sampling, which are present due to both instrumental errors but also due to sampling errors. In this applications the sampling errors are for example present since the state describes concentrations as averages in (large) grid boxes, while the observations concern point observations. The vectors $\mathbf{v}_k$ are assumed to be samples out of a random distribution with zero mean and covariance $\mathbf{R}_k$.

The analysis update of the ensemble members is proportional to the differences between the observations $\mathbf{y}_t$ and the observation simulation $\mathbf{H}_k\cdot\mathbf{x}_k^{b(i)}$ from the ensemble member following:

$$\mathbf{x}_k^{a(i)} = \mathbf{x}_k^{b(i)} + \mathbf{K}_k\cdot\left[\mathbf{y}_k - \mathbf{H}_k\cdot\mathbf{x}_k^{b(i)} + \mathbf{v}_k^{(i)}\right] \tag{2.15}$$

The difference between observations and simulations is distributed over the state elements using a matrix called the Kalman gain:

$$\mathbf{K}_k = \mathbf{P}_k^b\cdot\mathbf{H}_k^T\cdot[\mathbf{H}_k\cdot\mathbf{P}_k^b\cdot\mathbf{H}_k^T + \mathbf{R}_k]^{-1} \tag{2.16}$$

The Kalman gain is defined such that the sample covariance of the analysis ensemble is minimal with respect to l_2 matrix norm (Asch *et al.*, 2016). Note that the sample covariance $\mathbf{P}_k^b$ cannot be computed in this application given its large size ($\sim \mathcal{O}(10^6)\times\mathcal{O}(10^6)$). However, for the actual implementation it is sufficient to store only the factorization $\Delta\mathbf{X}$ from Eq. (2.13).

2.2.3. Covariance Localization

Due to the approximation of the state space covariance by a finite number of ensemble members, it is unavoidable that spurious correlations between elements of the state will appear. These spurious correlations can be removed by a procedure called localization (Ott *et al.*, 2004). The localization method used in this work is the *covariance localization* (Houtekamer and Mitchell, 2001). The covariance localization or Schur localization, focuses on the forecast error covariance matrix, cutting off correlations in the error covariances after a specified distance (Houtekamer and Mitchell, 2001; Petrie, 2008). The localization is implemented with a point wise multiplication called a Schur product and denoted by $\circ$:

$$[\mathbf{f} \circ \mathbf{P}^b]_{i,j} = [\mathbf{P}^b]_{i,j} \cdot [\mathbf{f}]_{i,j} \tag{2.17}$$

The Schur product theorem ensures that if $\mathbf{f}$ and $\mathbf{P}^b$ are positive semi-definite, then the Schur product, $\mathbf{f} \circ \mathbf{P}^b$, is positive semi-definite too. A cutoff function to fill $\mathbf{f}$ would be defined by $r \in \mathbb{R}^+ \rightarrow G(r/\rho)$, where r is the Euclidean distance between two state members and ρ is a length scaling called the localization radius (Sakov and Bertino, 2011). The localization radius is defined such that beyond this the correlation reduces from 1 and at a distance of more than $3.5 \cdot \rho$ the correlation reduces to zero (Petrie, 2008). The cutoff function utilized in this work has the following form:

$$f_{i,j} = \exp(-0.5 \cdot (r_{i,j}/\rho)^2) \tag{2.18}$$

This regularized covariance matrix $\mathbf{f} \circ \mathbf{P}^b$ is used in the EnKF analysis as well as in the generation of the posterior ensemble of perturbations, as a replacement for $\mathbf{P^b}$:

$$\mathbf{K} = (\mathbf{f} \circ \mathbf{P}^b) \cdot \mathbf{H}^T \cdot [\mathbf{H} \cdot (\mathbf{f} \circ \mathbf{P}^b) \cdot \mathbf{H}^T + \mathbf{R}]^{-1}$$

References

A. M. M. Manders, P. J. H. Builtjes, L. Curier, H. A. C. Denier Van Der Gon, C. Hendriks, S. Jonkers, R. Kranenburg, J. J. P. Kuenen, A. J. Segers, R. M. A. Timmermans, A. J. H. Visschedijk, R. J. W. Kruit, W. Addo, J. Van Pul, F. J. Sauter, E. Van Der Swaluw, D. P. J. Swart, J. Douros, H. Eskes, E. Van Meijgaard, B. Van Ulft, P. Van Velthoven, S. Banzhaf, A. C. Mues, R. Stern, G. Fu, S. Lu, A. Heemink, N. Van Velzen, and M. Schaap, *Curriculum vitae of the LOTOS–EUROS (v2.0) chemistry transport model,* Geosci. Model Dev **10**, 4145 (2017).

F. Sauter, E. V. der Swaluw, A. Manders-groot, R. W. Kruit, A. Segers, and H. Eskes, *TNO report TNO-060-UT-2012-01451*, Tech. Rep. (TNO, Utrecht, Netherlands, 2012).

A. M. M. Manders-Groot, A. J. Segers, S. Jonkers, M. Schaap, R. Timmermans, C. Hendriks, F. Sauter, R. W. Kruit, E. V. D. Swaluw, H. Eskes, and S. Banzhaf, *TNO report TNO2016 R10898*, Tech. Rep. (TNO, Utrecht, The Netherlands, 2016).

C. Fountoukis and A. Nenes, *Isorropiaii: A computationally efficient thermodynamic equilibrium model for k+-ca2+-mg2+-nh4 +-na+-so4 2–no3 –cl–h2o aerosols,* Atmospheric Chemistry and Physics **7**, 4639 (2007).

J. Danielson and D. Gesch, *Global Multi-resolution Terrain Elevation Data 2010(GMTED2010),* U.S. Geological Survey Open-File Report 2011–1073 **2010**, 26 (2011).

UPB and AMVA, *Inventario de Emisiones Atmosféricas del Valle de Aburrá - actualización 2015*, Tech. Rep. (Universidad Pontificia Bolivariana - Grupo de Investigaciones Ambientales, Area Metropolitana del Valle de Aburra, Medellín, 2017).

M. Ossés de Eicker, R. Zah, R. Triviño, and H. Hurni, *Spatial accuracy of a simplified disaggregation method for traffic emissions applied in seven mid-sized Chilean cities,* Atmospheric Environment **42**, 1491 (2008).

M. Haklay and P. Weber, *Openstreetmap: User-generated street maps,* IEEE Pervasive Computing **7**, 12 (2008).

D. Tuia, M. Ossés de Eicker, R. Zah, M. Osses, E. Zarate, and A. Clappier, *Evaluation of a simplified top-down model for the spatial assessment of hot traffic emissions in mid-sized cities,* Atmospheric Environment **41**, 3658 (2007).

C. D. Gómez, C. M. González, M. Osses, and B. H. Aristizábal, *Spatial and temporal disaggregation of the on-road vehicle emission inventory in a medium-sized Andean city. Comparison of GIS-based top-down methodologies,* Atmospheric Environment **179**, 142 (2018).

A. Jazwinski, *Stochastic processes and filtering theory*, Mathematics in science and engineering No. 64 (Acad. Press, New York, NY [u.a.], 1970).

A. L. Barbu, A. J. Segers, M. Schaap, A. W. Heemink, and P. J. H. Builtjes, *A multi-component data assimilation experiment directed to sulphur dioxide and sulphate over Europe,* Atmospheric Environment **43**, 1622 (2009).

G. Fu, F. Prata, H. Xiang Lin, A. Heemink, A. Segers, and S. Lu, *Data assimilation for volcanic ash plumes using a satellite observational operator: A case study on the 2010 Eyjafjallajökull volcanic eruption,* Atmospheric Chemistry and Physics **17**, 1187 (2017).

G. Evensen, *The Ensemble Kalman Filter: Theoretical formulation and practical implementation,* Ocean Dynamics **53**, 343 (2003).

G. Fu, *Improving volcanic ash forecasts with ensemble-based data assimilation*, Ph.D. thesis, TU Delft (2017).

M. Asch, M. Bocquet, and M. Nodet, *Data Assimilation*, 1st ed. (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016).

E. Ott, B. R. Hunt, I. Szunyogh, A. V. Zimin, E. Kostelich, M. Corazza, E. Kalnay, D. Patil, and J. A. Yorke, *A local ensemble Kalman filter for atmospheric data assimilation,* Tellus **56**, 415 (2004).

P. Houtekamer and H. Mitchell, *A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation,* American Meteorological Society **129**, 123 (2001).

R. E. Petrie, *Localization in the ensemble Kalman Filter*, Master, University of Reading (2008).

P. Sakov and L. Bertino, *Relation between two common localisation methods for the EnKF,* Computational Geosciences **15**, 225 (2011).

3

Forecasting PM_{10} and $PM_{2.5}$ via EnKF based Data Assimilation

In this chapter a data assimilation system for the LOTOS-EUROS chemical transport model has been implemented to improve the simulation and forecast of PM_{10} and $PM_{2.5}$ in a densely populated urban valley of the tropical Andes. The Aburrá Valley in Colombia was used as a case study, given data availability and current environmental issues related to population expansion. The data assimilation system is an Ensemble Kalman filter with covariance localization based on specification of uncertainties in the emissions. Observations assimilated were obtained from a surface network for the period March-April of 2016, a period of one of the worst air quality crisis in recent history of the region. In a first series of experiments, the spatial length scale of the covariance localization and the temporal length scale of the stochastic model for the emission uncertainty were calibrated to optimize the assimilation system. The calibrated system was then used in a series of assimilation experiments, where simulation of particulate matter concentrations was strongly improved during the assimilation period, which also improved the ability to accurately forecast PM_{10} and $PM_{2.5}$ concentrations over a period of several days.

Part of this chapter has been published in (Lopez-Restrepo *et al.*, 2020): Forecasting PM_{10} and $PM_{2.5}$ in the Aburrá Valley (Medellín, Colombia) via EnKF based Data Assimilation **Atmospheric Environment, 232, 117507**

3.1. Introduction

This study uses simulations of the LOTOS-EUROS (LE) chemistry transport model (CTM) for studying the atmospheric contaminant dynamics within the Aburrá valley, spanning a set of 10 municipalities including the city of Medellín.

LE is equipped with several Ensemble-based data assimilation applications focused on the reanalysis and forecasting of gasses over Europe (Manders *et al.*, 2017). In Barbu *et al.* (2009) the EnKF is used with the covariance localization technique for assimilating ground based observations to represent the dynamics of SO_4 and SO_2 over the European continent. In their work, two different sources of uncertainty were studied, the reaction rate in the production of SO_4 from SO_2 and the emissions of SO_2 and SO_4. The uncertainty was modeled as a colored noise process and estimated following the method presented in (Heemink and Segers, 2002) (explained in detail in the Section 2.2.1). Barbu *et al.* (2009) concluded that by improving the description of the emission and reaction rate uncertainties, the performance of the data assimilation was enhanced.

In (van Velzen and Segers, 2010), the performance of the data assimilation software package COSTA was evaluated with a LOTOS-EUROS application for a number of ensemble-based methods such as EnKF (without localization), ensemble square root filter (EnSRF) (Whitaker and Hamill, 2002), Complementary Orthogonal subspace Filter For Efficient Ensembles (COFFEE) (Heemink *et al.*, 2001) and the RRSQRT Kalman filter (Verlaan and Heemink, 1997). The model uncertainty was prescribed for emissions originating from different countries in the European domain.

A scheme of data assimilation using LOTOS-EUROS and a network of ground based sensors over Europe of O_3 is presented in (Curier *et al.*, 2012). A colored noise process was used to model the uncertainty in the NO_x and VOC emissions, the O_3 deposition rate and the exchange of O_3 between the troposphere and the stratosphere. The model performance and the quality of the forecasts generated improved significantly with data assimilation using the estimated emission factors.

In the Aburrá valley, observations of particulate material are available from *Sistema de Alerta Temprana del Valle de Aburrá* (SIATA), a ground-based sensor network with stations along the valley. A preliminary exercise is performed on assimilation of these observations within the simulations, and evaluating the forecast potential of this system. From the scientific point of view, this implementation represents a challenge due to the different sources of uncertainty present. The physical conditions of the region of interest such as the topography and the size of the valley demand an extra effort to conduct a regional high-resolution model simulation. Currently model inputs (emission inventory and meteorology) are not freely available with the desired resolution and quality, increasing the uncertainty present in the experiments. The results of the experiment suggest that the simulation and assimilation system is able to describe the dynamics of atmospheric pollutants in Medellín rather well, considering that the above mentioned issues remain challenging.

This chapter is organized as follows. In Section 3.2 we present the materials and methods, including the theoretical framework for the ensemble-based data assimilation technique and the covariance localization that was used for improving

the model results. Section 3.3 presents the experimental set up, and the data assimilation calibration with different radii for covariance localization, and several factors for the stochastic processes in the LE model. It also presents the observation error covariance matrix estimation from ground-based sensor network data. Section 3.4 presents the main results of the paper in terms of investigating the ability to forecast particulate matter concentrations over a few days. Section 3.5 offers some concluding remarks and outlines the needed future work.

3.2. Material and methods

3.2.1. The LOTOS-EUROS Model setup for Aburrá Valley

The LOTOS-EUROS simulations were performed using the modeling setup described in 2.1.2. For each of the domains, anthropogenic emissions were obtained from the global EDGAR inventory (Petrescu *et al.*, 2012). Although previous works have shown that there is a considerable gap in the information in EDGAR for the Colombian territory, this database is the only one available with all the necessary species to run the model in the selected domains (Gonzalez *et al.*, 2017; Pachón *et al.*, 2018; Nedbor-Gross *et al.*, 2018). The resolution of the EDGAR database is 10×10 km, which is approximately 10 times coarser than the resolution of the innermost domain. The low resolution of the emission data compared to the high resolution of the model simulation can produce an unrealistic spatial distribution of emissions and concentrations. This emphasizes the importance of considering emissions as a major source of uncertainty for the DA system. The behavior of our data assimilation scheme is studied using EnKF with 15 ensemble members ($N = 15$ in eqns. 2.9-2.13 in Section 2.2.2) for both periods of assimilation. Previous experiments in related works and using LOTOS-EUROS model showed that using an greater ensemble members the performance of the algorithm did not increase significantly to justify the additional computational cost and 12-15 ensemble members are sufficient to describe the local covariance and to produce assimilation with stable results (Barbu *et al.*, 2009; Curier *et al.*, 2012).

3.2.2. Ground based data for assimilation

The *Sistema de Alerta Temprana del Valle de Aburrá* (SIATA) network of sensors provides high quality measurements for different air pollutants in the atmosphere over the Aburrá Valley region, monitoring species such as O_3, SO_2, PM_{10}, $PM_{2.5}$ and PM_1. The network is distributed in the five most populated Aburrá Valley's municipality, with the majority of the measuring stations located in the city of Medellín. The distribution of the observation sites is shown in Figure 3.1.

In this work, only PM_{10} and $PM_{2.5}$ measurements were used for the assimilation experiments, obtained from 8 PM_{10} stations, 3 $PM_{2.5}$ stations, and 1 combined station on a total of 12 observation sites. The observation time series have an hourly temporal resolution, with full coverage for most days. Measurements for one station for each species (represented with a star in Figure 3.1) were used for validation, taking two stations with a considerable distance between them to obtain a acceptable spatial representation, namely *Universidad San Buenaventura* (located

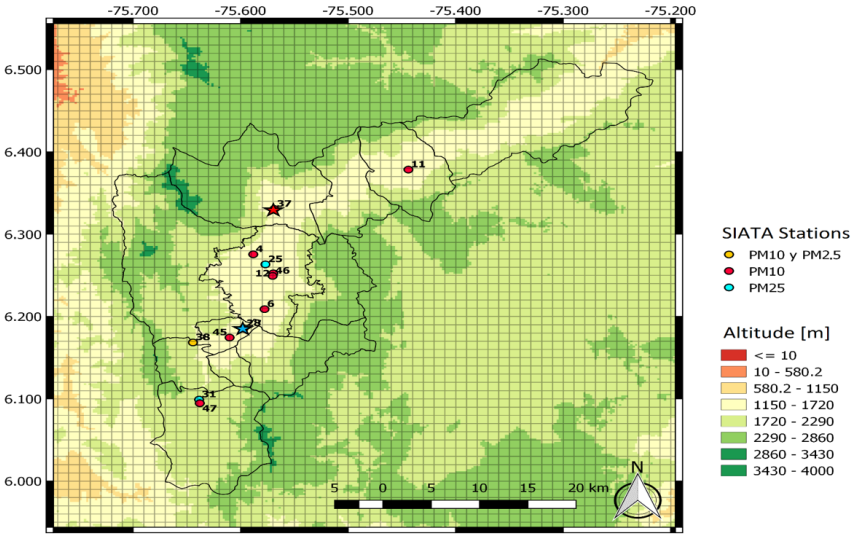


Figure 3.1: SIATA sensor network for PM_{10} and $PM_{2.5}$. The stars represent observation points for validation and the circles represent observations points for assimilation.

within the city of Medellín, near the geographic center of the valley; station 37 in Figure 3.1) for PM_{10}, and *Casa de la Justicia Itagui* (Southwest of the valley, in a mostly industrial zone; station 28 in Figure 3.1) for $PM_{2.5}$. During air quality crises, these stations tend to reach some of the highest values measured during the year. The metrics from section 3.2.3 are calculated only over these two stations.

3.2.3. Performance metrics

In this work, the performance of the LOTOS-EUROS simulations and the assimilation scheme were calculated by comparison with a subset of the ground observations not used in the assimilation. As described in Section 3.2.2, the collection of observations available in this study is rather small, and therefore only two time series were used to quantify the performance. Three metrics were computed to compare the simulations (assimilations) with the validation data; in addition, diurnal cycles were also compared.

The mean fractional bias (MFB) normalizes the bias for each model-observation pair using division by the average of the model and observation before taking the sample mean:

$$\text{MFB} = \frac{2}{M}\sum_{i=1}^{M}\frac{(H(c))_i - y_i}{(H(c))_i + y_i} \tag{3.1}$$

with M the number of elements in the set. In this application, M equals the

number of observations from all valid monitoring station data for the comparison time period of interest. The simulation $H(c)_i$ of an observation y_i is taken either from a model output, or from the ensemble mean in case of an assimilation run. The MFB ranges from -2 to $+2$, and has the advantage preventing the bias from being dominated by few high value observations/simulation pairs in case of strong variations, for example due to a strong diurnal cycle (Boylan and Russell, 2006).

The *root mean square error* (RMSE) represents the sample standard deviation of the differences between predicted values and observed values (equation 3.2). The RMSE penalizes a high variance as it gives errors with larger absolute values more weight than errors with smaller absolute values (Chai and Draxler, 2014):

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left((Hc)_i - y_i\right)^2} \tag{3.2}$$

The last metric is the *correlation coefficient* (Corr), which shows how the values from one data set (simulations) relate to value of a second data set (observations). A high value (approaching +1.0) is a strong direct relationship, values near 0.5 are considered moderate and values below 0.3 are considered to show weak relationships. A low negative value (approaching -1.0) is a strong inverse relationship, and values near 0.0 indicate little, if any, relationship. The correlation coefficient is calculated following (Yu *et al.*, 2006):

$$\text{Corr} = \frac{\sum_{i=1}^{M}\left((H(c)_i - \overline{(H(c)_i}\right)(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{M}\left((H(c)_i - \overline{(H(c)_i}\right)^2}\sqrt{\sum_{i=1}^{M}(y_i - \overline{y})^2}} \tag{3.3}$$

where the overline denotes a sample mean over the M elements of the validation set.

3.2.4. Standard model run results

In this section we described the simulated PM_{10} and $PM_{2.5}$ concentration for the two-week period between 31-March-2016 and 13-April-2016 using the model parameterization shown in Table 2.2 for domain D4. The evaluation against the validation stations are presented in figures 3.2 and 3.3. All the figures in this work are presented using the local time zone UTC-5. The statistical errors are shown in Table 3.1

Species	MFB	RMSE	CF
PM_{10}	-1.5	49.6369	0.3287
$PM_{2.5}$	-1.6	46.2302	0.3318

Table 3.1: Statistical error evaluation for PM_{10} and $PM_{2.5}$ via MFB, RMSE and CF for model standard Free-Run

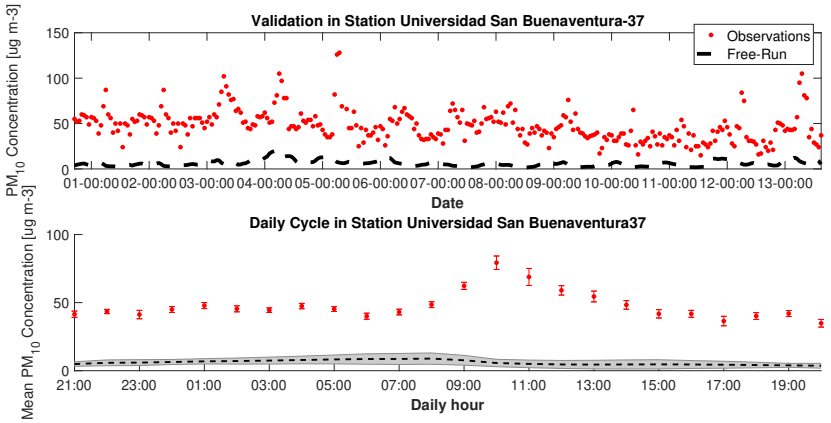


Figure 3.2: Time series and diurnal cycle of PM_{10} in validation site 37 for a standard model Free-Run. The time axis corresponds with the local time zone UTC-5.

It is evident that the model presents a considerable underestimation of the particulate matter concentration (around of 20 fold). These results are similar to the previous works of CTM implementation in Colombian cities (Kumar *et al.*, 2016). The causes of this gap can be attributed to two important factors: emissions and meteorology. As mentioned before, the EDGAR inventory is inaccurate over the Colombian territory (Gonzalez *et al.*, 2017) and the resolution is too coarse for the high-resolution model implementation. For these reasons, the emission of precursors and the particulate matter are considered as uncertainty parameters to be estimated in the DA system. The version of EDGAR used in this work (v4.2) only includes total particulate matter emissions, which in the model are distributed over the fine and coarse aerosol tracers. Therefore, only a single emission deviation field was used that was applied to all particulate matter emissions. The capability of the LE to use the last EDGAR version (v4.3) that differentiates between $PM_{2.5}$ and PM_{10} emissions is an upcoming feature of the model. NH_3 and SO_x emissions were estimated as precursors of secondary particulate matter. The mechanics of particle transport and the behavior of the boundary layer in the Aburrá Valley and its implications for concentration levels are not yet clear, nor is there a reliable high resolution meteorology for the region of interest: For this reason, we do not include meteorology as a source of uncertainty to estimate in the DA system.

3.3. Calibration of the data assimilation system

This section presents the results obtained from the data assimilation experiments with the LE model during a two-week episode. Simulations were conducted with the LE model in a nested domain configuration as described in Section 2.1.2. Default initial and boundary conditions were used, and data for assimilation was obtained and processed as described in Section 3.2.2. The goal of the experiments was to

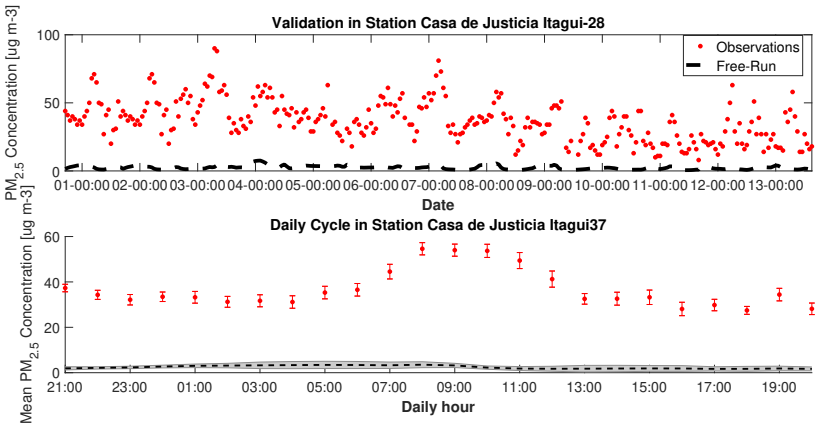


Figure 3.3: Time series and diurnal cycle of $PM_{2.5}$ in validation site 28 for a standard model Free-Run. The time axis corresponds to the local time (UTC-5).

obtain insight into the sensitivity of the assimilation system for the configuration parameters such as the temporal correlation of the emission uncertainty (Section 3.3.2), the localization length scale (Section 3.3.1), and the observation representation errors (Section 3.3.3). Also the impact of other parameters such as the standard deviation of the parameter uncertainty was evaluated, but for the chosen configuration their impact was minor. To see the impact of each configuration parameter, these are calibrated and analyzed independently. Based on the results, the best values for the assimilation parameters were selected for use in subsequent assimilation experiments. A series of emission deviation factors were obtained during the two-week episode using the calibrated assimilation system and used as nominal emissions for the next two-week test period. The forecast skill of the calibrated assimilation system was evaluated throughout the episode as described in Section 3.4. The summarized experimental setup is presented in the Figure 3.4.

3.3.1. Calibration of covariance localization radius

The covariance localization as described in Section 2.2.3 requires the definition of a localization radius ρ. In summary, the larger the radius chosen, the more observations are used to analyze a single element of the state. In this application, the state consists of concentrations and emission deviation factors, and the localization radius therefore has an impact on both. The influence of this parameter was evaluated by running the assimilation system with different values for ρ: 5, 10, 20, and 30 km. The temporal correlation length was fixed in $\tau = 3$ days for all the experiments. Figure 3.5 shows maps of the average value over the 2 week assimilation window emission deviation factor δe.

Figures 3.6 and 3.7 show time series of the average diurnal cycle of particulate matter concentrations in the two validation sites for the assimilation period.

For small localization radii, the concentrations in the validation sites were less

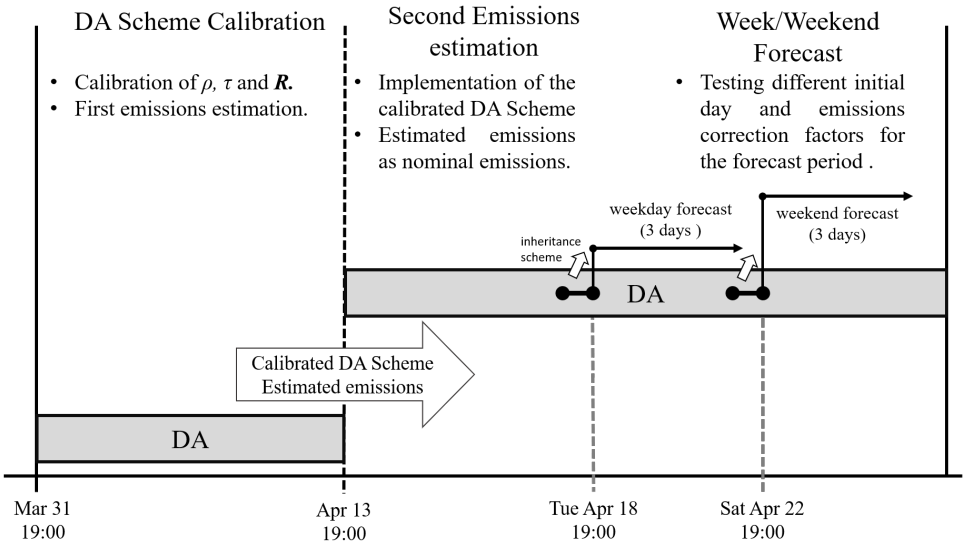


Figure 3.4: Graphic representation of the experimental setup.

ρ	MFB	RMSE	CF
5 km	-1.0718	49.6369	0.3287
10 km	-0.9220	46.2302	0.3318
15 km	-0.8564	44.8106	0.3273
30 km	-0.7815	43.4758	0.3082

Table 3.2: Statistical error evaluation for PM_{10} via MFB, RMSE and CF for ρ variation.

influenced by the analysis since the number of analyzed observations was limited for these locations. If the localization radius increases, the simulations become more in line with the observations, although even for a ρ of 30 km the simulations are lower than what is observed. In both stations, the assimilated model progressively approaches the observations towards the end of the assimilation window. The day cycles for both species show that the temporal dynamics are not significantly affected by the different values of ρ and the change is mostly reflected in the magnitude. For the available sensor network, the value of $\rho = 30$ km presented the best overall performance for both species. As shown in tables 3.2-3.3, the improvement of the error statistics related to the absolute error (MFB and RMSE) was more significant than the change in correlation factor (CF). The lack of accurate emissions inventories seems to have had a similar impact on simulations in all sites of the network, and therefore the best performance was obtained by changing emissions in the same way over the entire domain. It is expected that when using a sensor network with a higher spatial density, smaller values for ρ will become

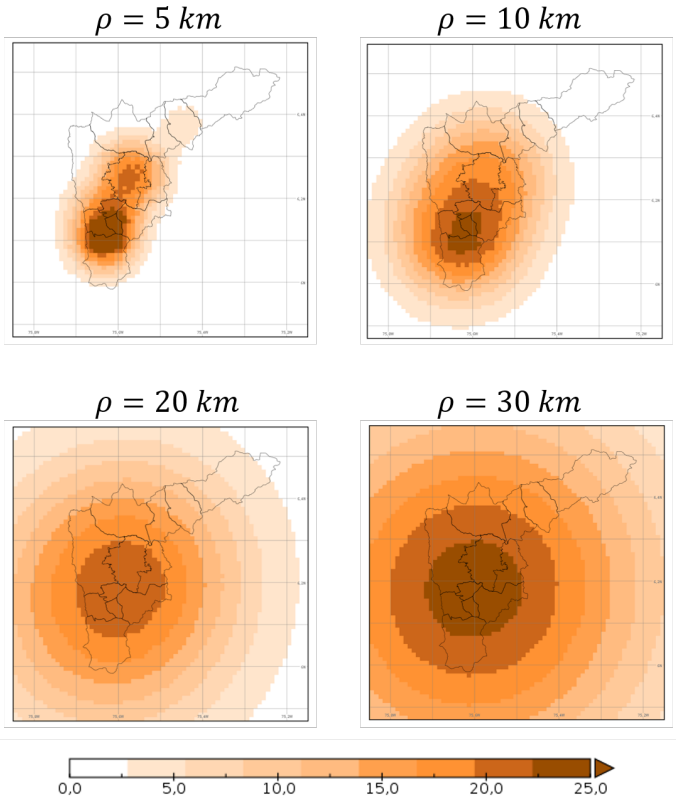


Figure 3.5: Maps of mean emission deviation factors for particle matter emissions during assimilation experiments with different localization radii.

beneficial.

3.3.2. Calibration of temporal correlation length

With a similar experiment as for the covariance localization, the temporal correlation parameter τ of the emission uncertainty described in section 2.2.1 was calibrated. The uncertainty on the emissions was modeled via equation (2.4). To evaluate the impact of the temporal correlation parameter, the assimilation experiment was performed with τ set to either 1, 2, 3, or 5 days. The localization radius was fixed in $\rho = 30$ km for all the experiments.

Figures 3.8 and 3.9 show the time series and average diurnal cycles of PM_{10} or $PM_{2.5}$ in the two validation sites, obtained from the observations, a standard model run, and analyzed ensemble means from assimilation experiments with different τ. Compared with the results shown in Figures 3.6 and 3.7 for variation of the localization radius, the impact of changes in the temporal correlation length are rather small. The assimilation results hardly differ from each other when τ changes,

ρ	MFB	RMSE	CF
5 km	-0.7264	25.7513	0.4118
10 km	-0.5232	23.1410	0.4204
15 km	-0.4456	22.2336	0.4171
30 km	-0.3731	21.8653	0.4019

Table 3.3: Statistical error evaluation for $PM_{2.5}$ via MFB, RMSE and CF for ρ variation.

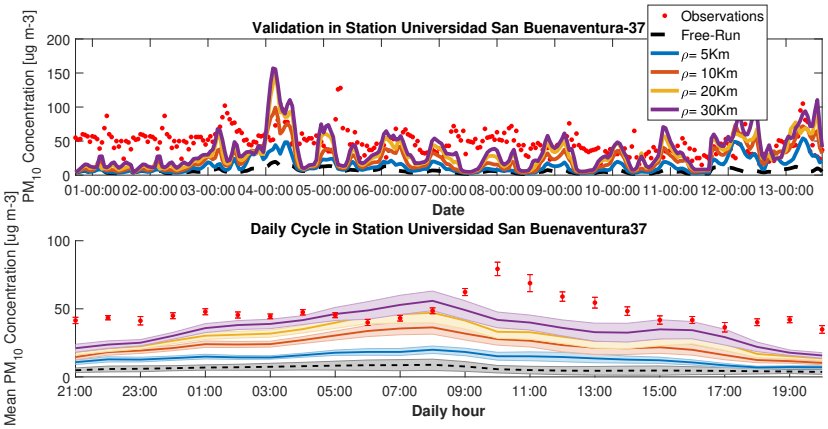


Figure 3.6: Time series and diurnal cycle of PM_{10} in validation site 37. Dots denote observations, dashed black lines are simulations by the standard model, and solid lines are analyzed ensemble means from the assimilation for different localization radii. The diurnal cycles were obtained from 13 samples for each hour. The shadows and the bars represent the standard deviation of the 13 samples. The time axis corresponds with the local time zone UTC-5.

indicating that in the current application this parameter is of minor importance.

Tables 3.4 and 3.5 show the values of the metrics defined in section 3.2.3 for the assimilation experiments with different τ. The MFB, RMSE and CF for both localization radius and correlation length showed good behavior in estimations for the PM_{10} and $PM_{2.5}$. Variations in the local analysis radius tended to diminish the MFB, RMSE and CF for the PM_{10} and $PM_{2.5}$ estimates in figures 3.4 and 3.5. The increase in correlation time does not seem to have improved statistical measurements and in general presents a smaller impact in the data assimilation performance than the localization radius.

3.3.3. Calibration of observation error covariance

Since the observation network described in section 3.2.2 has not been previously used for a data assimilation experiment, no suitable formulation for the observation error representation covariance was present yet. We implemented the method proposed in (Desroziers *et al.*, 2005) to estimate the observation error covariance matrix $\mathbf{R}$. Desroziers *et al.* (2005) showed that the relation:

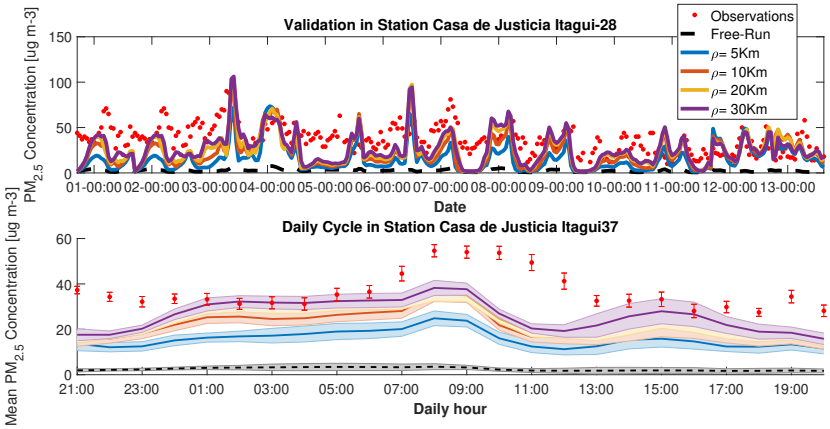


Figure 3.7: Time series and diurnal cycle of $PM_{2.5}$ in validation site 28. Lines as in Figure 3.6.

τ	MFB	RMSE	CF
1 day	-0.6353	40.8918	0.2603
2 day	-0.7815	43.4758	0.3082
3 day	-0.7096	42.1953	0.2834
5 day	-0.6832	41.6843	0.2802

Table 3.4: Statistical error evaluation for PM_{10} via MFB, RMSE and CF for both localization radius variation and correlation length

$$\mathbf{E}\left[\mathbf{d}_a^o(\mathbf{d}_f^o)^T\right] = \mathbf{R} \tag{3.4}$$

is valid if the matrices specified in

$$\mathbf{HK} = \mathbf{HP}^f\mathbf{H}^T\left(\mathbf{HP}^f\mathbf{H}^T + \mathbf{R}\right)^{-1} \tag{3.5}$$

are the true covariances for background and observation error. Here $\mathbf{K}$ is the Kalman gain, $\mathbf{d}_f^o$ is the difference between observations and forecast state in observation space and $\mathbf{d}_a^o$ is the difference between observations and analysis state in observation space. One application of this relationship is that it can be used to diagnose the observation error variance after the analysis cycle has been completed (Li *et al.*, 2009). In practice, the requirements for the relationship in Eq.(3.5) are never fully satisfied because the background covariance matrix is only an approximation of the real one. Nevertheless, the relationship could be used to obtain a useful first estimate of the observation error covariance matrix. For any subset of observations i with M observations, it is possible to compute an estimate of the error variance

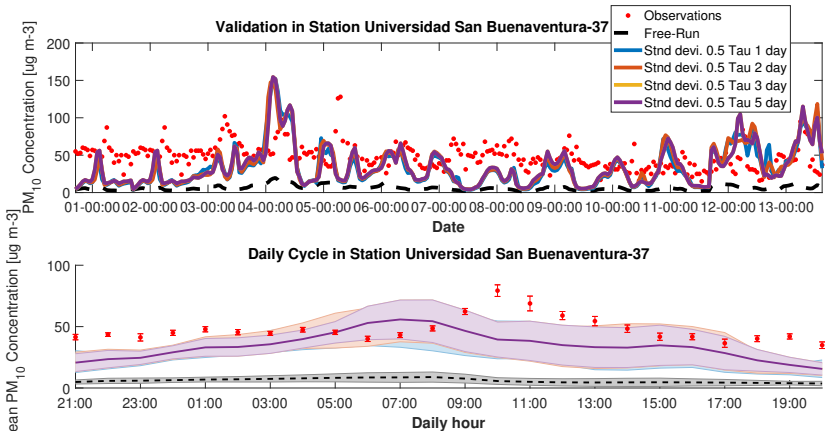


Figure 3.8: Time series and diurnal cycle of PM_{10} in validation site 37. Dots denote observations, dashed black lines are simulations by the standard model, and solid lines are analyzed ensemble means from the assimilation for different temporal correlation lengths. The diurnal cycles were obtained from 13 samples for each hour. The shadows and the bars represent the standard deviation of the 13 samples.

τ	**MFB**	**RMSE**	**CF**
1 day	-0.2183	38.0219	0.3455
2 day	-0.3731	21.8252	0.4019
3 day	-0.2854	21.8653	0.3774
5 day	-0.2513	21.4967	0.3746

Table 3.5: Statistical error evaluation for $PM_{2.5}$ via MFB, RMSE and CF for both localization radius variation and correlation length

$$\begin{aligned}\left(\hat{o}_o^2\right)_i &= \frac{(\mathbf{d}_a^o)_i^T(\mathbf{d}_f^o)_i}{M} \\ &= \sum_{j=1}^{M}\frac{(y_j - H(\mathbf{x^a})_j)(y_j - H(\mathbf{x^f})_j)}{M}\end{aligned} \tag{3.6}$$

where $\hat{o}_o^2$ correspond with the diagonal of the matrix $\mathbf{R}$.

The assimilation period from March 31 through April 13 was again used for calibration, in this case of $\mathbf{R}$. As an initial estimate the matrix $\mathbf{R}$ was filled with random Gaussian numbers to make the result independent of the initial value (Desroziers *et al.*, 2005; Li *et al.*, 2009)

For the subsequent experiment (test period, April 13-25), the off-line estimated matrix was used in the assimilation exercise. Once the DA scheme was calibrated, the estimated values for emissions correction factors were applied to the emissions

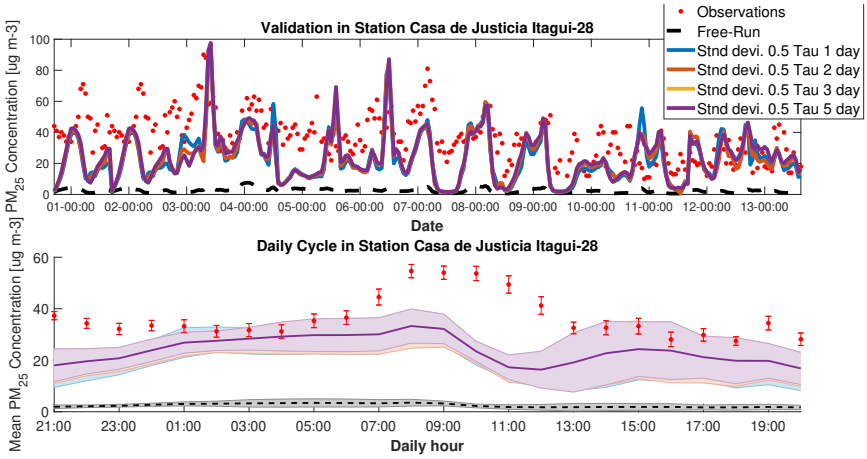


Figure 3.9: Time series and diurnal cycle of $PM_{2.5}$ in validation site 28. Lines as in Figure 3.8.

inventory in EDGAR V4.2 and taken as the nominal emissions inventory for the test period. The new DA iteration was done using the estimated values for $\mathbf{R}$, the radius size in the covariance localization scheme and τ.

3.4. Emissions estimation and particulate matter forecasting

A crucial requirement for an air quality simulation and assimilation system to contribute to the decision making process is that it be able to forecast pollution dynamics a few days in advance. The ability of the calibrated assimilation system to forecast concentrations of PM_{10} and $PM_{2.5}$ in the short term was evaluated during forecast experiments. Both weekend and weekday forecast starting points were assessed.

3.4.1. Model data assimilation with a calibrated scheme

Once the calibration process was completed, a new model run was conducted using the corrected emissions as nominal emission values, and a new 12-day data assimilation exercise was performed using the chosen fixed radius (for local analysis), time correlation length τ and the estimated observation error covariance $\mathbf{R}$ as was shown in Figure 3.4. In this second period the emissions were again estimated, but starting for the estimated emissions in the first period. Thus, the emissions were updated twice. It is important to note that experiments with other combinations of ρ and τ were performed but, in all the cases the results using the selected values in the previous section presented the best performance.The comparison between the nominal emissions of PM_{10} (from EDGAR data base) used in the first experiments

and the estimated emissions of PM_{10} used as nominal emissions for this new model runs is presented in Figure 3.10. Beyond the clear different in terms of magnitude and resolution between the estimated emissions and EDGAR that allow a better spacial representation of the emissions, the location of the hotspots in EDGAR appears in rural zones of the valley with minimal influence of human activity. The estimated emissions try to correct this behavior centring most of the emission in the urban zone of the valley, giving a more realistic representation.

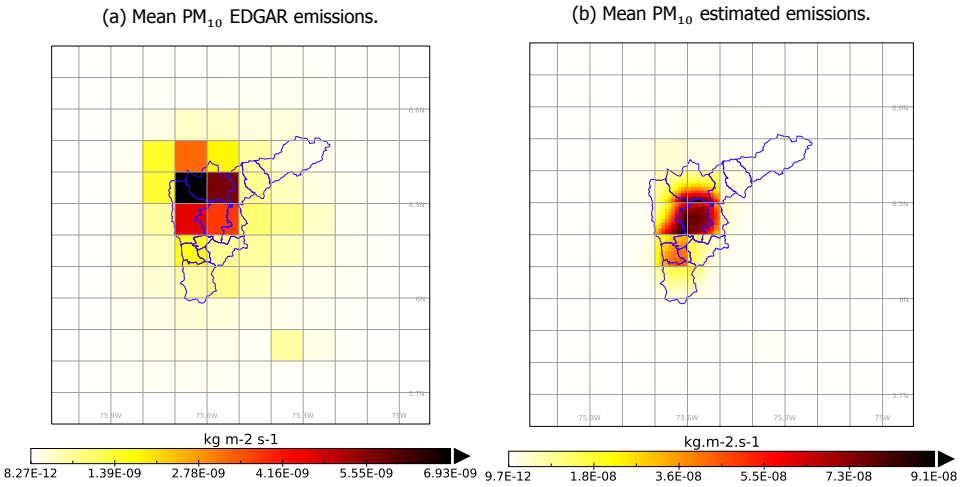


Figure 3.10: Comparison between EDGAR PM_{10} and estimated PM_{10} emissions.

The results of the assimilation for PM_{10} follow closely the measurements from validation station *Universidad San Buenaventura* (center of the valley) from April 13 at 19:00 UTC-5 through April 25 at 11:00 UTC-5 (see Figure 3.11). The peak near 18-00:00 (and in general almost all the day close to that hour) can be caused by an incorrect time profile in the emissions factors from EDGAR database, that does not reflect the real temporal dynamics of the emissions. Additionally, the meteorological fields can cause and increment of the concentrations levels. Note that the daily cycle for the assimilated model remains closer to the observations than the model without assimilation and with the previously estimated emissions.

Figure 3.12 shows a similar comparison for the $PM_{2.5}$ station. The model in a free run tends to over estimate the $PM_{2.5}$ concentrations (see peaks in 15 April at 23:00 UTC-5, 24 April at 22:00 and 25 April at 23:00 UTC-5). The results of the assimilation process offer a better average estimation. The daily cycle of $PM_{2.5}$ within the Aburrá valley is related to the industrial and mobile sources emissions profile and the meteorological conditions inside the valley.

The second period of assimilation provides a good representation not only of the dominant dynamics, but also offer an opportunity to forecast taking into account the profiles of emission sources. The next section will address the results from forecasts for the assimilated model with different radii in local analysis and correlation time lengths for the emissions.In the Appendix is presented a validation of the model for

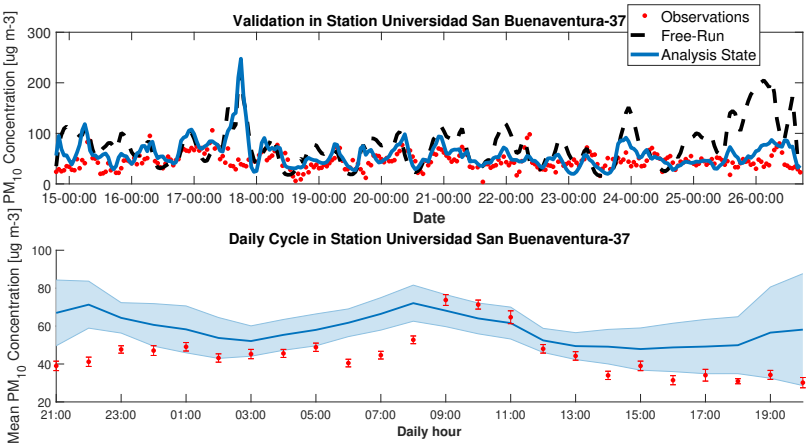


Figure 3.11: PM_{10} validation for the second DA iteration. Estimated emissions were used as nominal emissions, the estimated observation error covariance is used in the assimilation step. Red points are observations, solid black line is the free run model and the solid blue line is the analysis step for the assimilated model. The diurnal cycles were obtained from 13 samples for each hour. The shadows and the bars represent the standard deviation of the 13 samples. The time axis corresponds with the local time zone UTC-5.

O_3 and NO_2 concentrations for the second study period.

3.4.2. Forecasting PM profiles during weekdays and weekends

Using the second assimilation period (twice estimated emissions and the analysis state as initial condition, see Figure 3.4) three forecasts experiments were performed for up to three days, under the following scenarios: *i)* forecast starting on a Saturday night (19:00 UTC-5), with an assimilation window of the nine (9) days prior; *ii)* forecast starting on Tuesday night (19:00 UTC-5), with an assimilation window of the five (5) days prior; and *iii)* as in *ii)*, but using a localization radius of 5 km instead of 30 km.

Three different inheritance schemes (propagation of data assimilation information into forecast) for the emission correction factors were compared, namely:

1. Forecast default: Retaining only the state values from the end of the assimilation window. The correction factors estimated in the assimilation window are not used in the forecast.

2. Forecast hourly: Starting from the state values of the end of the assimilation window and using the hourly profile from the last 24 hours in the assimilation window.

3. Forecast average: Starting from the state values of the end of the assimilation window and using the average state values from the entire assimilation window.

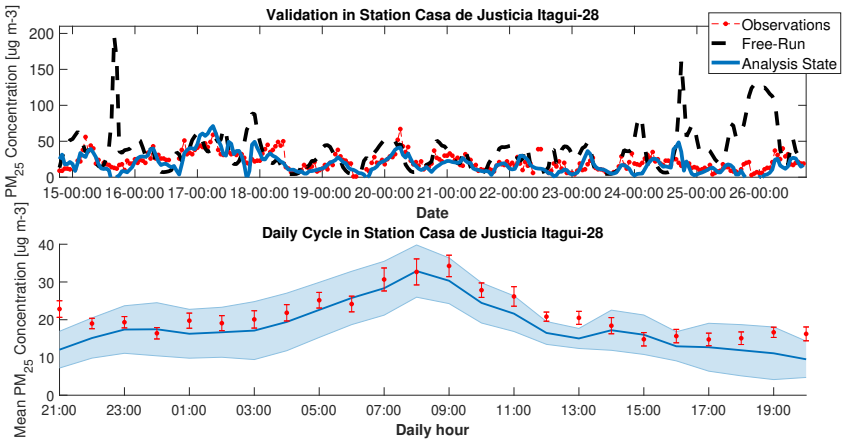


Figure 3.12: $PM_{2.5}$ validation for the second DA iteration. Estimated emissions were used as nominal emissions, the estimated observation error covariance was used in the assimilation step. Red points are observations, solid black line the free run model and solid blue line the analysis step for the assimilated model. The diurnal cycles were obtained from 13 samples for each hour. The shadows and the bars represent the standard deviation of the 13 samples. The time axis corresponds with the local time zone UTC-5.

For both PM species, forecasts starting on a weekend failed to reflect the observed dynamics. Forecasts initiating on a weekday were able, in general, to replicate the observed dynamics, having better performance in reproducing the dynamics of PM_{10} (Figure 3.14). All three inheritance schemes used in the PM_{10} weekday forecast (Figure 3.14) reproduced the observed dynamics, but the hourly scheme tracked the measured concentrations the closest. For the $PM_{2.5}$ weekday forecast, again the hourly profile tracked the measured concentrations more closely than the other two during the forecast period.

Figure 3.15 presents a comparison of the forecasts for PM_{10} and $PM_{2.5}$ under different local analysis radii. The smaller 5 km radius for local analysis, does not improve either forecast. We can interpret this phenomenon if we look at the forecast with 30 km radius in local analysis, taking into account that including more sensors in the assimilation it is possible to improve the correction factors in emissions. Consequently, if the emissions correction factors are higher for the latter two weeks in local analysis with 30 km radius, it can reduce the real emissions at a higher rate than the local analysis of 5 km radius via DA.

In order to provide quantitative measurements of forecast performance under various scenarios, the following error statistics were calculated and presented in the Figure 3.16: Mean Factoral Bias(MFB), Root Square Mean Error (RSME); and Correlation Factor (CF). The error statistics are calculated over a single forecast for each scenario and over the validations stations described in Figure 3.1. Since there were no considerable changes in the error statistics between the forecast days, the presented values correspond to the three-day average. Only PM_{10} are presented; the

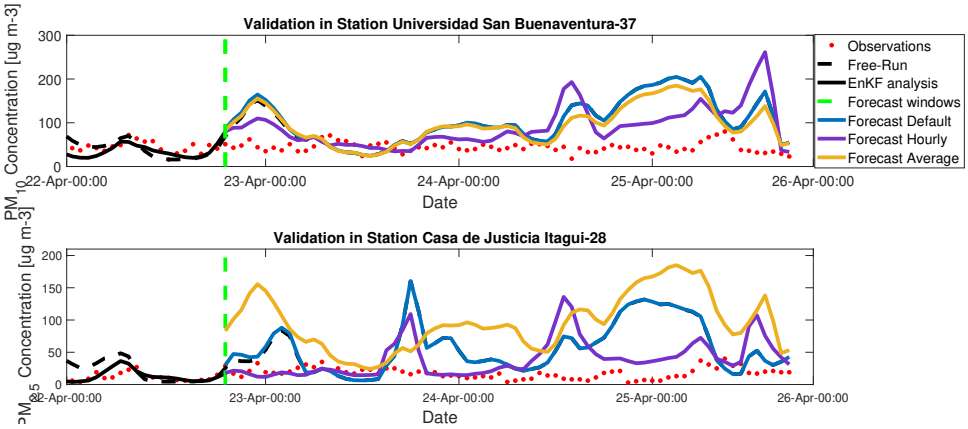


Figure 3.13: PM Forecast Starting on Saturday 19:00 UTC-5. The red points indicate observations; the dotted black line indicates the Free-Run; the solid black line shows the analysis of the EnKF; blue, purple and yellow lines show the forecasts under the different scenarios; vertical green line indicates the start of the forecasts.

behavior of $PM_{2.5}$ is very similar. Weekday forecasts (initiating on Tuesday) under a 30 km local analysis ratio scenario, presented the best error statistics. Independent of the inheritance scheme and the localization radius, weekend forecasts performed worse than weekday forecasts. For weekday forecasts, scenarios with localization radius of 5 km tended to perform worse than scenarios using a localization radius of 30 km.

3.5. Conclusions

Poor air quality is an environmental problem that many Colombian cities currently face. To avert the bi-annual deterioration in air quality due to the arrival of the Intertropical Convergence Zone, and in general to devise strategies to improve the quality of urban air, policy makers in Colombia and Northwest South America need accurate and reliable scientific information on atmospheric pollution dynamics for their decision making process. This study demonstrates that the LOTOS-EUROS model is suitable for use in regions of complex topography such as the Aburrá Valley, and paves the way for the creation of atmospheric pollution forecast systems fine tuned to the region that may assist the stated goal.

The use of regional, ground based atmospheric pollutant data from the SIATA sensor network, in data assimilation of the LOTOS-EUROS model via the use of the Ensemble Kalman Filter with covariance localization, improved the representation of PM_{10} and $PM_{2.5}$ dynamics and the estimation of their atmospheric concentrations within the Aburrá Valley.

Calibration of the radius for local analysis, the correlation time length τ and the estimation of the observation covariance error matrix $\mathbf{R}$, led to a better tuned DA scheme with improved performance, approaching more closely the available ob-

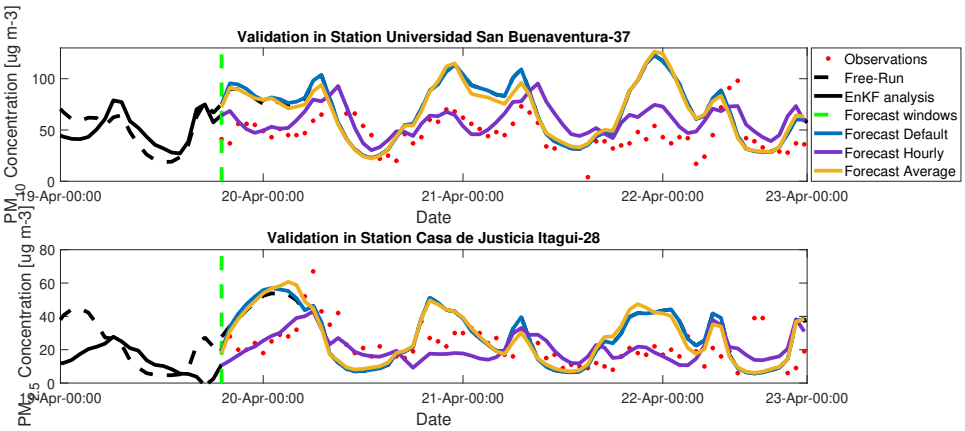


Figure 3.14: PM Forecast Starting on Tuesday 19:00 UTC-5. Lines as in Figure 3.13

servations. The estimation of an emission correction factor via data assimilation compensated for the scarcity in accurate and detailed emissions inventories for the region, enabling more accurate simulation results. Due to the coarse resolution of the emission inventory, and the rather low density of the sensor network available for 2016 within the area of interest, a large localization radius (30 km) performance better than a small radius (5 km).

Forecast performance was time and inheritance scheme sensitive, demonstrating that the temporal dynamics of pollutant emissions associated with the diurnal patterns of human activity need to be taken into account in the development of forecast systems. Inheritance schemes cognizant of complex system attributes (e.g., rugged topography, spatially heterogeneous and highly dynamic meteorology, etc.) may yield improved performance and increase the resolution and usability of air quality forecast systems.

Further research is needed using improved input data for the CTM , such as, for example, a local and more detailed emission inventory, and meteorology with a higher resolution that is better capable to represent the transport dynamics into the valley. Improvement of emission inventories and meteorological input is subject of current studies (see chapters 4 and 7). Additionally, a data assimilation scheme that also considers uncertainty in the meteorological variables and different emissions correction factors for each component could help to improve the presented results.

3.6. Supplementary

O_3, NO_x SO_2 are crucial for the secondary aerosol formation and the PM modeling (Barbu *et al.*, 2009; Manders *et al.*, 2009). In this Appendix (figures 3.17 and 3.18) a comparison is shown between the model concentrations for O_3 and NO_2 for the second period of Data Assimilation, using the calibrated DA scheme and the

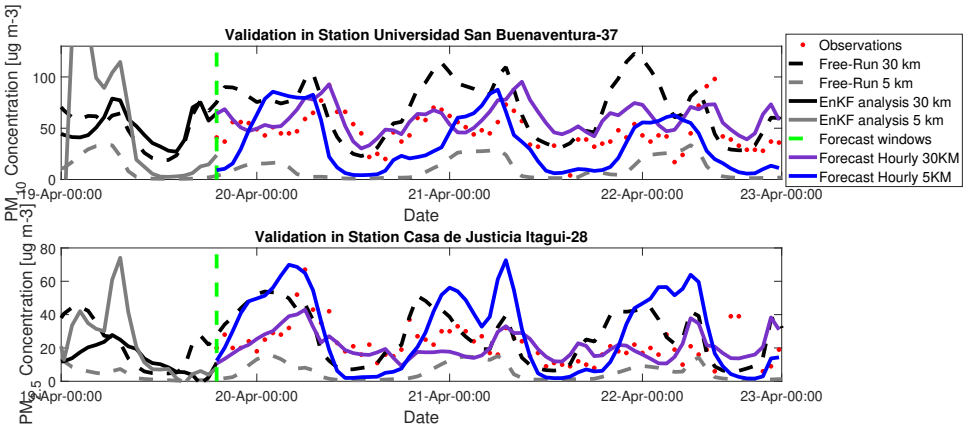


Figure 3.15: PM Weekday Forecast Comparisons for Different Radii. Red points depicts the observations, dotted black line the Free-Run using a 30 km radius, dotted grey line the Free-Run using a 5 km radius, solid black line the analysis of the EnKF using a 30 km radius, solid grey line the analysis of the EnKF using a 5 km radius, and purple and blue lines the forecast scenarios for 30 km and 5 km radius, respectively. Vertical green line indicates the beginning of the forecast window

estimated emissions (see Figure 3.4). Unfortunately, there is no data available from the SIATA network to evaluate the concentrations of NO_2 in the period of interest, and there are no others sources of high quality data over the region. The figures 3.17 and 3.18 show that in general the LE model tends to underestimate the O_3 and NO_2 concentrations, and not all the cycles are well captured by the model. These results support the idea that an improvement in the emission inventory and the meteorological fields are required to improve both the gases and the aerosol representation in the Aburrá Valley.

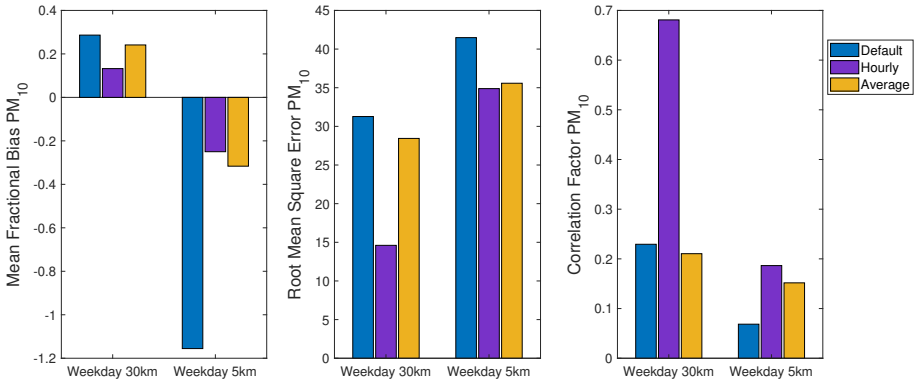


Figure 3.16: PM_{10} Forecast Error Statistics. Blue bars represent forecasts under the default inheritance scheme. Purple bars indicate forecasts under the hourly inheritance scheme. Yellow bars show forecasts under the average inheritance scheme.

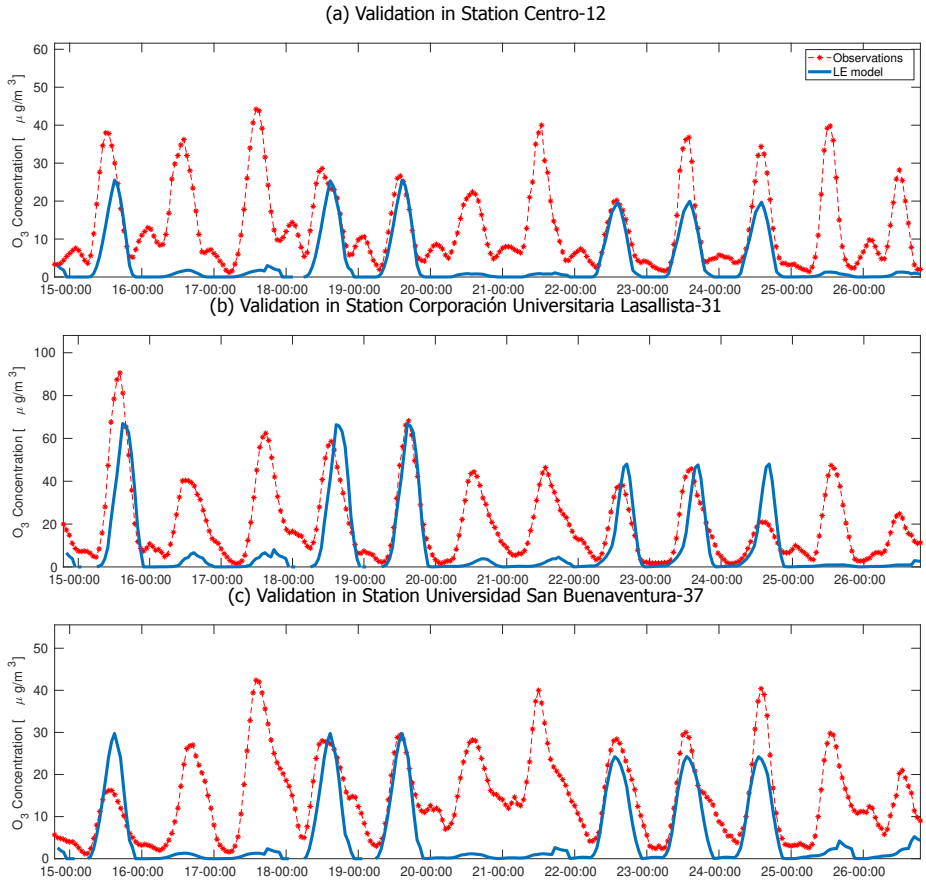


Figure 3.17: Comparison of LOTOS-EUROS O_3 concentration and SIATA observations. The time axis corresponds with the local time zone UTC-5.

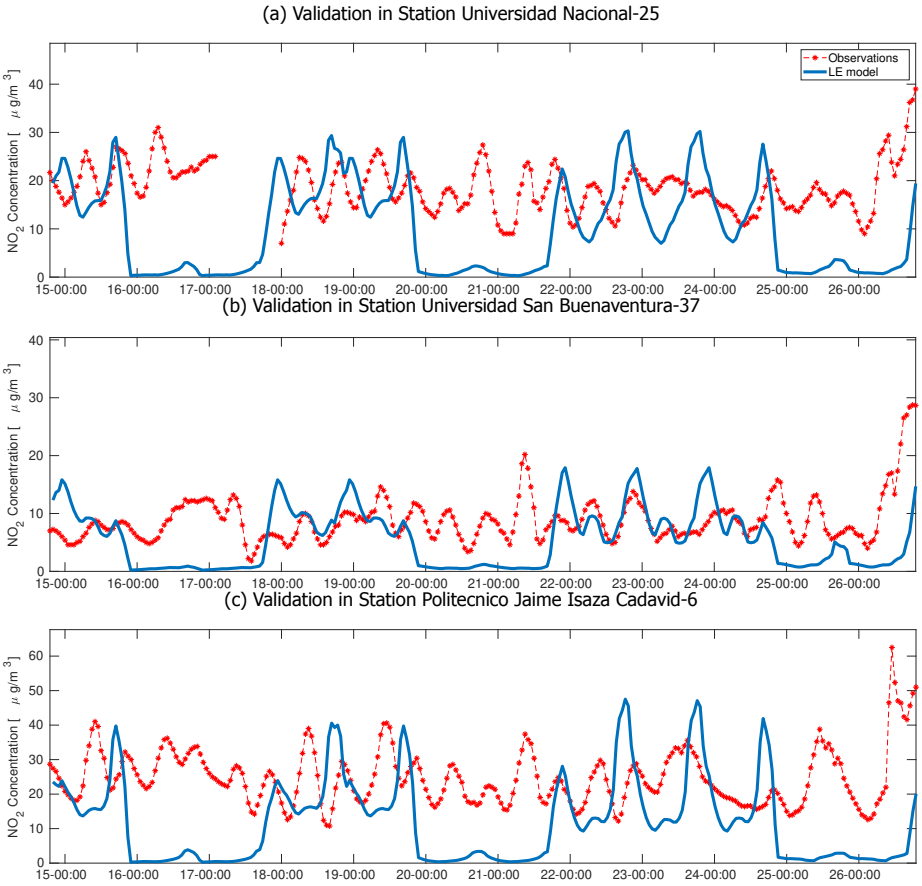


Figure 3.18: Comparison of LOTOS-EUROS NO_2 concentration and SIATA observations. The time axis corresponds with the local time zone UTC-5.

References

S. Lopez-Restrepo, A. Yarce, N. Pinel, O. Quintero, A. Segers, and A. Heemink, *Forecasting PM_{10} and $PM_{2.5}$ in the Aburrá Valley (Medellín, Colombia) via EnKF based Data Assimilation,* Atmospheric Environment **232**, 117507 (2020).

A. M. M. Manders, P. J. H. Builtjes, L. Curier, H. A. C. Denier Van Der Gon, C. Hendriks, S. Jonkers, R. Kranenburg, J. J. P. Kuenen, A. J. Segers, R. M. A. Timmermans, A. J. H. Visschedijk, R. J. W. Kruit, W. Addo, J. Van Pul, F. J. Sauter, E. Van Der Swaluw, D. P. J. Swart, J. Douros, H. Eskes, E. Van Meijgaard, B. Van Ulft, P. Van Velthoven, S. Banzhaf, A. C. Mues, R. Stern, G. Fu, S. Lu, A. Heemink, N. Van Velzen, and M. Schaap, *Curriculum vitae of the LOTOS–EUROS (v2.0) chemistry transport model,* Geosci. Model Dev **10**, 4145 (2017).

A. L. Barbu, A. J. Segers, M. Schaap, A. W. Heemink, and P. J. H. Builtjes, *A multi-component data assimilation experiment directed to sulphur dioxide and sulphate over Europe,* Atmospheric Environment **43**, 1622 (2009).

A. W. Heemink and A. J. Segers, *Modeling and prediction of environmental data in space and time using Kalman filtering,* Stochastic Environmental Research and Risk Assessment **16**, 225 (2002).

N. van Velzen and A. J. Segers, *A problem-solving environment for data assimilation in air quality modelling,* Environmental Modelling and Software **25**, 277 (2010).

J. S. Whitaker and T. M. Hamill, *Ensemble data assimilation without perturbed observations,* Monthly Weather Review **130**, 1913 (2002).

A. W. Heemink, M. Verlaan, and A. J. Segers, *Variance reduced ensemble kalman filtering,* Monthly Weather Review **129**, 1718 (2001).

M. Verlaan and A. W. Heemink, *Tidal flow forecasting using reduced rank square root filters,* Stochastic Hydrology and Hydraulics **11**, 349 (1997).

R. L. Curier, R. Timmermans, S. Calabretta-Jongen, H. Eskes, A. Segers, D. Swart, and M. Schaap, *Improving ozone forecasts over Europe by synergistic use of the LOTOS-EUROS chemical transport model and in-situ measurements,* Atmospheric Environment **60**, 217 (2012).

A. M. R. Petrescu, R. Abad-Viñas, G. Janssens-Maenhout, V. N. B. Blujdea, and G. Grassi, *Global estimates of carbon stock changes in living forest biomass: EDGARv4.3 - time series from 1990 to 2010,* Biogeosciences **9**, 3437 (2012).

C. M. Gonzalez, C. D. Gomez, N. Y. Rojas, H. Acevedo, and B. H. Aristizabal, *Relative impact of on-road vehicular and point-source industrial emissions of air pollutants in a medium-sized Andean city,* Atmospheric Environment **152**, 279 (2017).

J. E. Pachón, B. Galvis, O. Lombana, L. G. Carmona, S. Fajardo, A. Rincón, S. Meneses, R. Chaparro, R. Nedbor-Gross, and B. Henderson, *Development and evaluation of a comprehensive atmospheric emission inventory for air quality modeling in the megacity of Bogotá,* Atmosphere **9**, 1 (2018).

R. Nedbor-Gross, B. H. Henderson, M. P. Pérez-Peña, and J. E. Pachón, *Air quality modeling in Bogotá Colombia using local emissions and natural mitigation factor adjustment for re-suspended particulate matter,* Atmospheric Pollution Research **9**, 95 (2018).

J. W. Boylan and A. G. Russell, *Pm and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models,* Atmospheric Environment **40**, 4946 (2006), special issue on Model Evaluation: Evaluation of Urban and Regional Eulerian Air Quality Models.

T. Chai and R. R. Draxler, *Root mean square error (rmse) or mean absolute error (mae): Arguments against avoiding rmse in the literature,* Geoscientific Model Development **7**, 1247 (2014).

S. Yu, B. Eder, R. Dennis, S.-H. Chu, and S. E. Schwartz, *New unbiased symmetric metrics for evaluation of air quality models,* Atmospheric Science Letters **7**, 26 (2006), https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/asl.125 .

A. Kumar, R. Jimenez, L. C. Belalcazar, and N. Y. Roja, *Application of wrf-chem model to simulate pm10 concentration over bogota,* Aerosol and Air Quality Research **16**, 1206 (2016).

G. Desroziers, L. Berre, B. Chapnik, and P. Poli, *Diagnosis of observation, background and analysis-error statistics in observation space,* Quarterly Journal of the Royal Meteorological Society **131**, 3385 (2005).

H. Li, E. Kalnay, and T. Miyoshi, *Simultaneous estimation of covariance inflation and observation errors within an ensemble kalman filte,* Quarterly Journal Of The Royal Meteorological Society **135**, 4523 (2009).

A. M. Manders, M. Schaap, and R. Hoogerbrugge, *Testing the capability of the chemistry transport model LOTOS-EUROS to forecast PM10 levels in the Netherlands,* Atmospheric Environment **43**, 4050 (2009).

4

Urban Air Quality Modeling Using Low-Cost Sensor Network and Data Assimilation

In this study we show the evaluation of the operational Aburrá Valley's low-cost network against the official monitoring network. The results show that the $PM_{2.5}$ low-cost measurements are very close to those observed by the official network. We integrate low-cost observations with the chemical transport model LOTOS-EUROS using data assimilation. Two different configurations of the low-cost network were assimilated: using the whole low-cost network (255 sensors), and a high-quality selection using just the sensors with a correlation factor greater than 0.8 with respect to the official network (115 sensors). Both simulations assimilating the low-cost model outperform the model without assimilation and with assimilation of the official network. The capability to issue warnings for pollution events is also improved by assimilating the low-cost network with respect to the other simulations. Finally, the simulation using the high-quality configuration has lower error values than using the complete low-cost network, showing that it is essential to consider the quality and location and not just the total number of sensors. Our results suggest that with the current advance in low-cost sensors, it is possible to improve model performance with low-cost network data assimilation.

Part of this chapter has been published in:
(Lopez-restrepo *et al.*, 2021) *Urban Air Quality Modeling Using Low-Cost Sensor Network and Data Assimilation in the Aburrá Valley, Colombia*, **Atmosphere, 12, 91, 1–19**

4.1. Introduction

The integration of observations from dense networks of low-cost sensors into mathematical models through techniques such as data fusion or data assimilation enables a spatially continuous representation of concentration fields with significantly reduced bias (Lahoz and Schneider, 2014). These techniques provide an added value to the sensor observations by spatially interpolating between monitoring locations and at the same time adding value to the model by constraining the model with observations. Both sources of information can thus be combined in a mathematically objective manner with the goal of reducing the uncertainty inherent to both sources (Schneider *et al.*, 2017; Liu *et al.*, 2019; Popoola *et al.*, 2018). Although data assimilation is a more complex family of methods than data fusion or interpolation techniques, it is by far the most versatile and the robust of these approaches (Lahoz and Schneider, 2014).

This chapter seeks to implement the data assimilation technique *Ensemble Kalman Filter* (EnKF) (Evensen, 2003) to integrate data from a hyper-dense, low-cost $PM_{2.5}$ measuring network operated in Medellín (Colombia) and its neighboring municipalities of the Aburrá Valley (Hoyos *et al.*, 2019) into the Chemical Transport Model LOTOS-EUROS (Manders *et al.*, 2017). Data generated by the robust, official network of air quality monitoring stations in the Aburrá Valley were previously used for data assimilation in LOTOS-EUROS for modeling and forecasting PM dynamics in the valley (Lopez-Restrepo *et al.*, 2020). The goal with using data from the low-cost sensor network is to evaluate the impact of hyper-dense observations in the data assimilation approach and their viability as an alternative to monitoring $PM_{2.5}$ concentrations in developing countries. This study differs from previous studies such as (Schneider *et al.*, 2017; Popoola *et al.*, 2018; Ahangar *et al.*, 2019; Pournazeri *et al.*, 2014), in which a dispersion model was used to construct concentration maps or to estimate emissions from the measured concentration fields, and the integration of the model and observations was based on Kriging or other static approaches. In this work a dynamic data assimilation method is implemented to guide the model's concentration fields using the observations.

The main contributions from this chapter are as follows: 1) an evaluation of the low-cost sensor network against the official network; 2) the implementation of techniques for the assimilation of low-cost high-density data, focusing on the impact on the assimilated model results; and 3) a methodology for performing and evaluating PM forecasts with assimilated data over three-day windows, providing valuable information for decision makers.

4.2. Materials and methods

The period of interest for all data evaluations, simulations and data assimilation experiments spans from February 25 to March 15, 2019. During these days, the PM concentrations are higher due to the Northbound transit of the Inter-Tropical Convergence Zone.

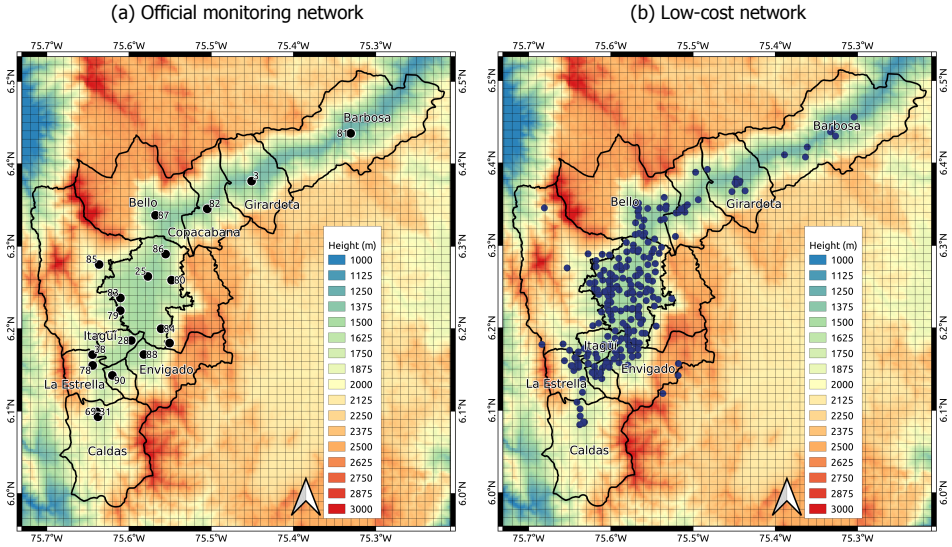


Figure 4.1: Spatial distribution of the hyper-dense low-cost network Citizen Scientist and official monitoring air-quality network for $PM_{2.5}$. The gray raster represent the LOTOS-EUROS model grid, the black lines are the boundaries of the municipalities borders, and the numbers are the official station numerations followed by SIATA.

4.2.1. Hyper-dense low-cost sensor network

In Medellín and its greater metropolitan area inside the Aburrá Valley, the *Sistema de Alerta Temprana del Valle de Aburrá* (SIATA) project operates the official high-end air quality monitoring network (henceforth *official network*, and a hyper-dense, low-cost air quality network developed within the Citizen Scientist program (henceforth *low-cost network*).

The official network provides high quality measurements for different pollutants in the atmosphere over the Aburrá Valley such as O_3, SO_2, PM_{10}, $PM_{2.5}$ and PM_1. The official network is distributed among the ten municipalities of the Valley, with the majority of the stations located within the city of Medellín (Figure 4.1, (a)). The low-cost network was created with the aim of engaging the community in issues surrounding air quality, and as an extension of the official network. As of writing, the low-cost network consists of 255 real-time $PM_{2.5}$ sensors across the Aburrá Valley and its hills. They are located in the premises of private homes and public or private institutions (Figure 4.1, (b)). The description of the network deployment is presented in (Hoyos *et al.*, 2019). Data were downloaded from SIATA's data portal[1]. Data from the official network for the corresponding dates were used for validation of both the low-cost network and the model simulations before and after data assimilation. Each station from the official network served as a reference point for all low-cost network sensors within a 2-km radius of the former. Performance of

[1]available at https://siata.gov.co/descarga_siata/index.php/index2/. Last accessed, December 2020.

the latter was evaluated using as metrics the Mean Fractional Bias (MFB), the Root Mean Square Error (RMSE) and the Pearson correlation coefficient (R) introduced in Section 3.2.3 (Chai and Draxler, 2014; Boylan and Russell, 2006; Shaocai *et al.*, 2006). When a low-cost sensors had more than one official station within a 2-km radius, the average value of the official measurements was used.

4.2.2. LOTOS-EUROS Model and Local Emission Inventory

All the simulations were conducted using the domain and experimental setup described in Section 2.1.2. The local emission inventory presented in Section 2.1.3 was used as emission input for all the simulations.

4.2.3. Ensemble Kalman Filter

The EnKF system in this application is configured to obtain estimates of both concentrations and emissions, following the stochastic representation and the EnKF implementation presented in Section 2.2. For all the simulations we used a correlation length τ of 1 day and a variance of the stochastic process σ of 0.5 following previous results (Lopez-Restrepo *et al.*, 2020). Additionally, we used a covariance localization radius ρ =5 km for all the simulations. We used an ensemble of N =25 members. Additional experiments with larger ensembles were performed without improvements in performance (not shown).

Two sets of low-cost sensors data were assembled: The first one included 255 sensors from the low-cost network that had a station from the official network within a 2-km radius. The second, higher quality one consisted of a subset of the previous set, including only those sensors whose data showed an R value equal or greater than 0.8 when evaluated against the official network.

We performed four different LOTOS-EUROS simulations:

1. a LOTOS-EUROS model simulation without data assimilation (henceforth *LE*);
2. a simulation with assimilation of data (observations) from the 14 stations of the official network (henceforth *LE-official*. The 14 stations were selected randomly and are indicated as red squares in Figure 4.4);
3. a simulation with assimilation of the data from the entire low-cost network (henceforth *LE-lowcost*)
4. a simulation with assimilation only of high-quality data from the low-cost network (henceforth *LE-lowcost-HQ*).

The 7 stations from the official network that were not used for data assimilation (green stars in Figure 4.4) were used as validation stations for all simulations.

4.2.4. Forecast experiments

Data assimilation can improve forecast performance mainly for two reasons: First, if the simulation is initialed with an assimilated field value, initial conditions at the start of the forecast window be a representation closer to reality than what the model alone may provide; second, the emission correction factors that were included in

the assimilation state (2.7) can be applied to the model during the forecast window to adjust the emissions in the same direction as during assimilation.

Forecasting experiments were conducted to evaluate the capabilities of the model with data assimilation to forecast PM concentrations in the valley up to three days. Simulations were carried out as above, with the assimilation schedule illustrated in Figure 4.2. Data assimilation was conducted up to the indicated date, with the three subsequent days representing the forecast window. The forecasting window started at 00:00 hours of the first day after the end of data assimilation. To bring the information obtained in the assimilation window into the forecast window, we used the hourly profile of the correction factor calculated from the last 24 hours of data assimilation. The experiments continued until all days between March 9 and March 13 (inclusive) had predictions as the first, second and third day of the forecast. The performance of the forecast was evaluated by calculating the Air Quality Index (AQI) according to the ranges established by the Metropolitan Area [2] and illustrated in Table 4.1; and comparing it to the AQI observed for the corresponding day. The comparison against the AQI rather than against plain PM concentrations facilitates the interpretation of the model forecast performance by decision makers and the general public. Additionally, this representation affords evaluating the ability of the model to predict warning-triggering episodes (AQI in orange, red or purple levels). Forecasts missing warning-triggering episodes (false negatives) are especially problematic in air quality management because the resulting inaction can lead to human exposure to dangerous concentrations of pollutants.

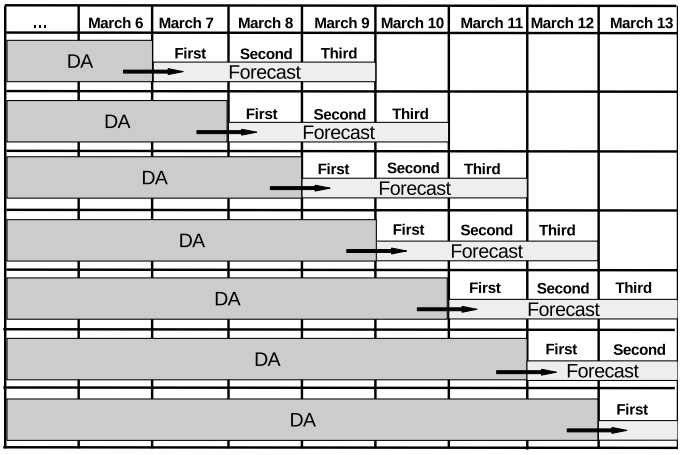


Figure 4.2: Graphic explanation of the experimental forecast setup. The arrows represent the inheritance of the last correction factor 24-hourly profile into the forecast. All simulations start at February 23 19:00 UTC-5. A spin-up of 5 previous days was taken for each simulation.

[2] available in Spanish https://www.metropol.gov.co/ambiental/calidad-del-aire/Documents/POECA/Plan_de_Acci%C3%B3n_POECA_Metropolitano_2019.pdf. Last accessed, October 2020.

Pollutant	Average time	Average Concentration ($\mu g/m^3$)				
		No warning		Warning		
		Green	Yellow	Orange	Red	Purple
$PM_{2.5}$	24 hours	0-12	13-37	38-55	56-150	≥ 151

Table 4.1: Air Quality Index (AQI) as defined for the Aburrá Valley with respect to $PM_{2.5}$ concentrations.

4.3. Results

4.3.1. Evaluation with low-cost sensor network

The performance of 145 sensors from the low-cost network was evaluated against data from the official network. The remaining 110 sensors did not have an official monitoring station within a 2-km radius. Figure 4.3 shows the histograms of the MFB, RMSE and R, and the geographical distribution of the performance values. For the majority (67%) of the low-cost sensors an MFB between -0.25 and 0.25 was obtained, with an average of about 0.2. Average RMSE was close to 8 $\mu g/m^3$, with most sensors presenting values below 15 $\mu g/m^3$. The majority (88%) of the sensors showed correlations with R values above 0.7. Observed errors fell within acceptable ranges (Boylan and Russell, 2006; Shaocai *et al.*, 2006). Zonal differences in measurement errors were observed. Locations in the South-central part of the city of Medellín (green ellipse on Figure 4.3.1 (d), (e), and (f)) contained most of the sensors with a R values lower than 0.5 and RMSE values grater than 15 $\mu g/m^3$. Those sensors are located in a dense urban area, while the closest monitoring stations is located in the outskirts of the city. Figure 4.4 shows the spatial distribution of the complete low-cost network and subset of 115 low-cost sensors with the highest quality data (as defined in section 4.2.3). The selection of the low-cost high quality is based in the results showed in Figure 4.3.1(b) and (e).

4.3.2. Evaluation of data assimilation runs

The concentration fields generated by the model simulations with or without data assimilation were compared to the observations from seven of the official monitoring stations (*validation stations*, green stars in Figure 4.4) to evaluate the performance of the data assimilation schemes. Figure 4.5 shows the temporal series for the simulated and observed $PM_{2.5}$ concentrations at four of the validation stations. The four selected stations represent downtown Medellín (station 25), residential areas (station 86), areas with high vehicular flow (station 88), and a peri-urban area in the outskirts of the city (station 85). Those stations summarize the behavior of all seven validation stations. The LE simulation consistently underestimated the concentrations observed at stations 85 and 88. At stations 25 and 86, the LE simulation results were close in magnitude between February 24 and March 3 and March 10 to March 15; between March 3 and March 10, the model presented values much lower than those observed. The day-to-day variability was reduced for this same period, as seen in stations 85 and 86. This inconsistent behavior suggests a poor representation of the meteorological dynamics that govern the dispersion and accumulation of $PM_{2.5}$ within the valley. Simulations using data assimilation showed

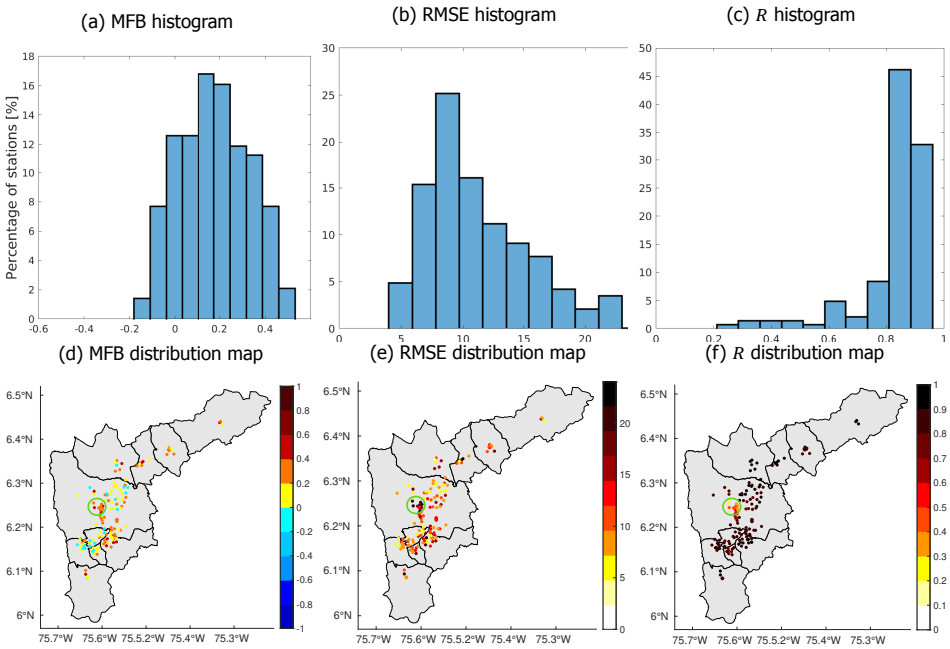


Figure 4.3: Evaluation of low-costs network against the official monitoring network for the period between 25-February-2019 and 15-March-2019.

noisier behaviors than the LE simulation. This process is commonly observed when applying the EnKF and obeys the stochastic nature and the handling of uncertainty inherent to the method (Evensen, 2003). However, those simulations managed to correct the large discrepancies present in the LE simulation. Both LE-official, LE-lowcost, and LE-lowcost-HQ represented more accurately the day-to-day variability of the observations than LE. In general terms, there was no evidence of a sizeable and persistent difference among the simulations with data assimilation throughout the entire period. Nevertheless, the LE-lowcost-HQ simulation reproduced with greater accuracy the concentrations observed in different periods, such as between February 26 and March 4 in station 25, between March 9 and March 14 in stations 85 and 86.

Figure 4.6 shows the diurnal cycles during the simulation period in the four selected validations stations. The diurnal cycle of the LE simulation differed from the observations in both magnitude and temporal behavior. The highest concentration peak that appears around 09:00 in all the stations is mainly due to traffic dynamics. In stations 25 and 88, the LE morning peak corresponded in time but not in magnitude with the observations; in stations 85 and 86, said peak appeared later in the simulations than in the observations. This time lag suggests a poor spatial representation of mobile emissions by the emissions inventory; or a deficiency it the wind fields in reproducing the valley dynamics, showing a late transport of the particulate material to these areas. The LE simulation did not capture the evening

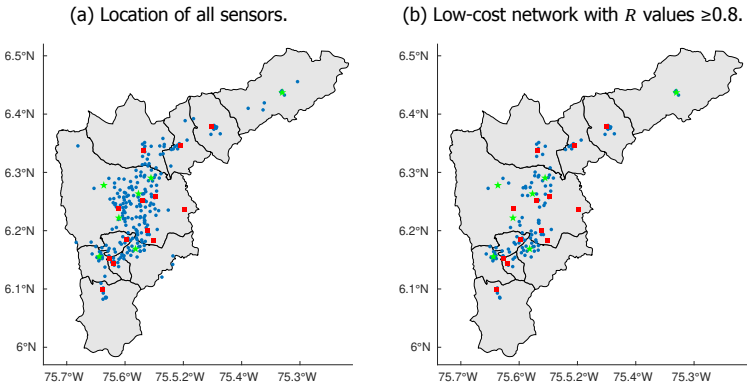


Figure 4.4: Spatial distribution of the different sets of sensors used for assimilation and validation. Blue dots indicate the location of the low-cost network sensors. Red squares correspond to the locations of the official monitoring stations that were used for data assimilation. Green stars indicate the stations from the official network whose data where used for validation of all model simulations.

peak shown by the observations around 21:00 hours. The simulations using data assimilation presented diurnal cycles closer to the observations than did the LE simulation. The LE-official simulation captured the time and magnitude of the morning peak in stations 85 and 86. In station 88, LE-official corrected the time lag in the morning peak seen in LE, and improved the estimated magnitudes albeit still falling short of the observed values. A different behavior was seen for station 25, where LE-official had low diurnal variability, with a slight underestimation in the morning, and an overestimation in the afternoon. The LE-lowcost and LE-lowcost-HQ simulations results resembled closely the diurnal behavior of the observations, especially the temporal component. In all the stations, both the morning and the evening peaks matched the observations. The observed concentrations for stations 25 and 88 fell inside the standard deviation range for the LE-lowcost simulation; the same simulation overestimated the concentrations between 11:00 and 19:00 for station 85, and underestimated the concentrations between 01:00 and 13:00 for station 86. The LE-lowcost-HQ simulation results were overall the closest to observations.

The averaged evaluation statistics among all the validation station are shown in Table 4.2. The simulation results without data assimilation (LE) underestimated the observed concentrations in all the validation stations. This was also seen in previous related works (Lopez-Restrepo *et al.*, 2020; Henao *et al.*, 2020). The RMSE value reflected a low correspondence between the observed and simulated concentrations when using the model without data assimilation. The correlation coefficient was low, meaning that the model was not able to capture the variations in diurnal and day-to-day concentrations. In contrast, the three simulations using data assimilation had MFB values close to 0, without a significant difference among them. The data assimilation was thus effective in reducing between the model and reality. The RMSE also improved when using data assimilation, decreasing by 24.4% in the LE-official, 32.8% in the LE-lowcost, and 36.2% in the LE-lowcost-HQ simulations relative to the RMSE of the LE simulation. The R values were all above

the criteria of good performance according with (Mogollón-sotelo *et al.*, 2020) Table 2, and based in (Boylan and Russell, 2006; EPA, 2000). Assimilation of either data set from the low-cost network resulted in improved error statistics when compared to the LE-official simulation.

Simulation	MFB	RMSE	R
LE	-0.65	27.38	0.42
LE-official	-0.07	20.69	0.64
LE-lowcost	0.08	18.39	0.76
LE-lowcost-HQ	0.06	17.46	0.82

Table 4.2: Mean Fractional Bias, Root Mean Square Error and Pearson Correlation Coefficient for simulated $PM_{2.5}$. Values are averaged over all the validation stations for the simulation period.

4.3.3. Evaluation of forecasts

Figure 4.7 shows a graphical evaluation of the model forecasts for March 12 as day 1, 2 or 3 within the forecasting window. Forecasts for all other days within the forecasting experiment behaved similarly. The observed AQIs and the values for the LE simulation are the same in all the graphs since all graphs illustrate the same calendar day (March 12). Similar to the results shown in section 4.3.2, the LE simulation underestimated $PM_2.5$ concentrations throughout the valley, yielding in most cases AQI lower than those reported. The AQI forecasts of the three simulations with data assimilation were consistently more accurate than the estimates from the simulation without assimilation (LE). There were no significant differences in performance among the three data assimilation simulations through the three forecast days. Their forecast accuracy decreased as the forecasting window advanced, as could be expected from the uncertainty inherent in the meteorological fields and nominal emission factors. All three simulations with data assimilation had similar spatial behavior, with a tendency to underestimate the AQI in the Northern and Eastern areas of the valley.

For public information on air quality, it is essential that a forecast correctly warns for a critical pollution event. Figure 4.8 shows the confusion matrix for LE-official, LE-lowcost, and LE-lowcost-HQ simulations in the data assimilation and forecast windows. The confusion matrix summarizes the percentage of true negatives, true positives, false negatives, and false positives (Kohavi and Provost, 1998). The data assimilation evaluation is performed just in the seven validation stations shown in Figure 4.4. The LE simulation does not offer a warning in any station in the assimilation nor forecast windows; for that reason, its confusion matrix is not presented. In the assimilation window, data assimilation simulations have a percentage of true negatives and true positives higher than 80%, and even higher than 90% in the case of the LE-lowcost-HQ. Both simulations using the low-cost network show lower false negative values than LE-official. The LE-lowcost-HQ has the highest accuracy in reproducing the warning-triggering events within the data assimilation window. The accuracy of the three simulations is lower in the forecast window than in the

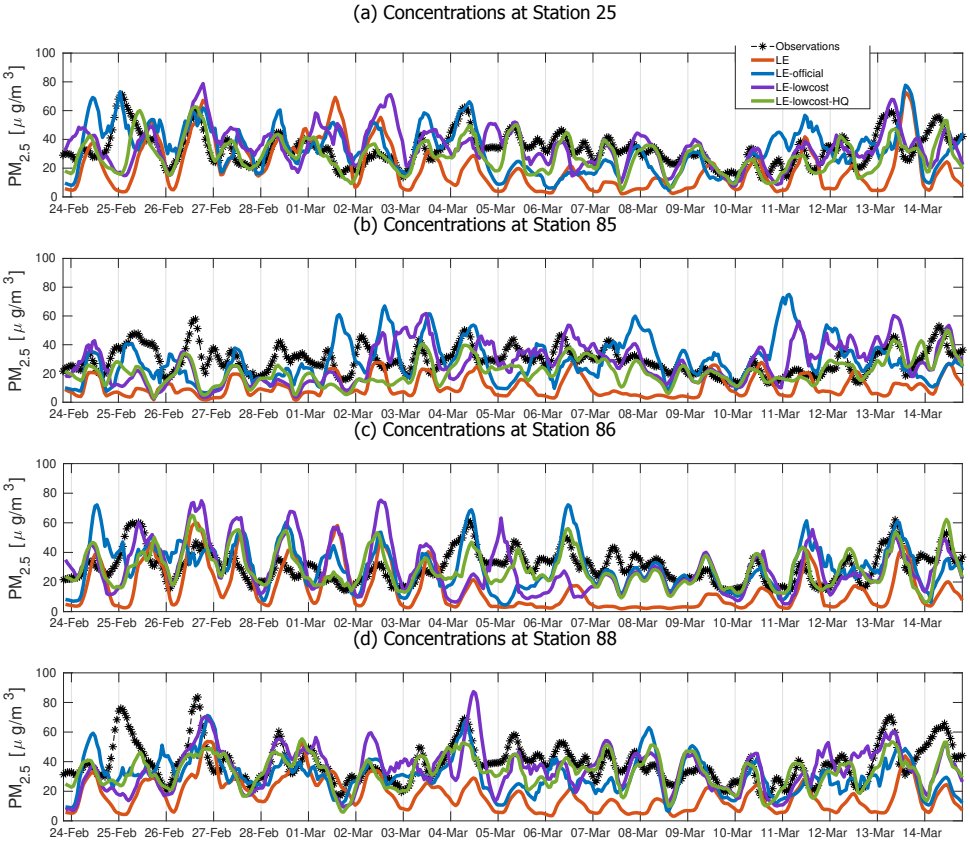


Figure 4.5: Temporal series of $PM_{2.5}$ concentrations from selected validation stations of the official network, LOTOS-EUROS without assimilation, LE-official, LE-lowcost and LE-lowcost-HQ. Time stamps are valid for local time (UTC-5). A spin-up of 5 previous days was taken for each simulation.

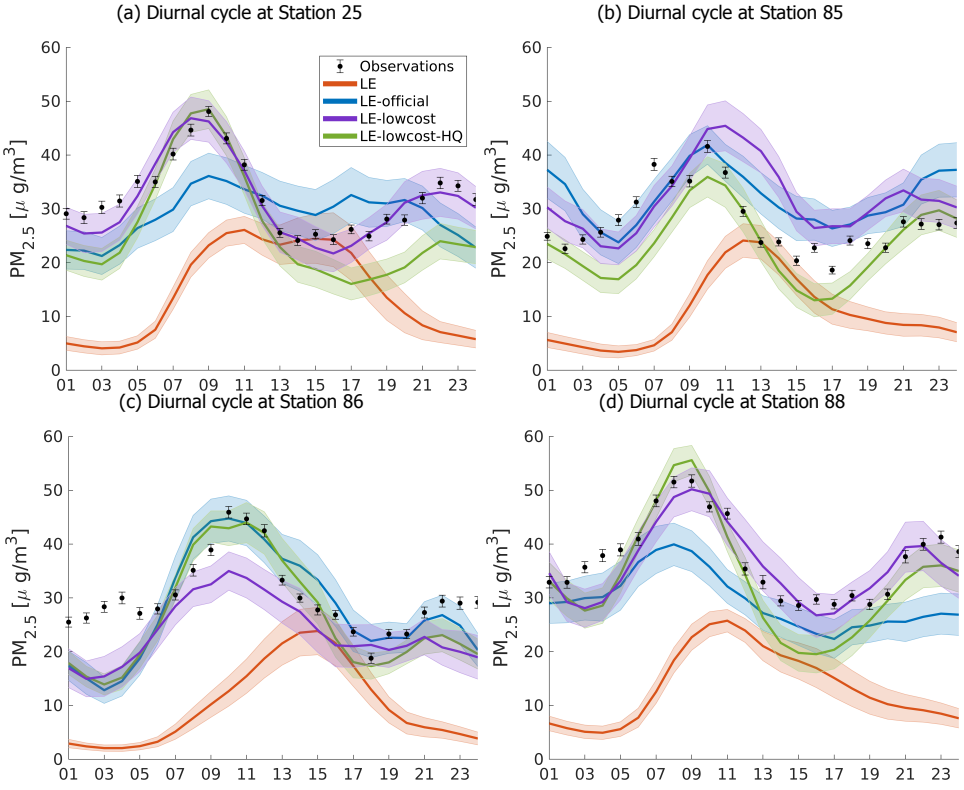


Figure 4.6: Diurnal cycle of $PM_{2.5}$ concentrations from selection stations of the official network, LOTOS-EUROS without assimilation, LE-official, LE-lowcost and LE-lowcost-HQ. The bars and the shadows represent the standard deviation over the simulation period. The time stamps are valid for local time (UTC-5).

assimilation window. The small percentage of false positives and high percentage of false negatives suggests that even using the estimated emissions inventory, the simulations continue to underestimate the observations. As observed within the data assimilation window, the two simulations assimilating data from the low-cost network (LE-lowcost and LE-lowcost-HQ) had better warning forecast performance than the LE-official simulation.

4.4. Discussion and comments

The experiments described in this chapter show that it is currently possible to develop low-cost networks with high performance even for cities with air quality problems such as Medellin. The high spatial density of the low-cost network allowed much higher spatial resolution than that attained with the official network. The errors in the low-cost sensors located within the green ellipse in Figure 4.3.1 (d), (e) and (f) represented spatial outliers. The increased errors observed in this sector of the Valley may be attributed to specific factors such as maintenance, characteristics of the infrastructure in which the sensors are located, differences in elevation relative to the official station against which they were evaluated, or particular meteorological conditions within the subregion of the Valley that may yield local heterogeneity in PM concentrations. Said green ellipse corresponds to a transition zone between peri-urban terrain and an expanding horizon of high-density residential buildings. The low-cost sensors are located in said buildings, while the official monitoring station is located in a school surrounded by forests. This may explain the apparent overestimation of the PM levels by the low-cost sensors and the low correlation values of their data.

Our results show a low correlation values and a high underestimation of the observed concentration by the LOTOS-EUROS model without assimilation. Similar behavior were observed in previous works (Lopez-Restrepo *et al.*, 2020; Henao *et al.*, 2020). In Henao *et al.* (2020) the WRF-Chem model in a sub-kilometer configuration was used to reproduce the CO dynamics in the valley. The emission inventory was obtained from the AMVA Official Emission Inventory (UPB and AMVA, 2017) and following a methodology similar to the presented in Section 2.1.3. Although the meteorological fields showed a high similarity with observations, the model underestimated the CO concentrations. The underestimation in both cases is attributed to mismatches in the official emission inventory and uncertainties generated by the simplifications of disaggregation methodologies. However, data assimilation notably improves the ability of LOTOS-EUROS to represent the magnitude and dynamics of $P_{2.5}$ within the Aburrá Valley. The assimilation of data from the low-cost network improves the correlation between the observed and the simulated concentrations to a greater extent than when data from the sparse official network is assimilated, both in terms of the RMSE and the R values. The errors left in the simulated concentrations after the assimilation of the low-cost network are within the performance goals for $PM_{2.5}$ representation formulated in (Boylan and Russell, 2006; Shaocai *et al.*, 2006; EPA, 2000; Chang and Hanna, 2004). The uncertainty present in the model causes the percentage of predicted alarm-triggering events related to high concentration of $PM_{2.5}$, to decrease to almost one half of the events

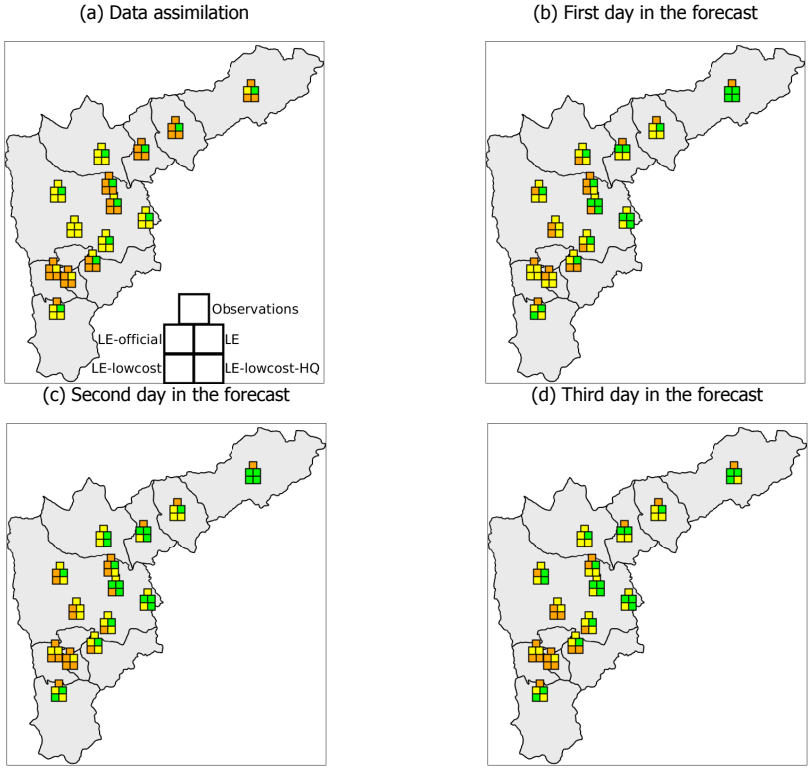


Figure 4.7: Evaluation of Air Quality Index (AQI) forecast capabilities of LOTOS-EUROS for the Aburrá Valley. All figures represents the forecasts for March 12 when it corresponded to the first (a), second (b) and third (c) day within the forecasting window. The five-square markers are located at the geographical location of each of the official stations used for comparisons. The upper-center square is the AQI calculated from the observed PM values, against which all other values are compared; the middle-left inner square is the AQI predicted by the LE-official simulation; the middle-right inner square is the AQI predicted by the model without assimilation; the bottom-left inner square the AQI predicted by the LE-lowcost simulation; and the bottom-right inner square is the AQI predicted by the LE-lowcost-HQ simulation. The AQI definition is as Table 4.1.

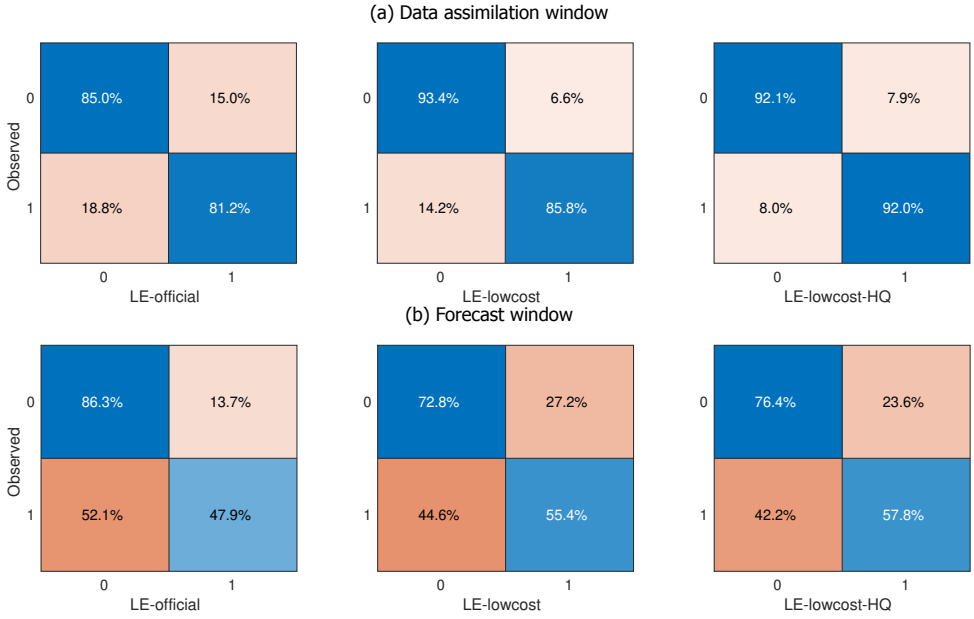


Figure 4.8: Comparison of confusion matrices for the data assimilation and forecast window depending on warning or no warning per station. The values are calculated across all the days of the corresponding window. The value of 0 corresponds with no warning, the value of 1 corresponds with a warning. For the LE simulation, there are no warnings in the data assimilation window nor forecast windows.

observed within the forecasting window (Figure 4.8). Our results highlight the persistent need to improve the inventories of nominal emissions, the meteorological data used, and to reduce other sources of uncertainty in the model in order to increase forecasting capacity. Nevertheless, the model's forecasting capacity is increased when observations are assimilated. The greater spatial coverage of the low-cost network contributed significantly to the improvements against the simulations assimilating data from the official network. The higher density of observations also allowed estimating emissions in more detail, as seen in Figure 4.6. The more detailed emission estimations also allowed a better reproduction of the concentrations in the forecast window even in the absence of data assimilation.

Although the LE-lowcost simulation used more observations than the LE-lowcost-HQ simulation (255 and 115, respectively), the location and quality of the additional observations played an important role. The LE-lowcost-HQ was defined using a high similarity criterion to the official network, so it was not as affected by observations with low quality as LE-lowcost. Comparisons between Figure 4.4 (a) and Figure 4.4 (b) reveal that the additional locations did not increase the spacial density considerably relative to the low-cost high quality sensors. Our results suggested that while a high observation density is essential for improving the performance of a model with data assimilation, it is crucial to consider other factors such as quality of the data and the location of the sensors. Different techniques of observation localization allow optimizing the number of sensors to improve the data assimilation or other data fusion techniques (Alexanderian *et al.*, 2016; King *et al.*, 2015; Mazzoleni *et al.*, 2017; Yildirim *et al.*, 2009). We highly recommend implementing these techniques in the development of a new low-cost network. Apart from minimizing the number of sensors and associated costs, the processing of a reduced number of observations requires less computational resources. As an example, the LE-lowcost simulation was 3.2 times slower than the LE-lowcost-HQ using the same computation configuration. Optimization of computational and time resources are especially important for operational systems.

Jointly with previous work (Schneider *et al.*, 2017; Popoola *et al.*, 2018; Ahangar *et al.*, 2019; Johnston *et al.*, 2019; Isakov *et al.*, 2019; Moltchanov *et al.*, 2015), our results can support and motivate the development of future low-cost networks and their integration in data fusion applications. According to the literature, North America, Europe, and China concentrate most of the current low-cost implementations, with experimental, citizen, and data dissemination purposes (Kumar and Gurjar, 2019; Morawska *et al.*, 2018). In developing countries, a low-cost network, together with a CTM and data assimilation can provide a valuable first approach to monitoring PM without the high cost of an official air quality network.

4.5. Conclusions

We represented a data assimilation application of a hyper-dense low-cost PM network and the chemical transport model LOTOS-EUROS in a urban setting. The low-cost network provided high quality data comparable to those provided by the official monitoring network. The performance of the model with assimilation of the spatially-dense data from the low-cost network improved both in terms of its

representation of the observed dynamics, as well as in its forecast capabilities, highlighting its value as an air-quality management tool. Our results support the idea than with the current advances in the low-cost sensors, it is possible to use low-cost networks and data assimilation to model and predict air quality in urban areas.

Although one of the main advantages of a low-cost network is that it could provide a hyper-dense networks with relative low costs, it is recommended to invest in the quality of the data (sensor quality, calibration, maintenance) and the study of optimal localization. High quality data and appropriate choices for the number and the location of the sensor strongly improves the data assimilation process and minimizes operational and computational costs.

References

S. Lopez-restrepo, A. Yarce, N. Pinel, O. Quintero, A. Segers, and A. W. Heemink, *Urban Air Quality Modeling Using Low-Cost Sensor Network and Data Assimilation in the Aburrá Valley , Colombia,* Atmosphere **12**, 1 (2021).

W. a. Lahoz and P. Schneider, *Data assimilation: making sense of Earth Observation,* Frontiers in Environmental Science **2**, 1 (2014).

P. Schneider, N. Castell, M. Vogt, F. R. Dauge, W. A. Lahoz, and A. Bartonova, *Mapping urban air quality in near real-time using observations from low-cost sensors and model information,* Environment International **106**, 234 (2017).

H. Y. Liu, P. Schneider, R. Haugen, and M. Vogt, *Performance assessment of a low-cost PM 2.5 sensor for a near four-month period in Oslo, Norway,* Atmosphere **10** (2019), 10.3390/atmos10020041.

O. A. Popoola, D. Carruthers, C. Lad, V. B. Bright, M. I. Mead, M. E. Stettler, J. R. Saffell, and R. L. Jones, *Use of networks of low cost air quality sensors to quantify air quality in urban settings,* Atmospheric Environment **194**, 58 (2018).

G. Evensen, *The Ensemble Kalman Filter: Theoretical formulation and practical implementation,* Ocean Dynamics **53**, 343 (2003).

C. D. Hoyos, L. Herrera-Mejía, N. Roldán-Henao, and A. Isaza, *Effects of fireworks on particulate matter concentration in a narrow valley: the case of the medellín metropolitan area,* Environmental Monitoring and Assessment **192**, 6 (2019).

A. M. M. Manders, P. J. H. Builtjes, L. Curier, H. A. C. Denier Van Der Gon, C. Hendriks, S. Jonkers, R. Kranenburg, J. J. P. Kuenen, A. J. Segers, R. M. A. Timmermans, A. J. H. Visschedijk, R. J. W. Kruit, W. Addo, J. Van Pul, F. J. Sauter, E. Van Der Swaluw, D. P. J. Swart, J. Douros, H. Eskes, E. Van Meijgaard, B. Van Ulft, P. Van Velthoven, S. Banzhaf, A. C. Mues, R. Stern, G. Fu, S. Lu, A. Heemink, N. Van Velzen, and M. Schaap, *Curriculum vitae of the LOTOS–EUROS (v2.0) chemistry transport model,* Geosci. Model Dev **10**, 4145 (2017).

S. Lopez-Restrepo, A. Yarce, N. Pinel, O. Quintero, A. Segers, and A. Heemink, *Forecasting PM_{10} and $PM_{2.5}$ in the Aburrá Valley (Medellín, Colombia) via EnKF based Data Assimilation,* Atmospheric Environment **232**, 117507 (2020).

F. E. Ahangar, F. R. Freedman, and A. Venkatram, *Using low-cost air quality sensor networks to improve the spatial and temporal resolution of concentration maps,* International Journal of Environmental Research and Public Health **16** (2019), 10.3390/ijerph16071252.

S. Pournazeri, S. Tan, N. Schulte, Q. Jing, and A. Venkatram, *A computationally efficient model for estimating background concentrations of nox, no2, and o3,* Environmental Modelling and Software **52**, 19 (2014).

T. Chai and R. R. Draxler, *Root mean square error (rmse) or mean absolute error (mae): Arguments against avoiding rmse in the literature,* Geoscientific Model Development **7**, 1247 (2014).

J. W. Boylan and A. G. Russell, *Pm and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models,* Atmospheric Environment **40**, 4946 (2006), special issue on Model Evaluation: Evaluation of Urban and Regional Eulerian Air Quality Models.

Y. Shaocai, E. Brian, D. Robin, C. Shao□Hang, and S. S. E., *New unbiased symmetric metrics for evaluation of air quality models,* Atmospheric Science Letters **7**, 26 (2006).

J. J. Henao, J. F. Mejía, A. M. Rendón, and J. F. Salazar, *Sub-kilometer dispersion simulation of a CO tracer for an inter-Andean urban valley,* Atmospheric Pollution Research **11**, 0 (2020).

C. Mogollón-sotelo, L. Belalcazar, and S. Vidal, *A support vector machine model to forecast ground-level* $pm_{2.5}$ *in a highly populated city with a complex terrain,* Air Quality, Atmosphere & Health (2020).

EPA, *Meteorological Monitoring Guidance for Regulatory Modeling Applications*, Tech. Rep. (U.S. ENVIRONMENTAL PROTECTION AGENCY, 2000).

R. Kohavi and F. Provost, *Applications of Machine Learning and the Knowledge,* Applications of Machine Learning and the Knowledge in Machine Learning **30**, 349 (1998).

UPB and AMVA, *Inventario de Emisiones Atmosféricas del Valle de Aburrá - actualización 2015*, Tech. Rep. (Universidad Pontificia Bolivariana - Grupo de Investigaciones Ambientales, Area Metropolitana del Valle de Aburra, Medellín, 2017).

J. C. Chang and S. R. Hanna, *Air quality model performance evaluation,* Meteorology and Atmospheric Physics **87**, 167 (2004).

A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, *A fast and scalable method for a-optimal design of experiments for infinite-dimensional bayesian nonlinear inverse problems,* SIAM Journal on Scientific Computing **38**, A243 (2016), https://doi.org/10.1137/140992564 .

S. King, W. Kang, and L. Xu, *Observability for optimal sensor locations in data assimilation,* International Journal of Dynamics and Control **3**, 416 (2015), cited By 5.

M. Mazzoleni, L. Alfonso, and D. Solomatine, *Influence of spatial distribution of sensors and observation accuracy on the assimilation of distributed streamflow data in hydrological modelling,* Hydrological Sciences Journal **62**, 389 (2017), cited By 6.

B. Yildirim, C. Chryssostomidis, and G. Karniadakis, *Efficient sensor placement for ocean measurements using low-dimensional concepts,* Ocean Modelling **27**, 160 (2009), cited By 55.

S. J. Johnston, P. J. Basford, F. M. Bulot, M. Apetroaie-Cristea, N. H. Easton, C. Davenport, G. L. Foster, M. Loxham, A. K. Morris, and S. J. Cox, *City scale particulate matter monitoring using LoRaWAN based air quality IoT devices,* Sensors (Switzerland) **19**, 1 (2019).

V. Isakov, S. Arunachalam, R. Baldauf, M. Breen, P. Deshmukh, A. Hawkins, S. Kimbrough, S. Krabbe, B. Naess, M. Serre, and A. Valencia, *Combining dispersion modeling and monitoring data for community-scale air quality characterization,* Atmosphere **10** (2019), 10.3390/atmos10100610.

S. Moltchanov, I. Levy, Y. Etzion, U. Lerner, D. M. Broday, and B. Fishbain, *On the feasibility of measuring urban air pollution by wireless distributed sensor networks,* Science of the Total Environment **502**, 537 (2015).

A. Kumar and B. R. Gurjar, *Low-Cost Sensors for Air Quality Monitoring in Developing Countries - A Critical View,* Asian Journal of Water, Environment and Pollution **16**, 65 (2019).

L. Morawska, P. K. Thai, X. Liu, A. Asumadu-Sakyi, G. Ayoko, A. Bartonova, A. Bedini, F. Chai, B. Christensen, M. Dunbabin, J. Gao, G. S. Hagler, R. Jayaratne, P. Kumar, A. K. Lau, P. K. Louie, M. Mazaheri, Z. Ning, N. Motta, B. Mullins, M. M. Rahman, Z. Ristovski, M. Shafiei, D. Tjondronegoro, D. Westerdahl, and R. Williams, *Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone?* Environment International **116**, 286 (2018).

5

An Efficient Ensemble Kalman Filter Implementation Via Shrinkage Covariance Matrix Estimation: Exploiting Prior Knowledge

In this chapter, we propose an efficient and practical implementation of the ensemble Kalman filter via shrinkage covariance matrix estimation. Our filter implementation combines information brought by an ensemble of model realizations, and that based on our prior knowledge about the dynamical system of interest. We perform the combination of both sources of information via optimal shrinkage factors. The method exploits the rank-deficiency of ensemble covariance matrices to provide an efficient and practical implementation of the analysis step in EnKF based formulations. Localization and inflation aspects are discussed as well. Experiments are performed to assess the accuracy of our proposed filter implementation by employing an Advection Diffusion Model and an Atmospheric General Circulation Model. The experimental results reveal that the use of our proposed filter implementation can mitigate the impact of sampling noise, and even more, it can avoid the impact of spurious correlations during assimilation steps.

Part of this chapter has been published in:
(Lopez-Restrepo et al.,2021) An Efficient Ensemble Kalman Filter Implementation Via Shrinkage Covariance Matrix Estimation: Exploiting Prior Knowledge, **Computational Geosciences,25, 985–1003**

5.1. Introduction

A dynamical system, approximately evolves according to some imperfect numerical model:

$$\mathbf{x}_{\text{current}} = \mathcal{M}_{t_{\text{previous}} \to t_{\text{current}}} \left(\mathbf{x}_{\text{previous}} \right), \tag{5.1}$$

where n and m are the model resolution and the number of observations, respectively, and $\mathcal{M} : \mathbb{R}^{n \times 1} \to \mathbb{R}^{n \times 1}$ is an imperfect model operator which mimics the behavior of a very highly non-linear system such as the ocean and/or the atmosphere.

On the former representation, the model operator maps the state variable into a sequential time steps realization of the behavior of the dynamical system. In most of the cases, there is a control variable included on the operator that related external inputs to the system and allows for the representation of the interactions between the system and the external world. The state variable may or may not be directly measurable and is used as a memory of the system. As seen in equation 5.1, the past behavior of the system affects its future development, but the lack of representation of the state variable may be a pitfall on the full representation of the real world. The relationship between the state space and the real noisy observation $\mathbf{y} \in \mathbb{R}^{m \times 1}$ is sometimes a useful tool for the proper understanding and representation of the full system.

Controllability is a property of the dynamical system that allows measuring the ability of a particular control input to manipulate all the states of the system, taking them from a point *A* to the point *B* in finite time. On the other hand, observability measures the ability of the particular sensor configuration to supply all the information necessary to estimate all the states of the system. State estimation and Parameter estimation are typically the main concerns in control and systems theory. They are required for the proper control law design and are mandatory for the full observability of the system.

In cases when there is a lack of observability, the problem of state estimation and parameter estimation arose, and it can be solved by means of the solution to the optimal filtering problem. That requires an analytical solution of the Bayes theorem by means of the Kushner or Zakai Equation. These are not feasible for non-linear and non-Gaussian systems. They are approximated most often via particle filters (Quintero M *et al.*, 2008, 2009). The linear and Gaussian case is solved by the well known Kalman filter, and its extension to non-linear and Gaussian cases can be found extensively in the literature. For Large scale systems, the solutions to complete the full observability of the system are not straight forward because the course of dimensionality and more sophisticated solutions to the optimal filtering problem were derived.

Sequential Data Assimilation (DA) is a statistical process that optimally combines information brought by an imperfect numerical forecast $\mathbf{x}^b \in \mathbb{R}^{n \times 1}$ and a real noisy observation $\mathbf{y} \in \mathbb{R}^{m \times 1}$ (Evensen, 2003; Anderson and Anderson, 1999) to estimate the actual state $\mathbf{x}^* \in \mathbb{R}^{n \times 1}$ of a dynamic system such as Equation 5.1. When

Gaussian assumptions are made over prior and observational errors via Bayes' rule, the posterior estimate has the form:

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B} \cdot \mathbf{H}^T \cdot \mathbf{A}^{-1} \cdot \mathbf{d} \in \mathbb{R}^{n \times 1}, \tag{5.2}$$

where $\mathbf{B} \in \mathbb{R}^{n \times n}$ is the background error covariance matrix, $\mathbf{d} = \mathbf{y} - \mathcal{H}\left(\mathbf{x}^b\right) \in \mathbb{R}^{m \times 1}$ is the vector of innovations (on the observations), $\mathcal{H}(\mathbf{x}) : \mathbb{R}^{n \times 1} \to \mathbb{R}^{m \times 1}$ is the observation operator (which maps vector states to observations), $\mathcal{H}(\mathbf{x}) \approx \mathcal{H}\left(\mathbf{x}^b\right) + \mathbf{H} \cdot \left[\mathbf{x} - \mathbf{x}^b\right] \in \mathbb{R}^{m \times n}$, $\mathbf{H} \in \mathbb{R}^{m \times n}$ is the Jacobian of $\mathcal{H}(\mathbf{x})$ at $\mathbf{x}^b$, the information matrix reads:

$$\mathbf{A} = \left[\mathbf{R} + \mathbf{H} \cdot \mathbf{B} \cdot \mathbf{H}^T\right] \in \mathbb{R}^{m \times m}, \tag{5.3}$$

and $\mathbf{R} \in \mathbb{R}^{m \times m}$ is the estimated data-error covariance matrix. In practice, an ensemble of model realizations can be employed to estimate the parameters $\mathbf{x}^b$ and $\mathbf{B}$ of prior error distributions. However, given the computational cost of a single model propagation, ensemble sizes are constrained by the hundreds while their underlying error distribution by the millions. Consequently, sampling errors impact the quality of analysis innovations: ensemble covariances are rank-deficient, and even more, they are ill-conditioned (Ott *et al.*, 2004; Anderson, 2001). Thus, spurious correlations among distant model components are developed in the ensemble covariance (Nino-Ruiz *et al.*, 2020). Localization methods are commonly employed during assimilation steps to mitigate the impact of sampling noise. In this context, well-known methods are covariance matrix localization, precision matrix localization, spatial domain localization, and observation impact localization. The selection of one method over the others relies on computational aspects. Yet another manner to mitigate the impact of spurious correlations is based on Shrinkage Covariance Matrix Estimation. In this family of covariance matrix estimators, the background error covariance matrix is estimated as the convex combination of a target matrix $\mathbf{T} \in \mathbb{R}^{n \times n}$, and the ensemble covariance $\mathbf{P}^b \in \mathbb{R}^{n \times n}$:

$$\widehat{\mathbf{B}} = \gamma \cdot \mathbf{T} + (1 - \gamma) \cdot \mathbf{P}^b \in \mathbb{R}^{n \times n}, \text{ for } \gamma \in [0, 1]. \tag{5.4}$$

The current literature proposes ensemble-based formulations via the covariance estimator (5.4) in which:

1. the target matrix $\mathbf{T}$ is diagonal (no prior structure is assumed for $\mathbf{B}$), and the weight γ is optimally computed via loss functions (Nino-Ruiz and Sandu, 2015; Nino-Ruiz and Sandu), or

2. the target matrix $\mathbf{T}$ is static (i.e., it retains climatological information), and the weight γ is ranged in $\gamma \in [0, 1]$ (Wang *et al.*, 2008, 2007).

We exploit the opportunity to include our prior knowledge about the structure of $\mathbf{B}$, the information brought by samples from the model dynamics, and the optimal estimation of γ. In this manner, we can obtain a covariance matrix estimator of $\mathbf{B}$ that optimally combines all sources of information. While several techniques have been proposed to reduce spurious correlations, most of them are designed for a specific

problem, and it is not possible to generalize them for other DA implementations (Fu *et al.*, 2017; Lu *et al.*, 2016). We are looking for a robust and generalizable manner to include previous knowledge of the system to a large scale Chemical Transport Model (CTM) for air quality purposes.

5.2. Preliminaries

In order to state the value of the current contribution, several questions must be solved to demonstrate the feasibility of the new data assimilation technique in an operational fashion (Zhu *et al.*, 2002): *Does the new method provide guidance that is of higher quality or more use than existing methods? Is the potential benefit of running a new technique cost-effective? Is the new method sufficient with respect to old methods?*. In this section, we discuss ensemble-based data assimilation methods and how those can be implemented in current operational settings. These concepts are necessary to develop our filter formulation.

5.2.1. Ensemble-Based Data Assimilation

In ensemble-based data assimilation, an ensemble of model realizations

$$\mathbf{X}^b = \left[\mathbf{x}^{b[1]}, \mathbf{x}^{b[2]}, \ldots, \mathbf{x}^{b[N]}\right] \in \mathbb{R}^{n \times N}, \tag{5.5}$$

is employed to estimate the parameters $\mathbf{x}^b$ and $\mathbf{B}$ of prior error distributions, where $\mathbf{x}^{b[e]} \in \mathbb{R}^{n \times 1}$ is the e-th ensemble member, for $1 \leq e \leq N$, and N stands for ensemble size. Hence:

$$\mathbf{x}^b \approx \overline{\mathbf{x}}^b = \frac{1}{N} \cdot \sum_{e=1}^{N} \mathbf{x}^{b[e]} \in \mathbb{R}^{n \times 1}, \tag{5.6}$$

and

$$\mathbf{B} \approx \mathbf{P}^b = \frac{1}{N} \cdot \Delta\mathbf{X} \cdot \Delta\mathbf{X}^T \in \mathbb{R}^{n \times n}, \tag{5.7}$$

where

$$\Delta\mathbf{X} = \mathbf{X}^b - \overline{\mathbf{x}}^b \cdot \mathbf{1}^T \in \mathbb{R}^{n \times N}, \tag{5.8}$$

is the matrix of member deviations, $\overline{\mathbf{x}}^b$ is the ensemble mean, $\mathbf{P}^b$ is the ensemble covariance, and $\mathbf{1}$ is a vector of consistent dimension whose components are all ones. Once an observation is available, the posterior state can be computed via the stochastic Ensemble Kalman Filter (EnKF) (Evensen, 2003):

$$\mathbf{X}^a = \mathbf{X}^b + \mathbf{P}^b \cdot \mathbf{H}^T \cdot \left[\mathbf{R} + \mathbf{H} \cdot \mathbf{P}^b \cdot \mathbf{H}^T\right]^{-1} \cdot \mathbf{D} \in \mathbb{R}^{n \times N}, \tag{5.9}$$

where the e-th column of the innovation matrix on the synthetic observations $\mathbf{D} \in \mathbb{R}^{n \times N}$ reads $\mathbf{d}^{[e]} = \mathbf{y} + \epsilon^{[e]} - \mathcal{H}\left(\mathbf{x}^{b[e]}\right) \in \mathbb{R}^{m \times 1}$, with $\epsilon^{[e]} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$. In practice, ensemble sizes are constrained by the hundreds, while model resolutions are

bounded by the millions, which mainly obey computational aspects. Consequently, the quality of analysis corrections can be impacted by spurious correlations. Hence, localization methods can be employed to mitigate the impacts of sampling errors. Well-known methods in this context are covariance matrix localization, spatial domain localization, and observation localization.

5.2.2. Shrinkage Covariance Matrix Estimation

A more robust family of covariance estimators under the DA case $n \gg N$ are the shrinkage based estimators (Touloumis, 2015; Couillet and McKay, 2014). This kind of estimators follow the form (Ledoit *et al.*, 2018):

$$\mathbf{B} \approx \widehat{\mathbf{B}}(\alpha) = \alpha \cdot \mathbf{T} + (1-\alpha) \cdot \mathbf{P}^b \in \mathbb{R}^{n \times n}, \tag{5.10}$$

where $\alpha \in [0, 1]$, and $\mathbf{T} \in \mathbb{R}^{n \times n}$ is known as the Target matrix. The resulting estimator is a convex combination of the ensemble covariance matrix and the predefined $\mathbf{T}$ matrix. When there is not available information about the structure of $\mathbf{B}$, an alternative for $\mathbf{T}$ is (Nino-Ruiz and Sandu):

$$\mathbf{T} = \frac{\mathbf{trace}\left(\mathbf{P}^b\right)}{n} \cdot \mathbf{I}, \tag{5.11}$$

where $\mathbf{I} \in \mathbb{R}^{n \times n}$ is the identity matrix. The value of α is chosen to minimize the loss function

$$\alpha^* = \arg\min_{\alpha} \mathbb{E}\left[\left\|\mathbf{B} - \widehat{\mathbf{B}}(\alpha)\right\|_F^2\right], \tag{5.12}$$

where $\|\bullet\|_F$ represents the Frobenius norm. For target matrices of the form (5.11), a distribution-free formulation for the optimal α^*_{LW} is proposed by Ledoit and Wolf in (Ledoit and Wolf, 2004a):

$$\alpha^*_{LW} = \min\left(\frac{\sum_{e=1}^{N}\left\|\mathbf{P}^b - \Delta\mathbf{x}^{[e]} \cdot \Delta\mathbf{x}^{[e]^T}\right\|_F^2}{N^2 \cdot \left[\mathbf{trace}\left(\mathbf{P}^{b\,2}\right) - \frac{\mathbf{trace}\left(\mathbf{P}^b\right)^2}{n}\right]}, 1\right), \tag{5.13}$$

where $\Delta\mathbf{x}^{[e]} \in \mathbb{R}^{n \times 1}$ denotes the e-th column of matrix (5.8). Based on the LW estimator, for Gaussian samples, the Rao-Blackwell Ledoit and Wolf (RBLW) one is proposed. In the RBLW estimator, the optimal weight is defined by:

$$\alpha^*_{RBLW} = \min\left(\frac{\frac{N-2}{n} \cdot \mathbf{trace}\left(\mathbf{P}^{b\,2}\right) + \mathbf{trace}^2\left(\mathbf{P}^b\right)}{(N+2) \cdot \left[\mathbf{trace}\left(\mathbf{P}^{b\,2}\right) - \frac{\mathbf{trace}^2\left(\mathbf{P}^b\right)}{n}\right]}, 1\right). \tag{5.14}$$

An EnKF implementation which exploits the special structure of this estimator is the EnKF based on the RBLW estimator (EnKF-RBLW) wherein the posterior ensemble

can be built as follows (Nino-Ruiz and Sandu, 2015; Nino-Ruiz and Sandu):

$$\widehat{\mathbf{B}}_{RBLW} = \alpha^*_{RBLW} \cdot \mu \cdot \mathbf{I} + (1 - \alpha^*_{RBLW}) \cdot \mathbf{P}^b, \tag{5.15a}$$

$$\mathbf{X}^a_{RBLW} = \mathbf{X}^b + \widehat{\mathbf{B}}_{RBLW} \cdot \mathbf{H}^T \cdot \left[\mathbf{R} + \mathbf{H} \cdot \widehat{\mathbf{B}}_{RBLW} \cdot \mathbf{H}^T\right]^{-1} \cdot \mathbf{D}, \tag{5.15b}$$

$$\mu = \frac{\mathbf{trace}\left(\mathbf{P}^b\right)}{n}. \tag{5.15c}$$

Since numerical models can be highly non-linear, Gaussian assumptions on prior members are commonly broken. This assumption can be relaxed in the EnKF context by employing, for instance, the LW estimator for the estimation of background error covariance matrices during assimilation steps (Nino-Ruiz *et al.*, 2021). Besides, different prior structures can be treated in $\mathbf{T}$ to enrich the covariance matrix estimation, this is, to account for prior information about the dynamical system.

5.3. An Ensemble Kalman Filter Via Shrinkage Covariance Matrix Estimation and Prior Knowledge

In this Section, a novel EnKF implementation that incorporates prior knowledge of the background error covariance matrix in a practical manner to improve the DA process is presented. The method is based on a shrinkage estimator using a general target matrix. An efficient and totally parallelizable implementation of the method for high-dimensional systems is also proposed.

5.3.1. Filter Derivation

As was mentioned above, shrinkage based covariance matrix estimators which allow the use of a target matrix $\mathbf{T}$ to structure the covariance matrix, are limited to a target matrix with identity matrix structure (Nino-Ruiz and Sandu; Stoica *et al.*, 2008). Although matrix identity structure can reduce the spurious correlations caused by the ill-conditioned approximation of the error covariance matrix (Nino-Ruiz and Sandu, 2015; Ledoit and Wolf, 2004b; Chen and Prinn, 2006), the assumption of a covariance structure without correlation between the states is not always valid or desirable. Using a general target matrix enables the incorporation of prior information about the system into the error covariance matrix. This prior information can be information about the system physics as for instance, parameters, topography, transport phenomena and environmental information, or knowledge about the covariance structure coming from experts or previous experiments. A close formulation to calculate the weight value α using a general target matrix $\mathbf{T}_{KA}$ is proposed in (Stoica *et al.*, 2008; Zhu *et al.*, 2011),

$$\alpha_{KA} = \min\left(\frac{\frac{1}{N^2} \cdot \sum_{e=1}^{N} \left\|\Delta\mathbf{x}^{[e]}\right\|^4 - \frac{1}{N} \cdot \left\|\mathbf{P^b}\right\|^2}{\left\|\mathbf{P}^b - \mathbf{T}_{KA}\right\|^2}, 1\right). \tag{5.16a}$$

and the KA (Knowledge-Aided) estimator is obtained using (5.16a) in

$$\hat{\mathbf{B}}_{KA} = \alpha_{KA} \cdot \mathbf{T}_{KA} + (1 - \alpha_{KA}) \cdot \mathbf{P}^b \in \mathbb{R}^{n \times n}, \tag{5.16b}$$

It is important to note that no assumptions about the structure of $\mathbf{T}_{KA}$ are made to calculate α_{KA}. This approach can be seen as an extension of that in (Ledoit and Wolf, 2004b; Chen *et al.*, 2010) to a general target matrix and is usable for complex-value data case [1]. Similarly than the EnKF-RBLW an implementation of the EnKF can be obtained using the KA shrinkage-based estimator presented in (5.16):

$$\mathbf{X}^a = \mathbf{X}^b + \hat{\mathbf{B}}_{KA} \cdot \mathbf{H}^T \cdot [\mathbf{R} + \mathbf{H} \cdot \hat{\mathbf{B}}_{KA} \cdot \mathbf{H}^T] \cdot \mathbf{D},$$

Since the target matrix $\mathbf{T}_{KA}$ in the EnKF-KA is not necessary a matrix with identity structure, information about the dynamical system can be integrated into the data assimilation process. The prior information is directly related to the error covariance of the model states; this means that it is possible to integrate information of the system and guide the dynamical relation between the states and the relation between states and observations. Although there are no restrictions in the structure of $\mathbf{T}_{KA}$, it is important to remarks that $\mathbf{T}_{KA}$ is still a covariance matrix, so all the conditions related have to be accomplished. In Section 5.4 are shown examples of how to select $\mathbf{T}_{KA}$ properly.

5.3.2. Domain Localization

Both most popular concepts of localization can be applied in the EnKF-KA approach: covariance localization (Hamill *et al.*, 2001; Houtekamer and Mitchell, 2001), and local domain analysis (Ott *et al.*, 2004). We explore the implementation of local domain analysis due to the advantages not only in the spurious correlation mitigation but also in the implementations. Since the mean idea of the EnKF-KA is to incorporate prior information of the system in the DA framework, it is inherent that this information has to be saved and available in all the DA process. In high-dimensional applications, it is not convenient and, in some cases, prohibited to save a matrix of the dimension of $\mathbf{T}_{KA} \in \mathbb{R}^{n \times n}$, and calculate $\mathbf{P}^b \in \mathbb{R}^{n \times n}$ directly. It is here where the concept of local domains is crucial for the implementations of the EnKF-KA for high-dimensional systems. In local domains, a box of radius r of components around the state of interest is created, and just the states and observations within this box (local domain) are used in the analysis step (Ott *et al.*, 2004; Sakov *et al.*, 2010; Hunt *et al.*, 2007). This process is repeated for all the state components, doing multiple local analysis (in a smaller dimension) instead of a unique and global analysis (in a higher dimension). Another advantage of this implementation is that it facilitates the parallelization of the analysis since each local analysis can be performed in an independent core (Nino-Ruiz and Sandu; Greybush *et al.*, 2011). The implementation of the EnKF-KA using local domains analysis is summarized in the next steps:

1. A local domain of radius r is created for any model component. The $k-th$ local domain is formed by n_r ($n_r << n$) and m_r observation. The use

[1] The reader can consult (Stoica *et al.*, 2008; Zhu *et al.*, 2011) for additional information.

of domain decomposition is applied, so that boundary information is shared across neighboring domains. In this manner, we preserve the continuous dynamics of some physical variables such as Temperature, Wind Components, and Pressure. Figure 5.1 illustrates this strategy. The background ensemble and the analysis ensemble into the box is denoted by $\mathbf{X}_k^b \in \mathbb{R}^{n_r \times N}$ and $\mathbf{X}_k^a \in \mathbb{R}^{n_r \times 1}$ respectively. The covariance model error into the box are denoted by $\mathbf{B}_k \in \mathbb{R}^{n_r \times n_r}$, the local observation is denoted by $\mathbf{y}_k \in \mathbb{R}^{m_r \times 1}$ with observation covariance $\mathbf{R}_k \in \mathbb{R}^{m_r \times m_r}$, and the local innovation matrix is denoted by $\mathbf{D}_k \in \mathbb{R}^{m_r \times N}$.

2. Compute the local sample covariance matrix $\mathbf{P}_k^b \in \mathbb{R}^{n_r \times n_r}$

$$\Delta\mathbf{X}_k^b = \mathbf{X}_k^b - \bar{\mathbf{x}}_k^b \cdot \mathbf{1}_N^T, \tag{5.17a}$$

$$\mathbf{P}_k^b = \frac{1}{(N-1)} \cdot \Delta\mathbf{X}_k \cdot (\Delta\mathbf{X}_k)^T. \tag{5.17b}$$

3. Define the local target matrix $\mathbf{T}_k \in \mathbb{R}^{n_r \times n_r}$. On this step, the use of previous knowledge of the model dynamics is required. Knowledge is understood as the human-based experience in front of a large scale model used to represent reality. Large scale models for atmospheric dynamics, weather, water and ocean, reservoir modeling are used normally by experts in their fields. Even if the data to be assimilated is measured, some details and specifications are not captured on the model or included on it. Other possible causes are that due to the spatial-temporal resolution chosen for the numerical solution of the equations, it does not allow to capture intrinsic relationships between the states. We suggest a matrix $\mathbf{T}_k$ built on the basis of that specific knowledge. Although $\mathbf{T}_k$ must meet all requirements of a covariance matrix, the main contribution is that the matrix $\mathbf{T}_k$ must not fulfill any requirement about its structure and also can change dynamically.

4. Estimate the local error covariance $\mathbf{B}_k$ throw the KA shrinkage-based estimator $\hat{\mathbf{B}}_k$ using

$$\alpha_k = \min\left(\frac{\frac{1}{N^2} \cdot \sum_{e=1}^{N} \left\|\Delta\mathbf{x}_k^{[e]}\right\|^4 - \frac{1}{N} \cdot \left\|\mathbf{P}_k^b\right\|^2}{\left\|\mathbf{P}_k^b - \mathbf{T}_k\right\|^2}, 1\right), \tag{5.18a}$$

$$\hat{\mathbf{B}}_k = \alpha_k \cdot \mathbf{T}_k + (1-\alpha_k) \cdot \mathbf{P}_k^b \in \mathbb{R}^{n_r \times n_r}. \tag{5.18b}$$

5. Perform the local analysis step

$$\mathbf{X}_k^a = \mathbf{X}_k^b + \hat{\mathbf{B}}_k \cdot \mathbf{H}_k^T \cdot [\mathbf{R}_k + \mathbf{H}_k \cdot \hat{\mathbf{B}}_k \cdot \mathbf{H}_k^T] \cdot \mathbf{D}_k. \tag{5.19}$$

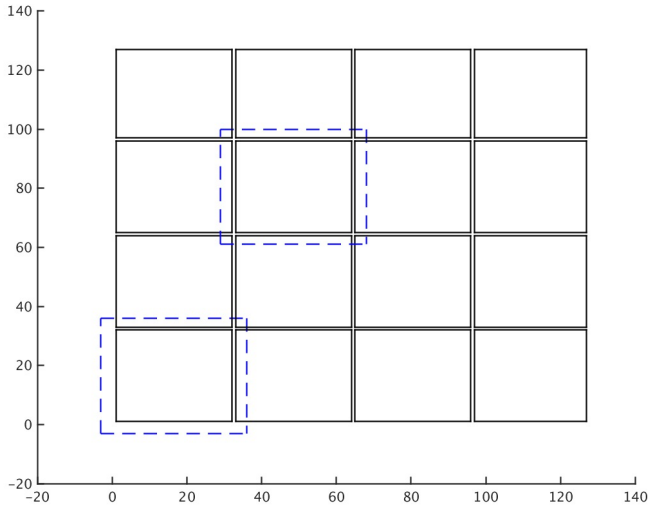


Figure 5.1: Domain decomposition is exploited to reduce the computational cost of our proposed method. Dashed regions denote the shared boundary information to be employed during assimilation steps

6. Once all the local analyses are performed, map those to the global domain. The global analysis state is then obtained. This does not mean to perform a new global analysis. In (Ott *et al.*, 2004) two map approaches are proposed. The first one uses only the analysis results at the center point of each local region to form the global analysis vectors. The second one uses the average of all the local analysis where a grid cell is involved in obtaining the global analysis.

Note that with a correct selection of r, the matrix computations in each local domain are inexpensive, so eq. (5.18) can be computed efficiently for high-dimensional systems.

5.3.3. Inflation Aspects

In the context of EnKF-KA, the covariance inflation can be efficiently performed increasing the dispersion of matrix (5.8) by a inflation factor β_{inf} :

$$\widehat{\Delta\mathbf{X}} = \beta_{\text{inf}} \cdot \Delta\mathbf{X} \in \mathbb{R}^{n\times N}, \tag{5.20}$$

and by noting that:

$$\mathbf{tr}\left(\beta_{\text{inf}}^2 \cdot \mathbf{P}^b\right) = \beta_{\text{inf}}^2 \cdot \mathbf{tr}\left(\mathbf{P}^b\right),$$

where **tr** represent the trace of the matrix. For instance, we can see that covariance inflation on the optimal factor (5.16a) reads:

$$\alpha_{KA}^{\text{inf}} = \min\left(\frac{\frac{1}{N^2}\cdot\sum_{e=1}^{N}\beta_{\text{inf}}^{8}\cdot\left\|\Delta\mathbf{x}^{[e]}\right\|^4 - \frac{1}{N}\cdot\beta_{\text{inf}}^{2}\cdot\left\|\mathbf{P}^b\right\|^2}{\left\|\beta_{\text{inf}}^{2}\cdot\mathbf{P}^b - \mathbf{T}_{KA}\right\|^2}, 1\right).$$

5.4. Experimental Settings

5.4.1. Results with an Advection Diffusion Model

This section illustrates the proposed EnKF-KA over simple a advection-diffusion process. The advection-diffusion governs the changes of a conservative property such as the concentration in a fluid environmental (Richardson and Mooney, 1975; Tirabassi, 1989). The advection-diffusion equation has been used as a simple model to study the behavior and transport of pollutant in the atmosphere. In two dimensions, the horizontal changes in the concentration of a determinate pollutant **C** in the atmosphere can be approximated as:

$$\frac{\partial C}{\partial t} = D_x\frac{\partial^2 C}{\partial x^2} - v_x\frac{\partial C}{\partial x} + D_y\frac{\partial^2 C}{\partial y^2} - v_y\frac{\partial C}{\partial y} + E(t), \tag{5.21}$$

where v_x and u_y are the north-south and west-east wind velocities respectively, D_x and D_y are the north-south and west-east diffusion coefficients respectively, and $E(t)$ are the emissions. The experimental settings are:

- The continuous advection-diffusion equation is discretizated in a 20×20 domain, obtaining a total of $n = 400$ states representing concentration in each cell.
- The boundary condition used for solving the experiment was the Dirichlet homogeneous zero or null value fixed in the contour.
- Ten emissions points are considered. Additionally, to represent a real scenario where the emissions are the most important uncertainty sources in the atmosphere chemistry modelation (Barbu *et al.*, 2009), uncertainty in every time in the emissions are considered.
- There is no considered uncertainty in initial conditions, boundary conditions, or parameter values.
- With the idea of simulating an imperfect representation of the model environment, an artificial valley is performed in the real scenario, where the true state $\mathbf{x}^*$ and observations $\mathbf{y}$ are taken. The artificial valley is created, increasing the diffusion coefficients and reducing the velocity winds components in a determinate number of cells. This implies that the interchange of pollutants between two locations, one inside and the other outside the valley, is considerably lower than two locations outside or inside the valley. The valley is not included in the model used for assimilation purposes. A graphical representation is shown in Figure 5.2.

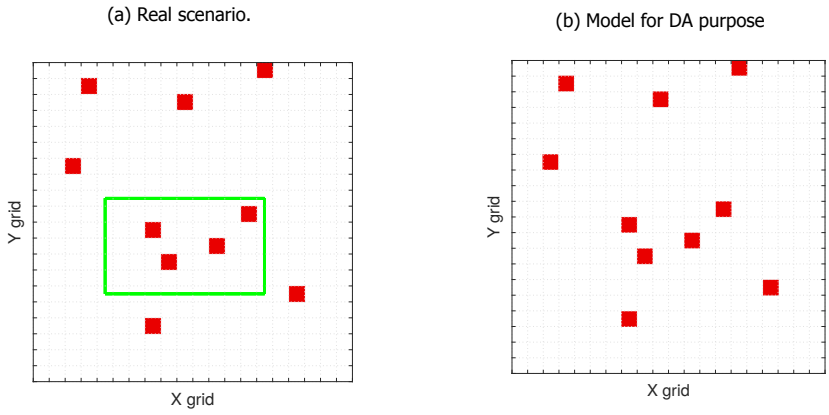


Figure 5.2: Comparison of the real scenario vs the model scenario. The green line represents the artificial valley. The red squares represent the emission points.

- A background ensemble is built perturbing the 10 emission points by drawing a sample from the Normal distribution,

$$\mathbf{x}^{b[e]} \sim \mathcal{N}(\overline{\mathbf{x}}^{b}, \rho_b^2 \cdot \mathbf{I}), \text{ for } 1 \leq e \leq N, \tag{5.22}$$

  where $\rho_b = 0.05$

- We propose three ensemble sizes for the experiments $N \in \{10, 50, 100\}$.

- The assimilation window consists of $M = 1000$ time steps. Two observation periods are proposed for the test, each time step and each ten time steps. We denote by $\delta t \in \{1, 10\}$ the elapsed time between two observations.

- The error statistics are associated with the Gaussian distribution,

$$\mathbf{y}_\ell \sim \mathcal{N}(\mathcal{H}_\ell(\mathbf{x}_\ell^*), \rho_o^2 \cdot \mathbf{I}), \text{ for } 1 \leq \ell \leq M, \tag{5.23}$$

  where $\rho_o = 0.001$.

- We consider two fractions of observed components $s \in \{0.12, 0.5\}$. The components are randomly chosen at each assimilation step.

- The L_2 norm of errors are utilized as a measure of accuracy at the assimilation step ℓ,

$$L_\ell = \sqrt{\left[\mathbf{x}_\ell^a - \mathbf{x}_\ell^*\right]^T \cdot \left[\mathbf{x}_\ell^a - \mathbf{x}_\ell^*\right]}, \tag{5.24}$$

  where $\mathbf{x}_\ell^*$ and $\mathbf{x}_\ell^a$ are the reference and the analysis solution respectively.

- The Root-Mean-Square-Error (RMSE) is used as a measure of performance, in average, on a given assimilation window,

$$\text{RMSE} = \frac{1}{M} \cdot \sum_{\ell=1}^{M} \lambda_\ell^2, \tag{5.25a}$$

where

$$\lambda_\ell = \left\| \mathbf{x}_\ell^a - \mathbf{x}_\ell^* \right\|_2 . \tag{5.25b}$$

- The percentage of non converge experiments (PNCE) is calculated for all the scenarios.

The idea is to incorporate the physical restrictions that the model does not capture, for this case, the artificial valley, via the EnKF-KA. If we use a standard distance-based localization for a state into the valley to cut the coming information from distant observations, the process will include both observations inside and outside the valley. With the EnKF-KA, we try to cut observations that are outside the valley, even if there are at the same distance, as is represented in Figure 5.3.

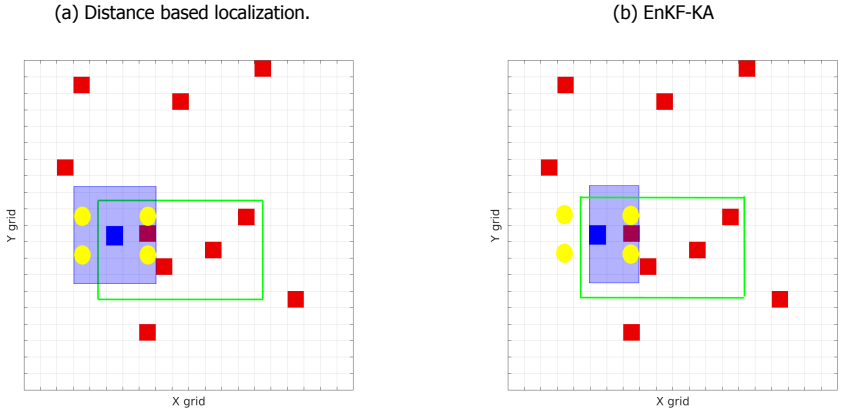


Figure 5.3: Comparison of the distance based localization approach vs the EnKF-KA. In the EnKF-KA the influence region is based on the distance and on the information about the system. The blue square represents the analysed state, the blue shadow the influence region, and the yellow circles represent the observations.

This is achieved by incorporating the physical restrictions (the topography of the interest domain) into the covariance estimation throw the target matrix $\mathbf{T}_{KA}$. The target matrix is built starting from a Gaspari-Cohn function (Gaspari and Cohn, 1999) and reducing to zero the covariance between the states inside and outside the valley. After this process, it is very important to test whether the final $\mathbf{T}_{KA}$ is still a positive semidefinite matrix. Note that the final covariance between the state inside and outside the valley will not be necessary zero because the final covariance matrix is a convex combination of $\mathbf{T}_{KA}$ and $\mathbf{P}^b$. In Figure 5.4 is shown an example of a $\mathbf{T}_{KA}$ matrix obtained using the proposed process for an influence radius $r = 4$.

The performance of the EnKF-KA is compared with the shrinkage-based EnKF-RBLW and the standard EnKF using covariance localization EnKF-CL with $r = 1$ (other influence radius were tested, but $r = 1$ presents the best performance) under

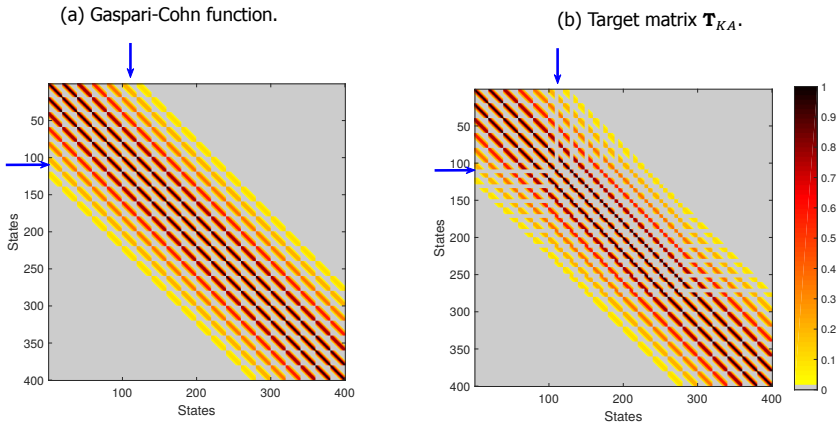


Figure 5.4: Graphical representation of the $\mathbf{T}_{KA}$ matrix. The arrows remark the state 110 which is located just in the inside border of the valley (represented as a blue square in Figure 5.3), and shows how the covariance between a states inside and the states outside the valley is fixed in 0.

the experimental setup presented below. A total of 20 experiments are performed for each scenario. The target matrix $\mathbf{T}_{KA}$ is built from a Gaspari-Cohn with $r = 1$ and following the mentioned process including physical restrictions of the valley. The magnitude of $\mathbf{T}_{KA}$ is computed according with the average of the trace of $\mathbf{P}^b$. In Figure 5.5 is shown the dynamical evolution of the L_2 norm for different scenarios. Figure 5.6 presents the values of the average RMSE for all the experiment scenarios and the PNCE for the EnKF-CL for the different ensemble members value. For the EnKF-RBLW and the EnKF-KA the $PNCE = 0\%$ for all the cases.

As is shown in the figures 5.5 and 5.6, the EnKF-KA presents lower error rates than the EnKF-RBLW and the EnKF-CL in almost all the scenarios. This shows how the integration of the physical restrictions can help the data assimilation process. It is interesting to evaluate the scenarios with a smaller number of ensemble members, where the differences among the three algorithms are more considerable. The RMSE value of the EnKF-KA in these scenarios is much lower than the EnKF-CL, showing that shrinkage-based estimators are more robust than the sample covariance matrix when $n >> N$. Since the ensemble statistics approximate the mean and the covariance of the state, the ensemble spread should describe the system uncertainty (Timmermans *et al.*, 2019; Houtekamer and Zhang, 2016). The ensemble spread should be as small as possible, reducing the estimation uncertainty, but enough to conserve the filter stability(Vrugt and Robinson, 2007). If the filter estimates the state uncertainty correctly, the ensemble spread should match with the RMSE when there are no model errors (Nan and Wu, 2011). Figure 5.7 shows the ensemble spread of each algorithm among assimilation steps for a specific experiment. It can be seen how all the algorithms reduce the ensemble spread after few assimilation steps, reducing the system uncertainty levels. The Free-Run keep similar uncertainty values among time because no new information is incorporated. Finally, the EnKF-KA presents the lowest spread values matching with the lowest

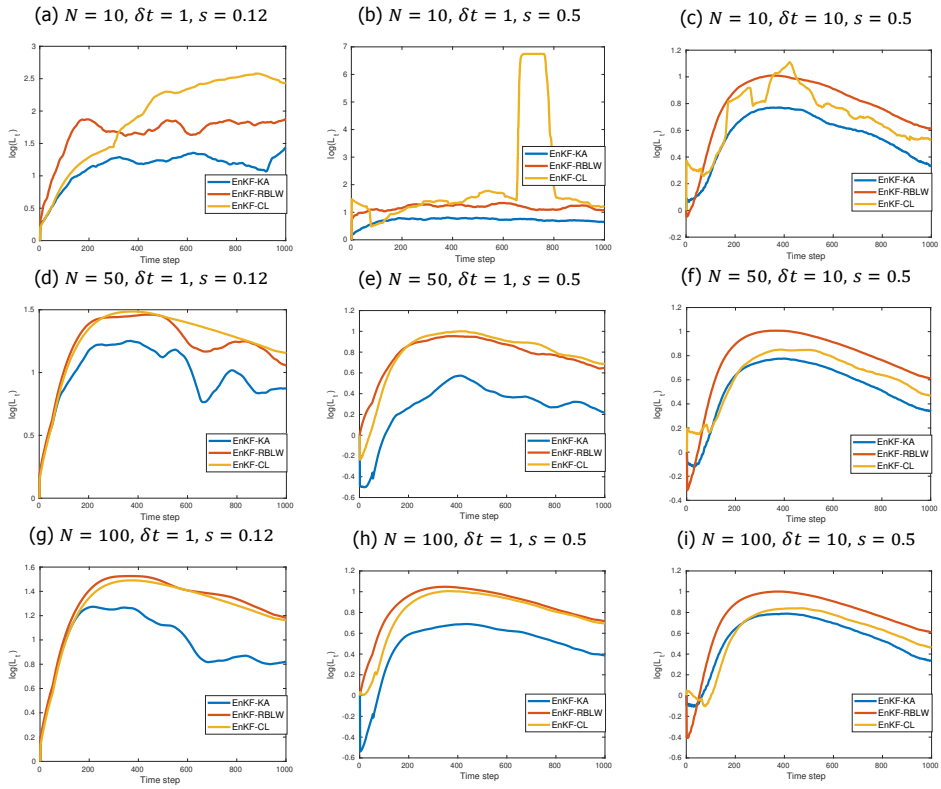


Figure 5.5: Comparisson of the performance among the EnKF-KA, EnKF-RBLW and EnKF-CL for some scenarios.

RMSE values, which means that the ENKF-KA can correctly reproduce the system uncertainty and improve estimation accuracy.

In Figure 5.8 is presented the time evolution of states in four different spacial location for one experiment scenario. It is evident that the EnKF-KA reproduces more accurately locations in the border of the artificial valley than the other methods, showing the effect of the incorporated information throw $\mathbf{T}_{KA}$.

An aspect that is important to remarks is the value of α_{KA} for different ensemble member values. The mean α_{KA} value for ensemble number of $N = 10$, $N = 50$ and $N = 100$ are $\bar{\alpha_{10}} = 0.698$, $\bar{\alpha_{50}} = 0.591$ and $\bar{\alpha_{100}} = 0.508$. With a small number of ensemble members the assumption of a poor estimation of the covariance throw the sample covariance matrix produces a higher value of α_{KA}, giving more weight to the target matrix than when the number of ensemble, and the quality of the estimation throw the sample covariance matrix, is higher.

5.4.2. Results with an Atmospheric General Circulation Model

SPEEDY (Simplified Parameterizations, privitivE-Equation DYnamics) is an Atmospheric General Circulation model (Bracco *et al.*, 2004; Miyoshi, 2011), which help

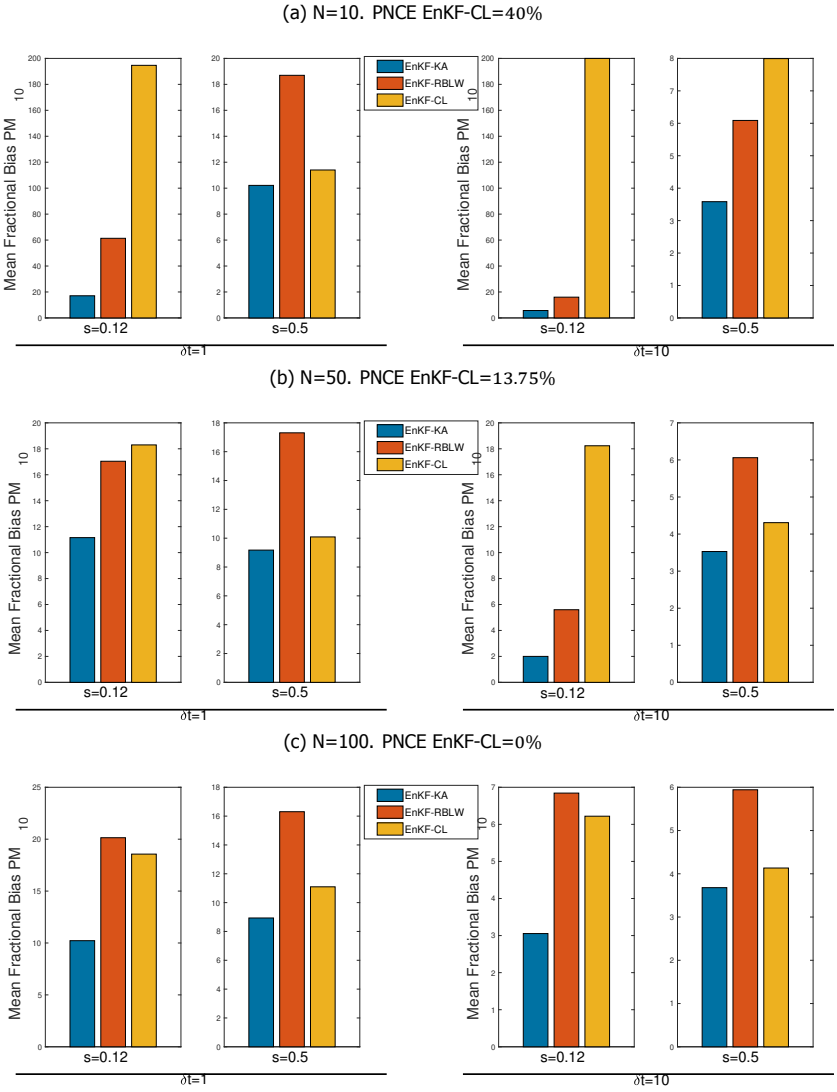


Figure 5.6: Comparisson performance for the different algorithms

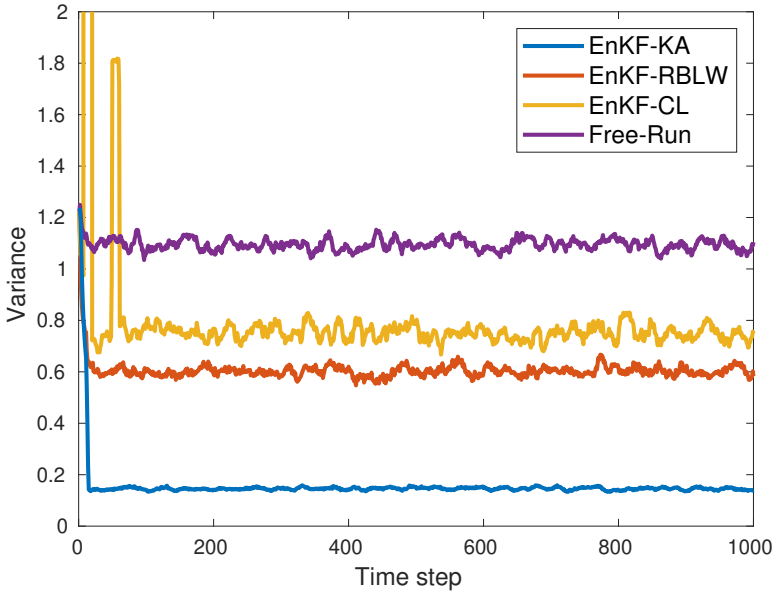


Figure 5.7: Ensemble spread for different algorithms. The graph corresponds with one experiment wit $N = 50$, $\delta t = 10$, and $s = 0.5$

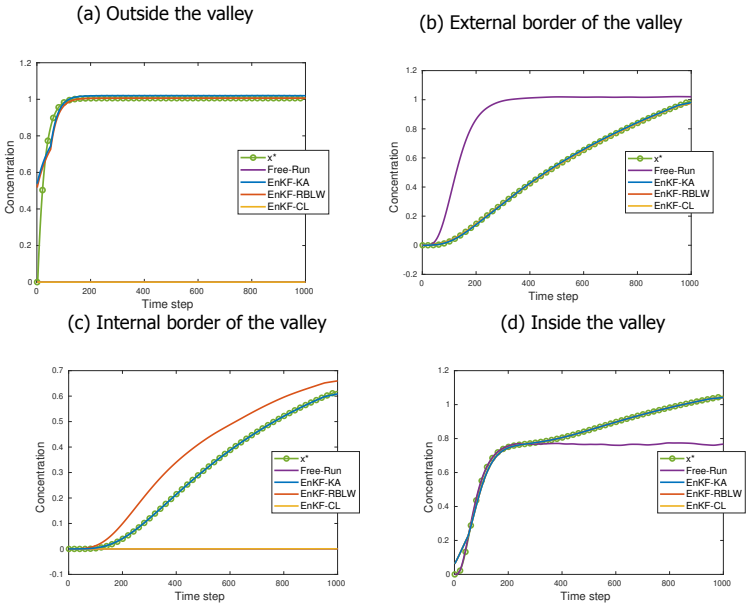


Figure 5.8: Time evolution of concentration for different locations. The graph corresponds with one experiment wit $N = 50$, $\delta t = 10$, and $s = 0.5$

us to study the performance of the EnKF-KA method in highly non-linear model scenario. The model consists of seven numerical layer, and at each one, a T-30 model resolution is employed (96×48 grid components) (Molteni, 2003; Kucharski *et al.*, 2006). The total number of physical variables at each numerical grid point is five, these are: the temperature T (K), the zonal u and the meridional v wind components (m/s), the specific humidity Q (g/kg), and the pressure ρ (hPa). We employ all physical variables into our data assimilation process. Note that, the model dimension in our settings reads $n = 133,632$. During our experiments, we consider ensemble sizes of $N = 10$ and $N = 20$, this applies for all numerical scenarios. Note that, model resolutions are $13,632$ and $6,685$ times larger than ensemble sizes ($n \gg N$), which takes to current DA operational settings. We follow the experimental settings presented in (Nino-Ruiz *et al.*, 2021; Kalnay *et al.*, 2007):

- Long term numerical integrations are applied to build the reference solution as well as the initial background ensemble (two years of a numerical simulation). We start with a system in equilibrium, and after adding a small perturbation, the numerical integration is performed.
- The experiments do not account for model errors.
- Standard deviations of observational errors are detailed in Table 5.1.
- We employ a highly sparse observational network. The observation coverage is 9% of the spatial resolution. This linear observation operator is shown in Figure 5.9. Note that this is an irregularly distributed, realistic observational network.

Model Variable	**Observational Error Standard Deviation**
Zonal Wind Component (u)	$1\ m/s$
Meridional Wind Component (v)	$1\ m/s$
Temperature (T)	$1\ (K)$
Specific humidity (q)	$0.0001\ (kg/kg)$
Surface pressure (ρ)	$100\ (Pa)$

Table 5.1: Observational error standard deviation.

- The inflation factor is $\beta_{\text{inf}} = 1.3$ for all experiments.
- We set up a total simulation time of two months with observations frequencies about 6 and 12 hours. We expect the non-linear dynamics of the SPEEDY model to impact the quality of analysis states as the observation frequency decreases.
- The Root-Mean-Square-Error (RMSE) is employed as a metric of accuracy for a given analysis $\mathbf{x}_\ell^a$ and a reference solution $\mathbf{x}_\ell^*$ (see eq. 5.25).

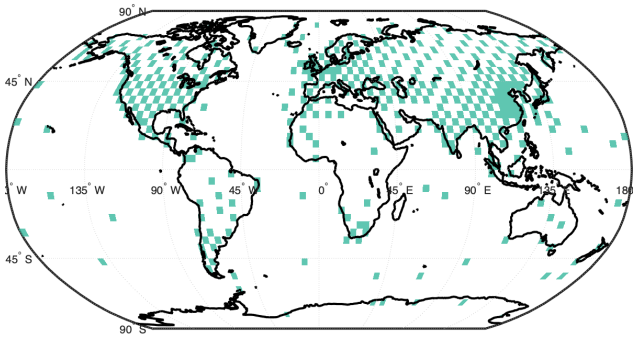


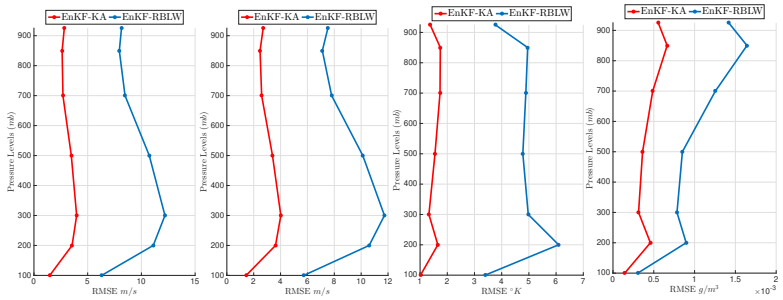
Figure 5.9: An irregularly distributed realistic observational network. 415 stations (9 % of all grid points) are located mostly over continents in the northern hemisphere.

5.4.3. Analysis Errors across Pressure Levels

Figures 5.10 and 5.11 show us the behavior of the proposed method against to EnKF-RBLW. The analysis was made using the RMSE metric for observation frequencies of 6 and 12 hours for u, v, and T model variables in different Pressure Levels. The numerical results show that EnKF-KA can be more accurate than EnKF-RBLW, this obeys the fact that the error correlations are driven by the physics and the numerical model's non-linear dynamics. Therefore, the underlying error distribution of wind components can be non-Gaussian as the frequency of observations decreases (long-term forecasts). This can apply to temperature fields as well. On the other hand, Gaussian assumptions can be valid for model variables such as the specific humidity. For this model variable, slight differences between analysis RMSE can be evidenced for the compared filter implementations. This can be expected since the RBLW covariance matrix estimator can perform well as the underlying error distribution of ensemble members is nearly Gaussian. Nevertheless, these small differences favor the proposed EnKF-KA formulation under the current experimental settings. In general, errors can grow faster across all pressure levels in model variables such as u, v, and T than those in variables that tend to preserve Gaussianity among assimilation steps (i.e., q).

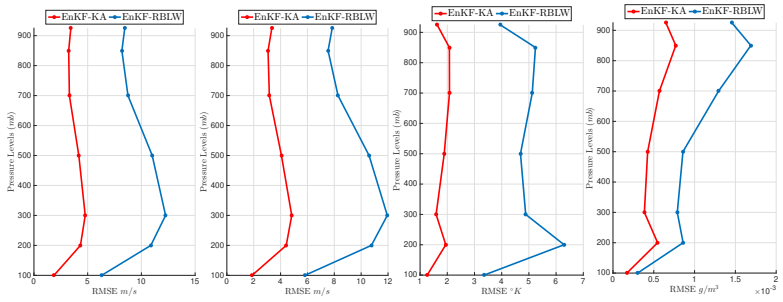
5.4.4. Evolution of Analysis Errors among Assimilation Steps

As we can see in figures 5.12 and 5.13, the initial errors decrease as observations are assimilated in each analysis step using the proposed method, for observation frequencies of 6 and 12 hours. It should be noted that the observation frequency affects the estimation quality but not the convergence of the EnKF-KA with the configuration of this experiment. On the other hand, the proposed method can outperform the EnKF-RBLW formulation as shown in the figures 5.12 and 5.13.



(a) u (b) v (c) T (d) q

Figure 5.10: Analysis RMSE at the all pressure levels temporally averaged for one month and a half after the initial spin-up period of two weeks. The number of ensemble members reads $N = 10$. The errors per layer are shown for observation frequencies of 6 h.



(a) u (b) v (c) T (d) q

Figure 5.11: Analysis RMSE at the all pressure levels temporally averaged for one month and a half after the initial spin-up period of two weeks. The number of ensemble members reads $N = 10$. The errors per layer are shown for observation frequencies of 12 h.

The fact that accurate analysis states can be estimated despite a highly sparse observational network shows that the dynamic system's background error correlations have been captured into the covariance matrix estimators.

5.4.5. Analysis RMSE for the Assimilation Window

Tables 5.2 and 5.3 shows the analysis RMSE of the EnKF-KA and the EnKF-RBLW using 6 and 12 hours for observation frequencies and ensemble sizes of $N = 10$ and $N = 20$. The RMSE values are computed for 60 days with an initial spin-up period of ten days. As can be seen, the analysis states of the EnKF-KA can improve on the results proposed by the EnKF-RBLW. This can be possible due to EnKF-KA uses a target matrix different from the identity matrix (used in EnKF-RBLW), and the EnKF-RBLW is performed under Gaussian assumptions over prior ensemble members. However, Gaussian assumptions on background errors can be broken by the numerical model's non-linear dynamics. The observational network in the experiment is sparse, about 9% of observations, which means that posterior

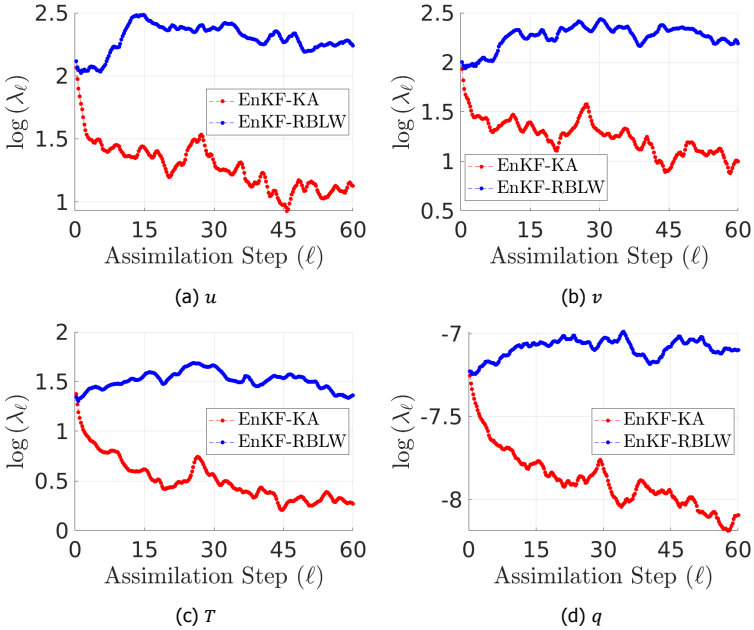


Figure 5.12: Evolution of analysis errors among assimilation steps for $N = 10$ and an observation frequency of 6 h. The l_2-norm of errors is displayed in the log-scale for ease of reading.

estimates' quality relies on background error correlations. The proposed method then improves the quality of the analysis results over the compared filter for sparse observational network and very small ensemble sizes in the experiment.

5.4.6. Uncertainty analysis

For sequential data assimilation based on Ensemble Kalman Filter is known that if the ensemble spread becomes very small or becomes very large, the filter falls into divergence, but also, the ensemble spread can be used to explore the uncertainty associated with the initial condition and the uncertainty associated to the formulation of the prediction model. Figure 5.14 shows the mean of ensemble variance among assimilation steps for u, v, T and Q variables in pressure level of 500 Pa. As expected, the ensemble variance decreases as EnK-KA is used for the analysis step. This means that the uncertainty decreases as the observations are assimilated. It should be noted that a covariance inflation factor of 1.3 was used in the experiment. In the same way, Figure 5.15 shows samples of the components taken for each of the physical variables of the model, it is possible to see how the differences between ensemble members decrease through the assimilation steps.

5.4.7. CPU-Time of Analysis Steps

Statistics of CPU-Times are computed across all analysis steps for both filter implementations. The reported times are shown in Table 5.4, where the average and

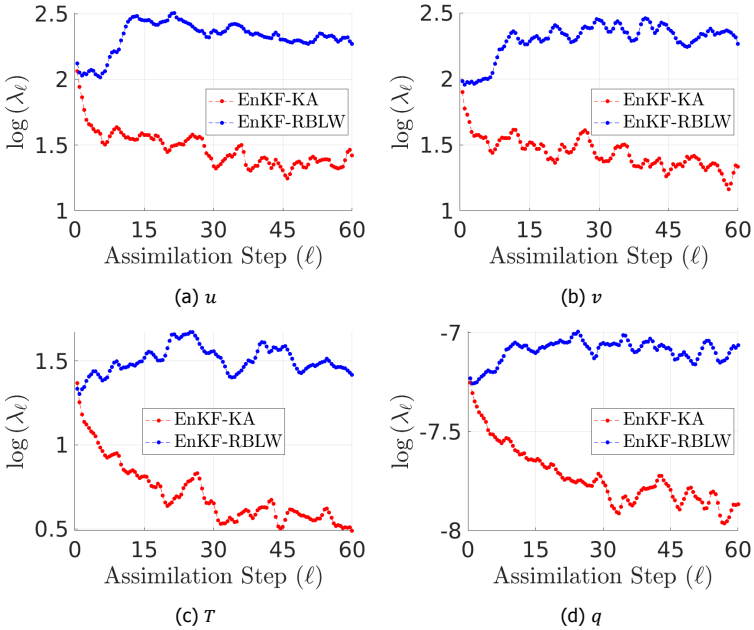


Figure 5.13: Evolution of analysis errors among assimilation steps for $N = 10$ and an observation frequency of 12 h. The l_2-norm of errors is displayed in the log-scale for ease of reading.

the variance of elapsed time for the analysis step computations are in seconds. The forecast step was realized using parallelism in a CPU with four cores; this means that up to four ensembles were forecast simultaneously.

Variable	Method	6 hours	12 hours
$u\ (m/s)$	EnKF-KA	3.682	4.359
	EnKF-RBLW	10.734	11.007
$v\ (m/s)$	EnKF-KA	3.619	4.284
	EnKF-RBLW	10.122	10.589
$T\ (K)$	EnKF-KA	1.655	2.001
	EnKF-RBLW	4.777	4.691
$Q\ (kg/kg)$	EnKF-KA	0.003	0.004
	EnKF-RBLW	0.008	0.008
$\rho\ (hPa)$	EnKF-KA	3.186	3.875
	EnKF-RBLW	10.330	11.110

Table 5.2: RMSE values in time for observation frequencies of 6 h and 12 h. As the frequency of observations is decreased, the EnKF-KA formulation can improve on the results of the EnKF-RBLW method. The number of ensemble members reads $N = 10$.

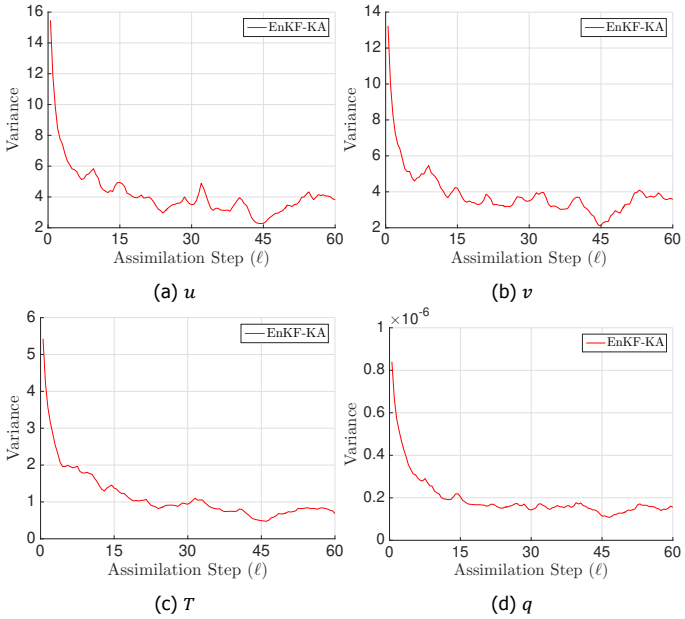
5

Variable	Method	6 hours	12 hours
$u\ (m/s)$	EnKF-KA	3.334	4.324
	EnKF-RBLW	10.232	10.595
$v\ (m/s)$	EnKF-KA	3.267	4.204
	EnKF-RBLW	10.278	10.085
$T\ (K)$	EnKF-KA	1.52	1.976
	EnKF-RBLW	4.391	4.431
$Q\ (kg/kg)$	EnKF-KA	0.003	0.004
	EnKF-RBLW	0.008	0.008
$\rho\ (hPa)$	EnKF-KA	2.863	3.848
	EnKF-RBLW	9.864	10.102

Table 5.3: RMSE values in time for observation frequencies of 6 h and 12 h. As the frequency of observations is decreased, the EnKF-KA formulation can improve on the results of the EnKF-RBWL method. The number of ensemble members reads $N = 20$.

Method	Average CPU-Time	Stand. Dev. CPU-Time
Analysis EnKF-KA	6.468	0.172
Analysis EnKF-RBLW	5.562	0.157
Forecast Step	4.327	0.209

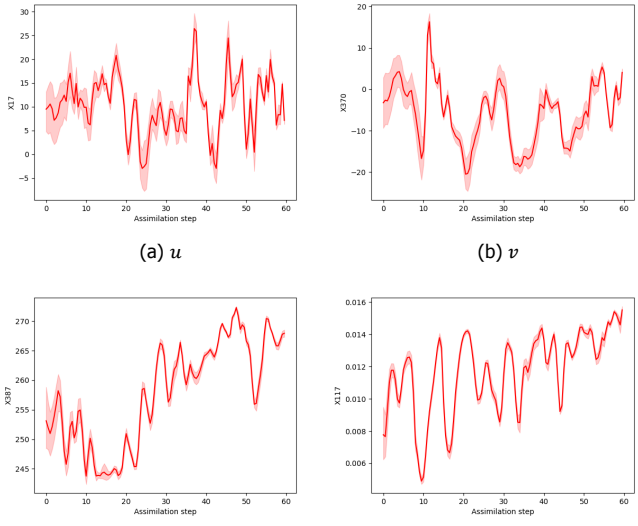
Table 5.4: Statistics of CPU-Time in seconds for the analysis steps of the compared filter implementations and the forecast step. The number of ensemble members reads 10.



(a) u (b) v

(c) T (d) q

Figure 5.14: Mean of variance among assimilation steps for $N = 10$, an observation frequency of 12 h and a pressure level of 500 Pa.



(a) u (b) v

(c) T (d) q

Figure 5.15: Ensembles of a component sample among assimilation steps for $N = 10$, an observation frequency of 12 h and a pressure level of 500 Pa

5.5. Conclusions

An efficient and practical implementation of the EnKF based on shrinkage covariance matrix estimation (EnKF-KA) was proposed in the present chapter. The proposed filter implementation exploits the information brought by an ensemble of model realization (numerical model dynamics) and our prior knowledge about the actual dynamical system (i.e., the prior structure of background error correlations). The EnKF-KA uses a target matrix with a general structure, representing a novel approach compared with the current shrinkage-based estimators that use an identity matrix as a target matrix. An efficient implementation for large systems is presented, taking advantage of the local domain decomposition. Experimental tests are performed by using an advection-diffusion model and an Atmospheric General Circulation Model. In both cases, the proposed method can outperform EnKF based on shrinkage covariance estimation where there is no prior information about error correlations, and the standard EnKF using covariance localization. The results support the idea that it is possible to use the information and prior knowledge of the system to improve the current ensemble-based DA method.

References

O. L. Quintero M, G. Scaglia, F. di Sciascio, and V. Mut, *Numerical methods based strategy and particle filter state estimation for bio process control,* in *2008 IEEE International Conference on Industrial Technology* (IEEE, 2008) pp. 1–6.

O. L. Quintero M, A. A. Amicarelli, G. Scaglia, and F. di Sciascio, *Control based on numerical methods and recursive bayesian estimation in a continuous alcoholic fermentation process,* BioResources **4**, 1372 (2009).

G. Evensen, *The Ensemble Kalman Filter: Theoretical formulation and practical implementation,* Ocean Dynamics **53**, 343 (2003).

J. L. Anderson and S. L. Anderson, *A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts,* Monthly Weather Review **127**, 2741 (1999).

E. Ott, B. R. Hunt, I. Szunyogh, A. V. Zimin, E. Kostelich, M. Corazza, E. Kalnay, D. Patil, and J. A. Yorke, *A local ensemble Kalman filter for atmospheric data assimilation,* Tellus **56**, 415 (2004).

J. L. Anderson, *An Ensemble Adjustment Kalman Filter for Data Assimilation,* Monthly Weather Review **129**, 2884 (2001).

E. D. Nino-Ruiz, L. G. Guzman-Reyes, and R. Beltran-Arrieta, *An adjoint-free four-dimensional variational data assimilation method via a modified cholesky decomposition and an iterative woodbury matrix formula,* Nonlinear Dynamics **99**, 2441 (2020).

E. D. Nino-Ruiz and A. Sandu, *Ensemble kalman filter implementations based on shrinkage covariance matrix estimation,* Ocean Dynamics **65**, 1423 (2015).

E. D. Nino-Ruiz and A. Sandu, Cluster Computing **22**, 2211.

X. Wang, D. M. Barker, C. Snyder, and T. M. Hamill, *A hybrid etkf–3dvar data assimilation scheme for the wrf model. part i: Observing system simulation experiment,* Monthly Weather Review **136**, 5116 (2008).

X. Wang, C. Snyder, and T. M. Hamill, *On the theoretical equivalence of differently proposed ensemble–3dvar hybrid analysis schemes,* Monthly Weather Review **135**, 222 (2007).

G. Fu, F. Prata, H. Xiang Lin, A. Heemink, A. Segers, and S. Lu, *Data assimilation for volcanic ash plumes using a satellite observational operator: A case study on the 2010 Eyjafjallajökull volcanic eruption,* Atmospheric Chemistry and Physics **17**, 1187 (2017).

S. Lu, H. X. Lin, A. W. Heemink, G. Fu, and A. J. Segers, *Estimation of Volcanic Ash Emissions Using Trajectory-Based 4D-Var Data Assimilation,* Monthly Weather Review **144**, 575 (2016).

Y. Zhu, Z. Toth, R. Wobus, D. Richardson, and K. Mylne, *THE ECONOMIC VALUE OF ENSEMBLE-BASED WEATHER FORECASTS,* Bulletin of the American Meteorological Society **83**, 73 (2002).

A. Touloumis, *Nonparametric stein-type shrinkage covariance matrix estimators in high-dimensional settings,* Computational Statistics & Data Analysis **83**, 251 (2015).

R. Couillet and M. McKay, *Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators,* Journal of Multivariate Analysis **131**, 99 (2014).

O. Ledoit, M. Wolf, *et al.*, *Optimal estimation of a large-dimensional covariance matrix under stein's loss,* Bernoulli **24**, 3791 (2018).

O. Ledoit and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices,* Journal of multivariate analysis **88**, 365 (2004a).

E. D. Nino-Ruiz, L. Guzman, and D. Jabba, *An ensemble kalman filter implementation based on the ledoit and wolf covariance matrix estimator,* Journal of Computational and Applied Mathematics **384**, 113163 (2021).

P. Stoica, J. Li, X. Zhu, and J. R. Guerci, *On using a priori knowledge in space-time adaptive processing,* IEEE Transactions on Signal Processing **56**, 2598 (2008).

O. Ledoit and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices,* Journal of Multivariate Analysis **88**, 365 (2004b).

Y. H. Chen and R. G. Prinn, *Estimation of atmospheric methane emissions between 1996 and 2001 using a three-dimensional global chemical transport model,* Journal of Geophysical Research Atmospheres **111**, 1 (2006).

X. Zhu, J. Li, and P. Stoica, *Knowledge-Aided Space-Time Adaptive Processing,* IEEE Transaction on Aerospace And Electronic Systems **47**, 1325 (2011).

Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, *Shrinkage algorithms for MMSE covariance estimation,* IEEE Transactions on Signal Processing **58**, 5016 (2010), 0907.4698 .

T. M. Hamill, J. S. Whitaker, and C. Snyder, *Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter,* Monthly Weather Review (2001), 10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2.

P. Houtekamer and H. Mitchell, *A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation,* American Meteorological Society **129**, 123 (2001), arXiv:0203058 [physics] .

P. Sakov, G. Evensen, and L. Bertino, *Asynchronous data assimilation with the EnKF,* Tellus, Series A: Dynamic Meteorology and Oceanography **62**, 24 (2010).

B. R. Hunt, E. J. Kostelich, and I. Szunyogh, *Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter,* Physica D: Nonlinear Phenomena **230**, 112 (2007).

S. J. Greybush, E. Kalnay, T. Miyoshi, K. Ide, and B. R. Hunt, *Balance and Ensemble Kalman Filter Localization Techniques,* Monthly Weather Review **139**, 511 (2011).

P. L. Richardson and K. Mooney, *The mediterranean outflow—a simple advection-diffusion model,* Journal of Physical Oceanography **5**, 476 (1975).

T. Tirabassi, *Analytical air pollution advection and diffusion models,* Water, Air, and Soil Pollution **47**, 19 (1989).

A. L. Barbu, A. J. Segers, M. Schaap, A. W. Heemink, and P. J. H. Builtjes, *A multicomponent data assimilation experiment directed to sulphur dioxide and sulphate over Europe,* Atmospheric Environment **43**, 1622 (2009).

G. Gaspari and S. E. Cohn, *Construction of correlation functions in two and three dimensions,* Quarterly Journal of the Royal Meteorological Society **125**, 723 (1999).

R. Timmermans, A. Segers, L. Curier, R. Abida, J. L. Attié, L. El Amraoui, H. Eskes, J. De Haan, J. Kujanpää, W. Lahoz, A. Oude Nijhuis, S. Quesada-Ruiz, P. Ricaud, P. Veefkind, and M. Schaap, *Impact of synthetic space-borne NO2 observations from the Sentinel-4 and Sentinel-5P missions on tropospheric NO2 analyses,* Atmospheric Chemistry and Physics **19**, 12811 (2019).

P. L. Houtekamer and F. Zhang, *Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation,* Monthly Weather Review (2016), 10.1175/MWR-D-15-0440.1.

J. A. Vrugt and B. A. Robinson, *Treatment of uncertainty using ensemble methods : Comparison of sequential data assimilation and Bayesian model averaging,* Water Resources Manager **43**, 1 (2007).

T. Nan and J. Wu, *Groundwater parameter estimation using the ensemble Kalman filter with localization,* Hydrogeology Journal **19**, 547 (2011).

A. Bracco, F. Kucharski, R. Kallummal, and F. Molteni, *Internal variability, external forcing and climate trends in multi-decadal AGCM ensembles,* Climate Dynamics **23**, 659 (2004).

T. Miyoshi, *The gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform kalman filter,* Monthly Weather Review **139**, 1519 (2011).

F. Molteni, *Atmospheric simulations using a gcm with simplified physical parametrizations. i: Model climatology and variability in multi-decadal experiments,* Climate Dynamics **20**, 175 (2003).

F. Kucharski, F. Molteni, and A. Bracco, *Decadal interactions between the western tropical pacific and the north atlantic oscillation,* Climate dynamics **26**, 79 (2006).

E. Kalnay, H. Li, T. Miyoshi, S.-C. Yang, and J. Ballabrera-Poy, *4dvar or ensemble kalman filter?* Tellus A: Dynamic Meteorology and Oceanography **59**, 758 (2007).

6

A Robust Ensemble-based Data Assimilation Method using Shrinkage Estimator and Adaptive Inflation

This chapter proposes a robust and non-gaussian version of the shrinkage-based EnKF implementation, the EnKF-KA. The proposed method is based on the robust H_∞ filter and on its ensemble time-local version the EnTLHF, using an adaptive inflation factor depending on the shrinkage covariance estimated matrix. This implies a theoretical and solid background to construct robust filters from the well-known covariance inflation technique. The method is tested using the Lorenz-96 model to evaluate the robustness and performance under different scenarios as ensemble size, observation error, errors in the model specifications, and ensemble gaussianity. The results suggest good robustness of the proposed method in all the evaluated cases compared with the standard EnKF, the shrinkage-based EnKF-KA, and the robust EnTLHF.

Part of this chapter is under review:
(Lopez-Restrepo et al.,2021) A Robust Ensemble-based Data Assimilation Method using Shrinkage Estimator and Adaptive Inflation, **Geophysical Research Letters**

6.1. Introduction

Data assimilation (DA) is a mathematical family of methods that allows the combination of observations and models. The model is used to fill observational gaps, and the observations constrain the model dynamics (Lahoz and Schneider, 2014; Bocquet *et al.*, 2015). In most of the DA methods, the aim is to minimize the estimated error variance. For instance, Kalman Filter (KF) is an optimal method that minimizes the mean-squared-error in the estimation. The KF is optimal when the following assumptions are fulfilled: the dynamic system is linear, and the observation and model uncertainties follow a Gaussian distribution (Kalman, 1960). The Ensemble Kalman Filter (EnKF) is a KF-based Monte Carlo approximation of the KF when the state space is large, and the model is non-linear (Evensen, 2003). The EnKF uses an ensemble of model realization to approximate the first and second background error moments, making it efficient for large-scale models and suitable in the presence of non-linearities. However, in real DA applications, the assumptions required to obtain the optimal solution may not be accurate, degrading the filter performance (Evensen, 2003; Houtekamer *et al.*, 2005). Additionally, small ensemble sizes may produce a poor approximation of the model uncertainty, causing a reduction in the filter accuracy or even filter divergence.

When the system conditions do not satisfy the KF-based methods requirement, a different approach is a robust filter or robust estimator. The robust filters emphasize the robustness of the estimation to have better tolerances to high uncertainty sources. Since its purpose is not the optimality in the estimation, the robust estimator does not require a strictly statistical representation of the system and the observations (Luo and Hoteit, 2011), showing a better performance than the KF-based methods in scenarios with a poor statistical uncertainty representation (Han *et al.*, 2009; Nan and Wu, 2017). There are several robust ensemble-based DA schemes based in different aspect such as H_∞ formulation (Han *et al.*, 2009), replacing the traditional L_2 norm (Roh *et al.*, 2013; Freitag *et al.*, 2013; Rao *et al.*, 2017), robust covariance estimation (Yang *et al.*, 2001; Nino-Ruiz *et al.*, 2018), and covariance inflation (Luo and Hoteit, 2011; Bai *et al.*, 2016). The approach that we propose uses a shrinkage-based covariance estimator that improves the model robustness and performance when the ensemble size is small. Additionally, our method incorporates adaptive covariance inflation closely related to the H_∞ formulation.

6.2. Ensemble time-local H_∞ filter

One of the most widely used robust filter is the H_∞ Filter (HF) (Hassibi *et al.*, 2000). The HF is based on the criterion of minimizing the supremum of the L_2 norm of the uncertainty sources (Han *et al.*, 2009). The HF ensures that the total energy of the estimation errors, is not larger than the uncertainty energy times a factor $1/\gamma$:

$$\sum_{t=0}^{M} ||\mathbf{x}_t^t - \mathbf{x}_t^a||^2_{\mathbf{S}_t} \leq \frac{1}{\gamma}\left(||\mathbf{x}_0^t - \mathbf{x}_0^a||^2_{\Delta_0^{-1}} + \sum_{t=0}^{M} ||\mathbf{u}_t||^2_{\mathbf{Q}_t^{-1}} + \sum_{t=0}^{M} ||\mathbf{v}_t||^2_{\mathbf{R}_t^{-1}} \right), \quad (6.1)$$

where $\mathbf{x}^t$ is the true state, $\mathbf{x}^a$ is the analysis state, $\mathbf{S}$ is a user-chosen matrix of weights, $\mathbf{u}$ and $\mathbf{v}$ are the model and observation uncertainty respectively, $\Delta_0, \mathbf{Q}$ and $\mathbf{R}$ are the uncertainty weighting matrices with respect to the initial conditions, model error and observations error, and M is the data assimilation windows length (Luo and Hoteit, 2011). To solve Equation 6.1, the cost function $\mathcal{J}^{\mathrm{HF}}$ is defined as:

$$\mathcal{J}^{\mathrm{HF}} = \frac{\sum_{t=0}^{M} ||\mathbf{x}_t^t - \mathbf{x}_t^a||^2_{\mathbf{S}_t}}{||\mathbf{x}_0^t - \mathbf{x}_0||^2_{\Delta_0^{-1}} + \sum_{t=0}^{M} ||\mathbf{u}_t||^2_{\mathbf{Q}_t^{-1}} + \sum_{t=0}^{M} ||\mathbf{v}_t||^2_{\mathbf{R}_t^{-1}}}. \tag{6.2}$$

Then Inequality 6.1 is equivalent to $\mathcal{J}^{\mathrm{HF}} \leq \frac{1}{\gamma}$. Let γ^* be the value such that:

$$\frac{1}{\gamma^*} = \underset{\{\mathbf{x}_t^a\}}{inf} \underset{\mathbf{x}_0,\{\mathbf{u}_t\},\{\mathbf{v}_t\}}{sup} \mathcal{J}^{\mathrm{HF}}, t \leq M, \tag{6.3}$$

the optimal HF is then achieved when $\gamma = \gamma^*$. In this formulation, the evaluation of γ^* is an application of the minimax rule (Berger, 1985), a strategy that aims to provide robust estimates and is different from its Bayesian counterpart (Luo and Hoteit, 2011). An Ensemble-based HF implementation for a nonlinear DA problem is the Ensemble time-local H$_\infty$ filter (EnLTHF) proposed by (Luo and Hoteit, 2011). In the EnLTHF a local cost function is proposed:

$$\mathcal{J}_t^{\mathrm{HF}} = \frac{||\mathbf{x}_t^t - \mathbf{x}_t^a||^2_{\mathbf{S}_t}}{||\mathbf{x}_0^t - \mathbf{x}_0||^2_{\Delta_0^{-1}} + ||\mathbf{u}_t||^2_{\mathbf{Q}_t^{-1}} + ||\mathbf{v}_t||^2_{\mathbf{R}_t^{-1}}}. \tag{6.4}$$

The local performance level γ_t satisfies:

$$\frac{1}{\gamma_t} \geq \frac{1}{\gamma_t^*} = \underset{\{\mathbf{x}_t^a\}}{inf} \underset{\mathbf{x}_0,\{\mathbf{u}_t\},\{\mathbf{v}_t\}}{sup} \mathcal{J}_t^{\mathrm{HF}}, \tag{6.5}$$

The EnLTHF can be expressed in terms of the EnKF algorithm using the notation of (Luo and Hoteit, 2011):

$$\begin{aligned}
\left[\mathbf{P}_t^a, \mathbf{K}_t\right] &= EnKF(\mathbf{x}_t^a, \mathbf{Q}_t, \mathbf{H}), && \text{(6.6a)}\\
\mathbf{G_t} &= \left[I_m - \gamma_t \cdot \mathbf{P}_t^a \cdot \mathbf{S}_t\right]^{-1} \cdot \mathbf{K}_t, && \text{(6.6b)}\\
\mathrm{x}_t^{a(i)} &= \mathrm{x}_t^{b(i)} + \mathbf{G_t} \cdot [\mathrm{y_t} - \mathbf{H_t} \cdot \mathrm{x}_\mathbf{t}^{\mathbf{b(i)}} + \mathbf{v_t^i}], && \text{(6.6c)}\\
\mathbf{x}_t^a &= \left(\sum_{i=1}^{N} \mathrm{x}_t^{a(i)}\right)/N, && \text{(6.6d)}\\
\left(\Delta_t^a\right)^{-1} &= (\mathbf{P}_t^a)^{-1} - \gamma_t \cdot \mathbf{S}_t, && \text{(6.6e)}
\end{aligned}$$

subject to the constraint

$$(\Delta_t^a)^{-1} = (\mathbf{P}_t^a)^{-1} - \gamma_t \cdot \mathbf{S}_t \geq 0, \tag{6.6f}$$

where the operator $EnKF(\cdot,\cdot,\cdot)$ means that $\mathbf{P}_t^a$ and $\mathbf{K}_t$ are obtained through the EnKF.

6.3. Robust Shrinkage-based Ensemble Kalman Filter

6.3.1. Adaptive inflation

A particular issue with ensemble-based DA algorithms is the covariance undersampling. Undersampling leads to further problems such as the ensemble collapse to an overconfident, but incorrect state, or even filter divergence (Anderson, 2001). The covariance inflation artificially increases uncertainties in the background covariance avoiding the underestimation of uncertainties, and undersampling (Bellsky and Mitchell, 2018). The magnitude of the inflation depends to a large degree on each system and application (Houtekamer and Zhang, 2016).
In equation 6.6e, the presence of the extra term $-\gamma_t \cdot \mathbf{S}_t$ inflates the EnKF covariance matrix. In this way, it is possible to interpretate the EnTLHF as an EnKF formulation with a specific value of inflation. This implies a theoretical and solid background to construct robust filters. Consider the case where $\mathbf{S} = \mathbf{I}_n$, that corresponds with an inflation of the analysis covariance matrix eigenvalues. To satisfy the constraint 6.6f, or what is equivalent, to make $(\Delta_t^a)^{-1}$ semi-definite positive, consider the SVD decomposition of $\mathbf{P}_t^a$

$$\mathbf{P}_t^a = V_t \cdot \Sigma_t \cdot U_t, \tag{6.7}$$

where $\Sigma_t = diag(\sigma_{t,1}, ..., \sigma_{t,n})$ is a diagonal matrix with all the eigenvalues of $\mathbf{P}_t^a$ in descending order, that is, $\sigma_{t,1} \geq \sigma_{t,2} \geq .... \geq \sigma_{t,n}$ and γ_t is a variable that satisfies

$$\sigma_{t,1}{}^{-1} - \gamma_t \geq 0,$$

that corresponds with

$$\gamma_t \leq \frac{1}{\sigma_{t,1}},$$

guaranteeing that $(\Delta_t^a)^{-1}$ is semi-definite positive. It is convenient to introduce a performance level coefficient (PLC) c by defining

$$\gamma_t \leq \frac{c}{\sigma_{t,1}}. \tag{6.8}$$

In contrast to conventional inflation schemes, γ_t is adaptive in time even for a fixed c value, and it is directly related with the analysis covariance matrix.

6.3.2. EnTLHF-KA

According to sections 6.2 and 6.3.1, with a specific structure and inflation value, it is possible to obtain a robust version of the EnKF. Although the EnTLHF has shown to have a better performance than the EnKF in scenarios with high uncertainty (Luo and Hoteit, 2011; Altaf *et al.*, 2013; Triantafyllou *et al.*, 2013), the limitations of the EnKF with respect to the ensemble size and the ensemble normality distribution are inherited in its robust version. When the ensemble size is small $N << n$, sampling errors can have impact on the quality of covariances matrix estimation causing problems such as filter divergence and spurious correlations

(Evensen, 2003; Houtekamer and Zhang, 2016) . Even though many localization techniques have been developed to mitigate those problems, it usually prohibits its implementation in high dimensional applications (Sakov and Bertino, 2011). The shrinkage-covariance estimator methods have shown a better performance than the classical sampling covariance matrix in scenarios with small ensemble size and non-gaussianities (Chen *et al.*, 2009; Nino-Ruiz and Sandu, 2015, 2017; Ledoit and Wolf, 2018). We propose a robust implementation of the EnKF-KA shrinkage-based method following the principles of the EnTLHF and the adaptive inflation denoted EnTLHF-KA. The EnTLHF-KA can be obtained similarly to the EnLTHF by taking as base the EnKF-KA:

$$\left[\hat{\mathbf{B}}_t^a, \mathbf{K}_t\right] = \text{EnKF-KA}(\mathbf{x}_t^a, \mathbf{T}_{KA}, \mathbf{H}), \tag{6.9a}$$

$$\mathbf{G_t} = \left[I_m - \gamma_t \cdot \hat{\mathbf{B}}_t^a \cdot \mathbf{S}_t\right]^{-1} \cdot \mathbf{K}_t, \tag{6.9b}$$

$$\mathrm{x}_t^{a(i)} = \mathrm{x}_t^{b(i)} + \mathbf{G_t} \cdot [\mathrm{y_t} - \mathbf{H_t} \cdot \mathrm{x}_\mathbf{t}^{\mathbf{b(i)}} + \mathbf{v_t^i}], \tag{6.9c}$$

$$\mathbf{x}_t^a = \left(\sum_{i=1}^{N} \mathrm{x}_t^{a(i)}\right)/N, \tag{6.9d}$$

where the operator EnKF-KA$(\cdot,\cdot,\cdot)$ represents the EnKF-KA shrinkage-based method (see Section 5.3). For an specific PLC, the inflation value is obtained using the equation 6.8.

6.4. Results and discussion

6.4.1. Numerical experiments

The Lorenz-96 is one of the most used benchmarks for testing data assimilation algorithms. The model is highly non-linear and with a strong relationship between the states. The Lorenz-96 dynamics are described by: (Lorenz and Emanuel, 1998; Gottwald and Melbourne, 2005):

$$\frac{dx_j}{dt} = \begin{cases} (x_2 - x_{n-1}) \cdot x_n - x_1 + F & \text{for } j = 1, \\ (x_{j+1} - x_{j-2}) \cdot x_{j-1} - x_j + F & \text{for } 2 \leq j \leq n-1, \\ (x_1 - x_{n-2}) \cdot x_{n-1} - x_n + F & \text{for } j = n, \end{cases} \tag{6.10}$$

where n is the state number choosen as 40, and F is the external force. For consistency, periodic boundary conditions are assumed. We take the next considerations for the numerical experiments:

- The assimilation window consist of $M = 500$ observations.
- The number of observed components is $m = 20$, representing and 50% of the model components.
- The observation statistics are associated with the Gaussian distribution,

$$\mathbf{y}_t \sim \mathcal{N}(\mathbf{H} \cdot \mathbf{x}_t^a, \rho_o^2 \cdot \mathbf{I}), \text{ for } 1 \leq t \leq M, \tag{6.11}$$

where $\rho_o = 0.001$, and $\mathbf{H}$ is a linear operator that randomly chooses the m observed components.

- To avoid random fluctuations, each experiment is repeated 20 times ($L = 20$).
- We compare the performance and robustness of the EnTLHF-KA against the non-robust methods EnKF and EnKF-KA, and the robust method EnTLHF.
- We take the Root-Mean-Square-Error (RMSE) of L experiments as a measure of performance (see eq. 5.25).
- We choice a PLC value $c = 0.5$ for all the experiments, following Luo and Hoteit (2011). Other c values have been tested (not reported here), but no performance improvements were obtained.

6.4.2. Robustness against Ensemble members

When the state dimension is large, it is important to test the performance with relative small ensemble sizes. We evaluate both the accuracy and the robustness of the EnTLHF-KA with respect to the ensemble size. For this case we set the observation error $\delta = 1 \times 10^{-3}$, the observation frequency $f = 1$, and the external force $F = 8$. The ensemble size $N \in [10, 20, 50, 100, 1000]$. Figure 6.1 presents the RMSE value for those values of N.

The EnTLHF-KA has more constant RMSE values for different N. The other methods present variation in its performance when the ensemble size changes. In general, the RMSE values decrease for larger N values for all the methods. For $N = 10$, the EnTLHF-KA presents a superior performance compared to the others, followed by the EnKF-KA. This behavior is attributed to the shrinkage-based estimator used in both methods, that have shown a better covariance estimation when $N << n$ (Nino-Ruiz and Sandu, 2017; Lopez-Restrepo *et al.*, 2021). However, the adaptive inflation factor of the EnTLHF, and the ENTLHF-KA improves these methods' performance against its non-robust counterpart. For larger ensemble size, both EnTLHF-KA and EnKF-KA tend to converge to the EnTLHF and EnKF respectively, since the sampling ensemble matrix represents a good estimator for the covariance matrix and $\hat{\mathbf{B}}_{KA}$ converge to $\mathbf{P}^a$. Due to the good estimation of $\mathbf{B}$ by $\mathbf{P}^a$, and all the EnKF assumptions are satisfied, the non-robust methods present lower RMSE value for large ensemble size. This example clarifies the different advantages and disadvantages of the robust approach compared to the optimal approach. Although the EnTLHF-KA performance is not the best in all the scenarios, its robustness allows it to have low RMSE values in all the scenarios.

6.4.3. Robustness against observation error

Figure 6.2 shows the RMSE value when $\delta \in [1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}]$. The other model parameters are: $N = 20$, $f = 1$, and $F = 8$. The idea now is to evaluate the impact of the observation error in the new robust EnTLHF-KA. It can be seen that the performance of the non-robust methods is affected by the increase of the observation error, causing divergence of the EnKF-KA. This kind of behavior

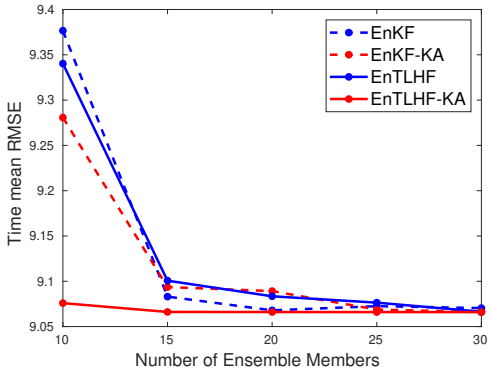


Figure 6.1: Error evaluation of the robust and non-robust methods respect to the ensemble member number.

is one of the main reasons for the development of the new robust techniques (Rao *et al.*, 2017). The observation error's impact is much lower in the robust methods, and the performance is almost constant, especially in the EnTLHF-KA. When $\delta = 1\times^{-4}$, the EnKF and the EnKF-KA perform better than its robust counterpart, but the robust filters hold a good performance even for large observation errors.

6.4.4. Robustness against model errors

To evaluate the EnTLHF-KA robustness respect to model errors, we compare the method's performance when $F \in [6, 7, 8, 9, 10]$. $F = 8$ corresponds with the assumption of a perfect model. Figure 6.3 presents the RMSE value for each F value and the comparison among the four filters. The RMSE values remain almost constant for both robust filters, with smaller values for the EnTLHF-KA. The adaptive inflation makes the analysis covariance matrix larger in the robust filters that in its non-robust counterpart, given the same background covariance. Consequently, the EnTLHF and the EnTLHF-KA put more weight in the observations, convenient when

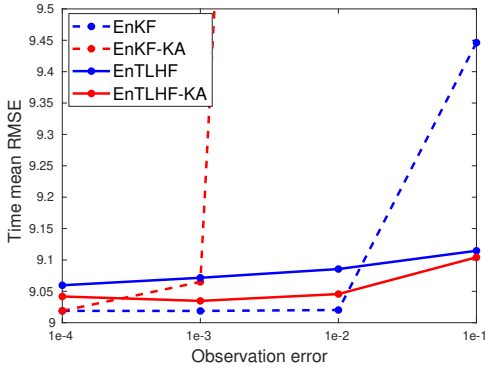


Figure 6.2: Error evaluation of the robust and non-robust methods respect to the observation error.

there are larger model errors.

6.4.5. Robustness against ensemble distribution

The standard EnKF assumes that the ensemble state has a Gaussian distribution. This assumption is especially essential because the state covariance $\mathbf{B}$ is approximated by the ensemble sample covariance $\mathbf{P}^b$. Although the ensemble at t_0 is Gaussian, non-linearities in the model dynamics can modify the ensemble distribution, causing the approximation of $\mathbf{B}$ by $\mathbf{P}^b$ to lose accuracy. Figure 6.4 presents an evaluation of the ensemble distribution for different times steps using the Lorenz-96 model. We use the Shapiro-Wilk to evaluate the gaussianity of each state variable (Shapiro and Wilk, 1965). We take an initial Gaussian ensemble of 100 members as reference. After 15-time steps, some variables begin to change its initial distribution, and after 30-time steps, the Gaussian assumption is not valid anymore for the ensemble.

We perform different experiments varying the observation frequency or the number of time steps between two available observations. Figure 6.5 shows the time averaged RMSE for the EnKF, EnKF-KA, EnTLHF and the EnTLHF-KA using a observation frequency $f \in [1, 5, 10, 20, 30, 50]$ times steps. We set an ensemble size of $N = 20$, an observation error of $\delta = 1 \times 10^{-3}$, and the external force $F = 8$. The EnKF performance decreases considerably when f increases, and after the value of $f = 30$ the method diverges. This result illustrates the importance of the Gaussian distribution for obtaining a good representation of $\mathbf{B}$ throw $\mathbf{P}^b$. The adaptive inflation increases EnTLHF robustness and performance, even when both EnKF and EnTLHF are using the same approximation of $\mathbf{B}$. Nevertheless, the EnTLHF performance decrease considerably when $f = 50$. In contrast, EnKF-KA and EnTLHF-KA use a shrinkage-based estimator for $\mathbf{B}$. The KA estimator does not assume a Gaussian distribution, as other shrinkage-based estimators do (Ledoit and Wolf, 2018; Nino-Ruiz *et al.*, 2021). As a result, the EnKF-KA presents better performance than EnKF for large f values, and similar error levels than EnTLH without incorporating adaptive inflation. In the case of the EnTLHF-KA, the combination of both the

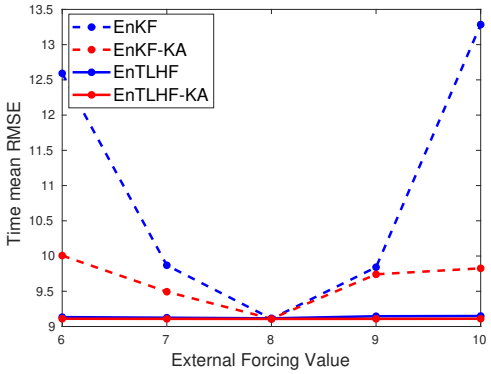


Figure 6.3: Error evaluation of the robust and non-robust methods respect errors in the model.

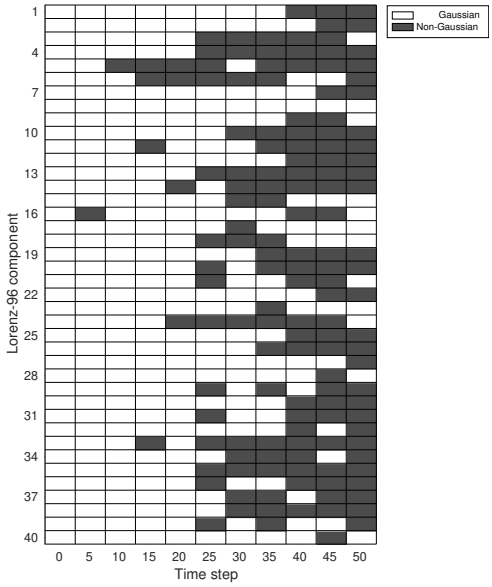


Figure 6.4: Shapiro-Wilk test for each Lorenz component at different time step. The ensemble size is 100. The white color represents that the null-hypothesis is not rejected (the ensemble for that specific variable is Gaussian). The grey color represents that the null-hypothesis is rejected (the ensemble for that specific variable is non-gaussian).

shrinkage-based estimator and the adaptive inflation produces high robustness and performance even when the ensemble distribution is non-gaussian.

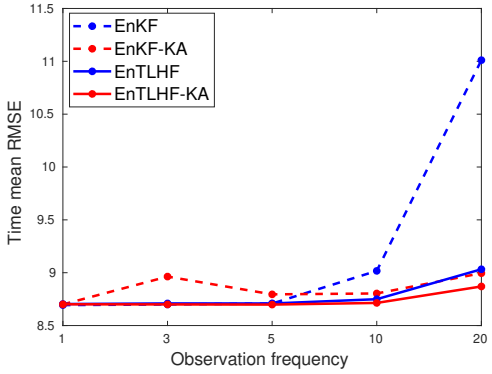


Figure 6.5: Error evaluation of the robust and non-robust methods respect to the observation frequency.

6.5. Conclusions

We propose a robust version of the shrinkage-based EnKF-KA algorithm using adaptive inflation derived from the concept of H_∞ filter (EnTLHF-KA). The EnTLHF-KA uses a covariance estimator that allows the incorporation of prior information and does not assume a Gaussian distribution in the background. Using numerical experiments, we compared the proposed method's robustness and performance against the standard EnKF, the shrinkage-based EnKF-KA, and the robust filter EnTLHF. The EnTLHF-KA has lower RMSE values in conditions with high observation error and model errors than the other methods. When the number of ensembles is small, the shrinkage estimator gives a better approximation of the background covariance matrix than the sample covariance matrix, generating lower errors in both shrinkage-based algorithm, especially in the EnTLHF-KA. The combination of the non-gaussian shrinkage estimator and the adaptive inflation grant a higher robustness to the EnTLHF-KA when the ensemble distribution is non-gaussian. All these characteristics make the EnTLHF-KA a suitable option in applications with highly non-linear models, high observation frequency, and computational restrictions in the number of ensembles.

References

W. A. Lahoz and P. Schneider, *Data assimilation: Making sense of Earth Observation,* Frontiers in Environmental Science **2**, 1 (2014).

M. Bocquet, H. Elbern, H. Eskes, M. Hirtl, R. Aabkar, G. R. Carmichael, J. Flemming, A. Inness, M. Pagowski, J. L. Pérez Camaño, P. E. Saide, R. San Jose, M. Sofiev, J. Vira, A. Baklanov, C. Carnevale, G. Grell, and C. Seigneur, *Data assimilation in atmospheric chemistry models: Current status and future prospects for coupled chemistry meteorology models,* Atmospheric Chemistry and Physics **15**, 5325 (2015), 9809069v1 [arXiv:gr-qc] .

R. E. Kalman, *A new approach to linear filtering and prediction problems,* Transactions of the ASME–Journal of Basic Engineering **82**, 35 (1960).

G. Evensen, *The Ensemble Kalman Filter: Theoretical formulation and practical implementation,* Ocean Dynamics **53**, 343 (2003).

P. L. Houtekamer, H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, *Atmospheric data assimilation with an ensemble kalman filter: Results with real observations,* Monthly weather review **133**, 604 (2005).

X. Luo and I. Hoteit, *Robust Ensemble Filtering and Its Relation to Covariance Inflation in the Ensemble Kalman Filter,* Monthly Weather Review **139**, 3938 (2011), arXiv:1108.0158 .

Y. Han, Y. Zhang, Y. Wang, S. Ye, and H. Fang, *A new sequential data assimilation method,* Science in China, Series E: Technological Sciences **52**, 1027 (2009).

T.-c. Nan and J.-c. Wu, *Application of ensemble H-infinity filter in aquifer characterization and comparison to ensemble Kalman filter,* Water Science and Engineering **10**, 25 (2017).

S. Roh, M. G. Genton, M. Jun, I. Szunyogh, and I. Hoteit, *Observation Quality Control with a Robust Ensemble Kalman Filter,* Monthly Weather Review **141**, 4414 (2013).

M. A. Freitag, N. K. Nichols, and C. J. Budd, *Resolution of sharp fronts in the presence of model error in variational data assimilation,* Quarterly Journal of the Royal Meteorological Society **139**, 742 (2013).

V. Rao, A. Sandu, M. Ng, and E. D. Nino-Ruiz, *Robust data assimilation using l1 and huber norms,* SIAM Journal on Scientific Computing **39**, B548 (2017).

Y. Yang, H. He, and G. Xu, *Adaptively robust filtering for kinematic geodetic positioning,* Journal of Geodesy **75**, 109 (2001).

E. Nino-Ruiz, H. Cheng, R. Beltran, E. D. Nino-Ruiz, H. Cheng, and R. Beltran, *A Robust Non-Gaussian Data Assimilation Method for Highly Non-Linear Models,* Atmosphere **9**, 126 (2018).

Y. Bai, Z. Zhang, Y. Zhang, and L. Wang, *Inflating transform matrices to mitigate assimilation errors with robust filtering based ensemble Kalman filters,* Atmospheric Science Letters **17**, 470 (2016).

B. Hassibi, T. Kailath, and A. Sayed, *Array algorithms for h∞ estimation,* Automatic Control, IEEE **45**, 702 (2000).

J. O. Berger, *Statistical decision theory and Bayesian analysis*, Springer Series in Statistics (Springer, New York, 1985).

J. L. Anderson, *An ensemble adjustment kalman filter for data assimilation,* Monthly Weather Review **129**, 2884 (2001).

T. Bellsky and L. Mitchell, *A shadowing-based inflation scheme for ensemble data assimilation,* Physica D: Nonlinear Phenomena **380-381**, 1 (2018).

P. L. Houtekamer and F. Zhang, *Review of the ensemble Kalman filter for atmospheric data assimilation,* Monthly Weather Review **144**, 4489 (2016).

M. U. Altaf, T. Butler, X. Luo, C. Dawson, T. Mayo, and I. Hoteit, *Improving short-range ensemble kalman storm surge forecasting using robust adaptive inflation,* Monthly Weather Review **141**, 2705 (2013).

G. Triantafyllou, I. Hoteit, X. Luo, K. Tsiaras, and G. Petihakis, *Assessing a robust ensemble-based Kalman filter for efficient ecosystem data assimilation of the Cretan Sea,* Journal of Marine Systems **125**, 90 (2013).

P. Sakov and L. Bertino, *Relation between two common localisation methods for the EnKF,* Computational Geosciences **15**, 225 (2011).

Y. Chen, A. Wiesel, and A. O. Hero, *Shrinkage estimation of high dimensional covariance matrices,* in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2009) pp. 2937–2940.

E. D. Nino-Ruiz and A. Sandu, *Ensemble kalman filter implementations based on shrinkage covariance matrix estimation,* Ocean Dynamics **65**, 1423 (2015).

E. D. Nino-Ruiz and A. Sandu, *Efficient parallel implementation of DDDAS inference using an ensemble Kalman filter with shrinkage covariance matrix estimation,* Cluster Computing **22**, 2211 (2017).

O. Ledoit and M. Wolf, *Optimal estimation of a large-dimensional covariance matrix under stein's loss,* Bernoulli **24**, 3791 (2018).

E. N. Lorenz and K. A. Emanuel, *Optimal Sites for Supplementary Weather Observations: Simulation with a Small Model,* Journal of the Atmospheric Sciences **55**, 399 (1998), https://journals.ametsoc.org/jas/article-pdf/55/3/399/3437184/1520-0469(1998)055_0399_osfswo_2_0_co_2.pdf .

G. A. Gottwald and I. Melbourne, *Testing for chaos in deterministic systems with noise,* Physica D: Nonlinear Phenomena **212**, 100 (2005).

S. Lopez-Restrepo, E. D. Nino-Ruis, A. Yarce, O. L. Quintero, N. Pinel, A. Segers, and A. W. Heemink, *An Efficient Ensemble Kalman Filter Implementation Via Shrinkage Covariance Matrix Estimation: Exploiting Prior Knowledge,* Computational Geosciences **25**, 985–1003 (2021).

S. Shapiro and M. Wilk, *An analysis of variance test for normality (complete samples†,* Biometrika **52**, 591 (1965), https://academic.oup.com/biomet/article-pdf/52/3-4/591/962907/52-3-4-591.pdf .

E. D. Nino-Ruiz, L. Guzman, and D. Jabba, *An ensemble Kalman filter implementation based on the Ledoit and Wolf covariance matrix estimator,* Journal of Computational and Applied Mathematics **384** (2021), 10.1016/j.cam.2020.113163.

7

Using a robust data assimilation method to improve $PM_{2.5}$ modeling in the Aburrá Valley

The implementation of the shrinkage-based techniques EnKF-KA and EnTLHF-KA in an air quality application using the LOTOS-EUROS model over the Aburrá Valley is described in this chapter. The EnKF-KA is an EnKF implementation that requires a target covariance matrix to integrate previously obtained information and knowledge directly into the data assimilation. The EnTLHF-KA is a robust variant of the EnKF-KA based on the H∞ filter, using an adaptive inflation factor. In the spatial distribution of the $PM_{2.5}$ concentrations along the valley, both methods outperform the well-known LETKF. In contrast to the other simulations, the ability to issue warnings for high concentration events is also increased. Finally, the simulation using EnTLHF-KA has lower error values than using EnKF-KA, indicating the advantages of robust approaches in high uncertainty systems.

Part of this chapter is under preparation:
(Lopez-Restrepo et al.,2021) Using a robust data assimilation method to improve $PM_{2.5}$ modeling in the Aburrá Valley

7.1. Introduction

The uncertainty in CTM simulations could be reduced by improvement of the emission inventory and the upgrade of meteorological data. Alternatively one could incorporate ground data, satellite information or vertical in the simulations using Data Assimiaiton (DA) techniques to reduce the uncertainty, and this is approach is used in this study (Fu *et al.*, 2017; Lu *et al.*, 2016; Jin *et al.*, 2018; Lopez-Restrepo *et al.*, 2020). In Lopez-Restrepo *et al.* (2020), data assimilation over the Aburrá Valley has been applied using the LOTOS-EUROS CTM, building on earlier applications (Fu *et al.*, 2017; Lu *et al.*, 2016; Jin *et al.*, 2018). Figure 7.1 shows a map of the Metropolitan Area of the Aburrá Valley in Colombia, as well as maps of emission correction factors that were obtained in the data assimilation experiments. The circular patterns in the correction factors originate from the use of a traditional localization scheme using a defined radius of influence, that is used to remove spurious correlations from an ensemble covariance. For the complex topography of the valley, this is obviously not an optimal choice. In this study, shrinkage-based techniques will be used to improve the covariance description, taking into account the knowledge on the topography.

7.2. LOTOS-EUROS model and observations

The period of interest for all data evaluations, simulations and data assimilation experiments spans from February 25 to March 15, 2019. During these days, the PM concentrations are higher due to the Northbound transit of the Inter-Tropical Convergence Zone.

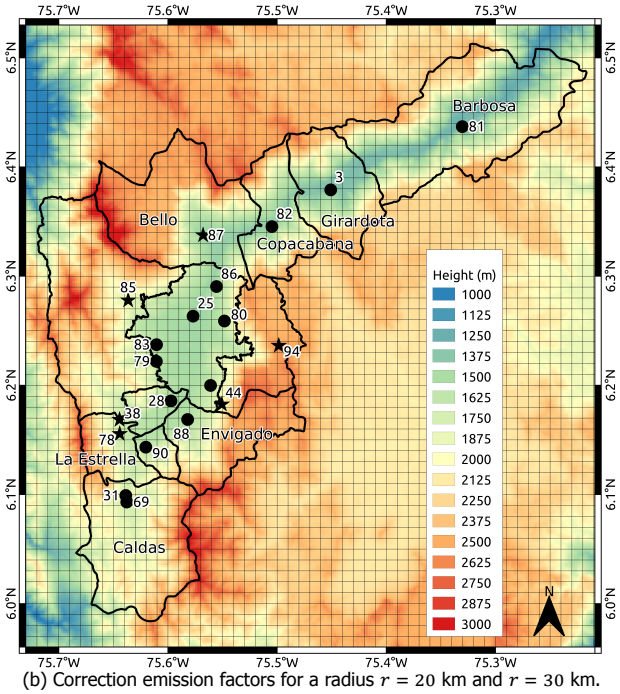
7.2.1. Simulation setup

All the simulations were conducted using the domain and experimental setup described in Section 2.1.2. The local emission inventory presented in Section 2.1.3 was used as emission input for all the simulations. Additionally, the simulations in the domain of interest (D4) where performed using the meteorological fields coming from the Weather Research and Forecasting (WRF) model (Skamarock *et al.*, 2008). The description of the WRF meteorology is presented in Section 7.2.2.

7.2.2. WRF meteorology

The WRF model is a numerical weather prediction and atmospheric simulation system designed for research and operational applications (Skamarock *et al.*, 2008). The WRF simulations are suitable to understand the behaviour of meteorological variables in a domain like the Aburrá Valley. The WRF model has been used over Colombia in previous studies(Misenis and Zhang, 2010; Carvalho *et al.*, 2012; Tuccella *et al.*, 2012; Hu *et al.*, 2013; Dillon *et al.*, 2016; Kumar *et al.*, 2016; Henao *et al.*, 2020). The configuration of the nested domains usedin this study is shown in the Figure 7.2 and described in Table 7.1. The settings used for the WRF simulations are summarized in Table 7.2.

(a) The topography of the Aburrá Valley and observations distribution.



(b) Correction emission factors for a radius $r = 20$ km and $r = 30$ km.

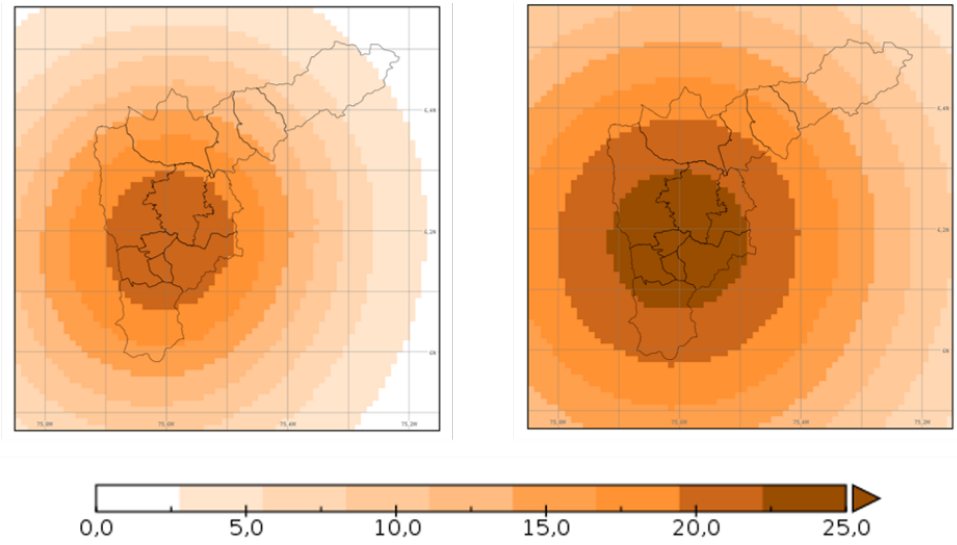


Figure 7.1: Topography and correction emission factors for different localization radius using and standard localization technique (Lopez-Restrepo *et al.*, 2020).

7

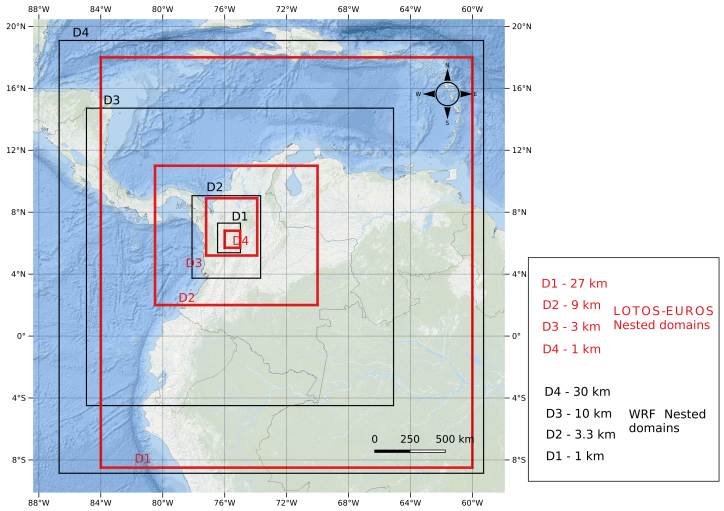


Figure 7.2: WRF and LOTOS-EUROS model nested domain configuration.

7

Domain	Latitude	Longitude	Resolution	Number of cells
D1	-8.864, 19.091	-86.694, -59.275	0.3°	90 x 93
D2	-4.946, 14.719	-84.929, -65.091	0.1°	193 x 193
D3	3.734,9.064	-78.108,-73.677	0.033°	130 x 157
D4	5.379,7.294	-76.458, -74.981	0.011°	130 x 169

Table 7.1: WRF model domains description.

Category	Parameter	Selection in WRF
Domain settings	Coordinate system	mercator True latitude 1: 36°. True latitude 2: 60°. Standard. longitude:-98°.
	Vertical setting	35.
	Nesting	Two way.
Input data	Land use	MODIS.
Initial-boundary conditions	Name of model	NCAR-GFS.
	Grid resolution	32 levels + 5 soil levels.
Physic Settings	Radiation scheme	CAM scheme.
	Microphysics	Single moment 6-class.
	Surface layer options	Layer: Monin-Obukhov. Physics: Thermal Diffusion. Scheme: soil temperature. only, using five layers.
	PBL Scheme	MYJ.
	Cumulus option	KF.

Table 7.2: WRF model set up.

7.2.3. Assimilation and validation network

We used the hyper-dense low-cost network deployed and operated by the *Sistema de Alerta Temprana del Valle de Aburrá* (SIATA) as observations for the data assimilation methods. The low-cost network consists of 255 real-time $PM_{2.5}$ sensors across the Aburrá Valley and its hills. The distribution of the low-cost network is shown in Figure 4.1. For validation, we used the independent official monitoring network of the metropolitan area. The official network has 80 measurement sites that observer particulate matter at hourly frequency (Hoyos *et al.*, 2019). The set of validation sites is split in two sets: the stations located in the bottom part of the valley (BS, represented by circles in Figure 7.1), and the stations located in the city's outskirts or hills (OS, represented by stars in Figure 7.1). The objective of this division is to evaluate the simulations performance in regions where the $PM_{2.5}$ concentration regimes are different.

7.3. Data assimilation system

We performed a total of four different LOTOS-EUROS simulations:

1. a LOTOS-EUROS model simulation without data assimilation (henceforth *LE*);
2. a DA simulation using the LETKF introduced in Section 7.3.1 (henceforth *LE-LETKF*);
3. a DA simulation using the shrinkage-based EnKF-KA developed in Chapter 5

(Lopez-Restrepo *et al.*, 2021) (henceforth *LE-KA*);

4. a DA simulation using the robust and shrinkage-based EnTLHF-KA developed in Chapter 6 (henceforth *LE-Robust*).

All the simulations were evaluated using the two validation station's sets, and the performance metrics Mean Fractional Bias (MFB), Root Mean Square Error (RMSE), and Pearson Correlation Factor (R), shown in Section 3.2.3. The three ensemble-based algorithms estimate both concentrations and emissions, following the stochastic representation presented in Section 2.2. For all the methods, an ensemble size N of 25 members and a localization radius r of 5 km where used.

7.3.1. LETKF

One of the most commonly used implementations of the EnKF method is the local ensemble transform Kalman filter (LETKF) (Ott *et al.*, 2004), where the assimilation process is performed independently for each model variable. Around each model variable (grid point), a sub-domain of radius r is constructed and the assimilation process is carried out within the local domain. Each local analysis is mapped onto the global domain to obtain the global analysis and the assimilation is completed. In the assimilation process, all the information found within the sub-domain (i.e., observed components and error correlations) is used. The analysis state could be obtained following the implementation by (Shin *et al.*, 2016) :

$$\Delta\mathbf{X} = \mathbf{X}^b - \overline{\mathbf{x}}^b \cdot \mathbf{1}^T \in \mathbb{R}^{n\times N}, \tag{7.1a}$$

$$\Delta\mathbf{Y} = \mathbf{H} \cdot \Delta\mathbf{X} \tag{7.1b}$$

$$\mathbf{P}^a = \left[\Delta\mathbf{Y}^T \cdot \mathbf{R}^{-1} \cdot \Delta\mathbf{Y} + (m-1) \cdot \mathbf{I} \right]^{-1}, \tag{7.1c}$$

$$\mathbf{D} = \mathbf{y} - \mathbf{H} \cdot \overline{\mathbf{x}}^b, \tag{7.1d}$$

$$\mathbf{w}^a = \mathbf{P}^a \cdot \mathbf{Y}^T \cdot \mathbf{R}^{-1} \cdot \mathbf{D}, \tag{7.1e}$$

$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^b + \Delta\mathbf{X} \cdot \mathbf{w}^a, \tag{7.1f}$$

$$\mathbf{X}^a = \mathbf{X}^b \cdot \left[(n-1) \cdot \mathbf{P}^a \right]^{1/2}, \tag{7.1g}$$

where n, m, and N are the model resolution, the number of observations, and the number of ensemble members respectively, $\mathbf{X}^b \in \mathbb{R}^{n\times N}$ and $\mathbf{X}^a \in \mathbb{R}^{n\times N}$ are the background and analysis ensemble, $\overline{\mathbf{x}}^b \in \mathbb{R}^{n\times 1}$ and $\overline{\mathbf{x}}^a \in \mathbb{R}^{n\times 1}$ are the background and analysis ensemble means, $\mathbf{y} \in \mathbb{R}^{m\times 1}$ is the observation vector, $\mathbf{P}^a \in \mathbb{R}^{n\times n}$ is the analysis ensemble covariance matrix, $\mathbf{H} \in \mathbb{R}^{n\times m}$ is the observation operator, $\mathbf{R} \in \mathbb{R}^{m\times m}$ is the estimated data-error covariance matrix, and $\mathbf{1}$ is a vector of consistent dimension whose components are all ones. In the LETKF algorithm, the above analysis is applied per grid cell. The algorithm becomes:

1. Compute in each domain simulated observations for all ensemble members.
2. Collect per domain also the observations from neighbouring domains that are within r distance

3. Loop over grid cells.
   (a) Select observations and simulations that are within range r.
   (b) Compute analysis weights $\mathbf{w}^a$.
   (c) Apply the analysis with the ensemble elements for the selected grid cell.
4. Once all the local analyses are performed, map those to the global domain.

Note that the background error covariance matrix approximation in the LETKF is the sample covariance matrix (5.7), therefore for large radii of influence, the quality of the LETKF results could be influenced by spurious correlations.

7.3.2. EnKF-KA and EnTLHF-KA

The shrinkage-based algorithm EnKF-KA (described in Chapter 5) and the robust EnTLHF-KA (introduced in Chapter 6) were implemented to be used with the LOTOS-EUROS model. The aim of these algorithms is to improve the model representation in the complex orography conditions of the Aburra Valley.
For the EnTLHF-KA We choice a PLC value $c = 0.5$ based in the results showed in Section 6.4.

Both shrinkage-based algorithms required a target matrix $\mathbf{T}_{KA}$ to compute the covariance matrix $\mathbf{B}$ according to (5.16). The matrix $\mathbf{T}_{KA}$ should guide the covariance structure in $\mathbf{B}$ by limiting the spurious correlations between elements at large distance (Nino-Ruiz and Sandu, 2015), or in the case of the EnKF-KA and the EnTLHF-KA, to incorporate previously obtained knowledge directly in the DA process (Lopez-Restrepo *et al.*, 2021). For this application, we are interested in using the target matrix to represent the valley's complex orography in the covariance estimation. Previous works have shown issues to reproduce the pollutant dynamics into the Aburrá valley due to the limited representation of the valley in the simulation model (Lopez-Restrepo *et al.*, 2020; Henao *et al.*, 2020). Even with high-resolution meteorological simulations, it is still challenging to capture the transport of pollutants in the narrow valleys (Rendón *et al.*, 2020).

The main purpose of the $\mathbf{T}_{KA}$ matrix is to reduce the covariance between elements in the state that are distant in the vertical direction but close in the horizontaldirection. Thus, observations located in the bottom part of the valley (where the pollutant concentration are higher) should not have a high impact in the city's outskirts (where the concentrations are lower) and vice versa. A first version of the target matrix $\mathbf{T}^*_{KA}$ was built using a fourth-order-polynomial covariance function as described in Gaspari and Cohn (1999), reducing the correlation as function of vertical distance, with zero correlation for vertical distance exceeding 600 m. Other distances were tested too, without significant changes in the result. The chosen formulation preserves the dependency on horizontal distance that is necessary to remove the spurious correlations. To ensure that $\mathbf{T}_{KA}$ is positive semidefinite, we applied the method presented in (Higham, 1988) to obtain a positive semidefinite matrix that is closest to $\mathbf{T}^*_{KA}$ in the Frobenius norm.

Figure 7.3 illustrates the influence area of the Gaspari-Cohn based covariance matrix, the $\mathbf{T}^*_{KA}$ covariance matrix, and the $\mathbf{T}_{KA}$ covariance matrix for two locations.

The influence area corresponds with a row (or column) of the covariance matrix. It is possible to see how the proposed $\mathbf{T}^*_{KA}$ matrix (Figure 7.3 (c)) follows the valley shape according to the orography shown in Figure 7.3 (b) unlike the Gaspari-Cohn covariance matrix (Figure 7.3 (a)). Additionally, there are no significant modifications between the $\mathbf{T}_{KA}$ (Figure 7.3 (d)) and the $\mathbf{T}^*_{KA}$ matrix. Finally, the $\mathbf{T}_{KA}$ matrix is used as the target matrix for both EnKF-KA and EnTLHF-KA methods. Note that the final covariance between the state inside and outside the valley will not be necessary zero because the final covariance matrix $\mathbf{B}_{KA}$ is a convex combination of $\mathbf{T}_{KA}$ and $\mathbf{P}^b$.

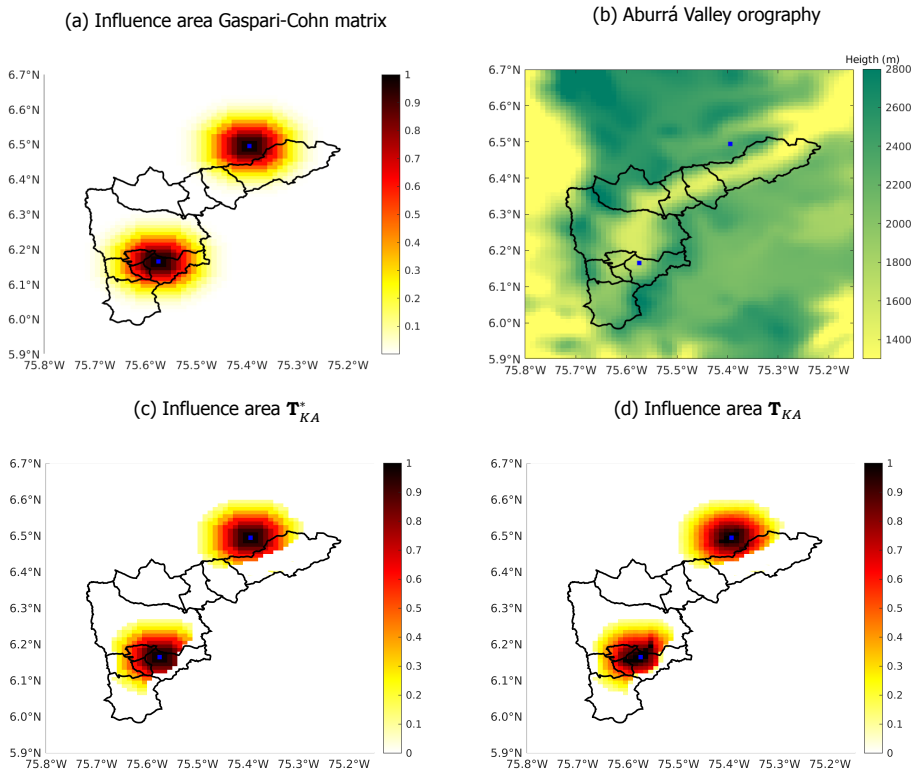


Figure 7.3: Comparison of the influence area of two selected states (blue dots) between a distance depended localization, and the target covariance matrix based in the distance and the orography.

7.3.3. Forecast experiments

The new covariances have been evaluated with forecast experiments, in which a model simulation over a limited number of days is performed using information from the assimilation. DA will primarily boost the accuracy of the forecast for two reasons. First, if the simulation is initialized with an assimilated state, the initial conditions at the beginning of the forecast window are closer to reality than the

model alone can provide. Second, it is possible to apply the emission correction factors included in the assimilation state (2.7).

Forecasting experiments were performed to test the model's capability to predict the PM concentrations in the valley up to to three days ahead. We applied the methodology proposed in Section 4.2.4, with all days from March 9 to 13 having predictions as the first, second and third day of a forecast. We are especially interested in evaluating the ability of the model to predict warning-triggering episodes (AQI in orange, red, or purple levels, see Table 4.1). All forecast simulations used the estimated emission correction factors from the last assimilation day, in each of the three forecast day. This inheritance scheme has shown the best option for the LE implementation over the Aburrá Valley ((Lopez-Restrepo *et al.*, 2020), and also Chapter 3).

7.4. Results

7.4.1. Evaluation of LE simulations

The concentration fields produced by model simulations with or without data assimilation were compared with the observations from official monitoring stations (Figure 7.1), dividing the study into stations at the bottom of the valley (BS stations) and stations at the outskirts of the city (OS stations). The averaged assessment statistics over the validation station are shown in Table 7.3. In all validation stations, the simulation results without data assimilation (LE) underestimated the observed concentrations. This is for example reflected in a high RMSE value. The correlation coefficient was low, which means that the model could not fully capture the temporal variations at hourly and daily scale. An improvement is observed when the LE simulation is compared with previous results using ECMWF meteorology (Lopez-restrepo *et al.*, 2021) (Chapter 4, Table 4.2 MFB=-0.65, RMSE=27.38, and R=0.42). The three simulations using data assimilation had MFB values similar to 0 for the BS stations (bottom of the valley), without a noticeable difference. DA was thus successful in reducing the discrepancy between the model and observations. The RMSE also decreased by 45.03% in the LE-LETKF, 41.57% in the LE-KA, and 41.91% in the LE-Robust simulations compared to the RMSE of the LE simulation. According to Mogollón-sotelo *et al.* (2020) Table 2 and based on EPA (2000); Boylan and Russell (2006), the R values were all above the criterion for good results. In contrast, over the OS stations (outskirts of the city) the simulations using the shrinkage-based methods presented better statistics respect to the LE-LETKF. For instance, the RMSE's improvements in OS stations using shrinkage-based methods are 15.02% for the LE-KA and 22.22% for the LE-Robust compare with the LE-LETKF. In general, all DA simulations showed lower scores in the OS stations than in the BS stations, mainly because of the poor representation in these areas by the background simulation (LE simulation) and the lack of close observations. Even so, the LE-Robust looks more robust among all the stations.

Figure 7.4 shows diurnal cycles in the four chosen validation stations during the simulation phase. Those stations illustrate the differences between BS and OS, and are representative for all validation stations. The LE diurnal cycle differs from the

Simulation	MFB			RMSE			R		
	BS	OS	Total	BS	OS	Total	BS	OS	Total
LE	-0.42	-1.2	-0.55	20.23	21.12	21.11	0.61	0.41	0.57
LE-LETKF	0.03	0.26	0.08	11.12	17.50	13.93	0.86	0.63	0.81
LE-KA	-0.02	-0.09	-0.02	11.82	14.87	12.88	0.84	0.71	0.82
LE-Robust	0.02	-0.03	0.01	11.75	13.61	12.22	0.84	0.78	0.83

Table 7.3: Statistical evaluation of different simulations. BS corresponds with stations located in the bottom of the valley. OS corresponds with stations located in the outskirts of the city. The total value is calculated over all the validation stations.

observations in magnitude in the BS stations, and in the OS stations in both magnitude and temporal behavior. The highest peak of concentration in the BS stations around 09:00 is primarily due to traffic dynamics and is partially captured by the LE simulation. For example, the LE morning peak emerged faster in the simulations at station 44 than in the observations. This time lag could be due to a poor spatial representation of mobile sources in the emission inventory, or a failure by the meteorology or the model to reproduce the dynamics of the valley, indicating a premature transport of particulate matter to these regions. In comparison, at 22:00 hours, the LE simulation presents the highest point at station 44 (Figure 7.4 (c)), which does not correspond with the observations. The LE simulation in the other OS station 85 (Figure 7.4 (d)), cannot fit the observation interval, indicating a late morning peak and a minimum around 21:00 that does not appear in the measurements. The LE simulation show a general underestimation of concentrations, with a better replication of the $PM_{2.5}$ dynamics at the bottom of the valley.

The simulations using data assimilation presented diurnal cycles closer to the observations, with a marked difference in performance between BS stations and OS stations. In the BS stations (Figure 7.4 (a) and (b)), the three methods showed very similar daily cycles capturing the magnitude and the variability of the observations with high accuracy. These simulations corrected the concentration underestimation presented in the LE simulation and improved the temporal profile. Unlike in the BS stations, in the OS stations, the three DA methods showed different results. The LE-LETKF tends to overestimate the concentrations and has different diurnal variability concerning the observations. In station 44, the LE-LETKF persistently displayed higher values than the observed, and a low variability around the day, with small peaks and valleys. In station 85, the LE-LETKF showed higher concentration values than the observations, and the morning peak appears later (similar to the LE simulation). The discrepancy in the magnitude and the lack of representation of the temporal variability suggest that the LE-LETKF simulation assimilates observations located in regions where the PM presents a different temporal behavior than those grid cells located in the outskirts. On the other hand, the two simulations using the shrinkage-based covariance estimator and the target matrix $\mathbf{T}_{KA}$ (LE-KA and LE-Robust) improve the performance in the OS stations. The LE-KA simulation showed a similar temporal variability in both OS stations, although a concentration underestimation. The LE-Robust displayed a high agreement between the simulated

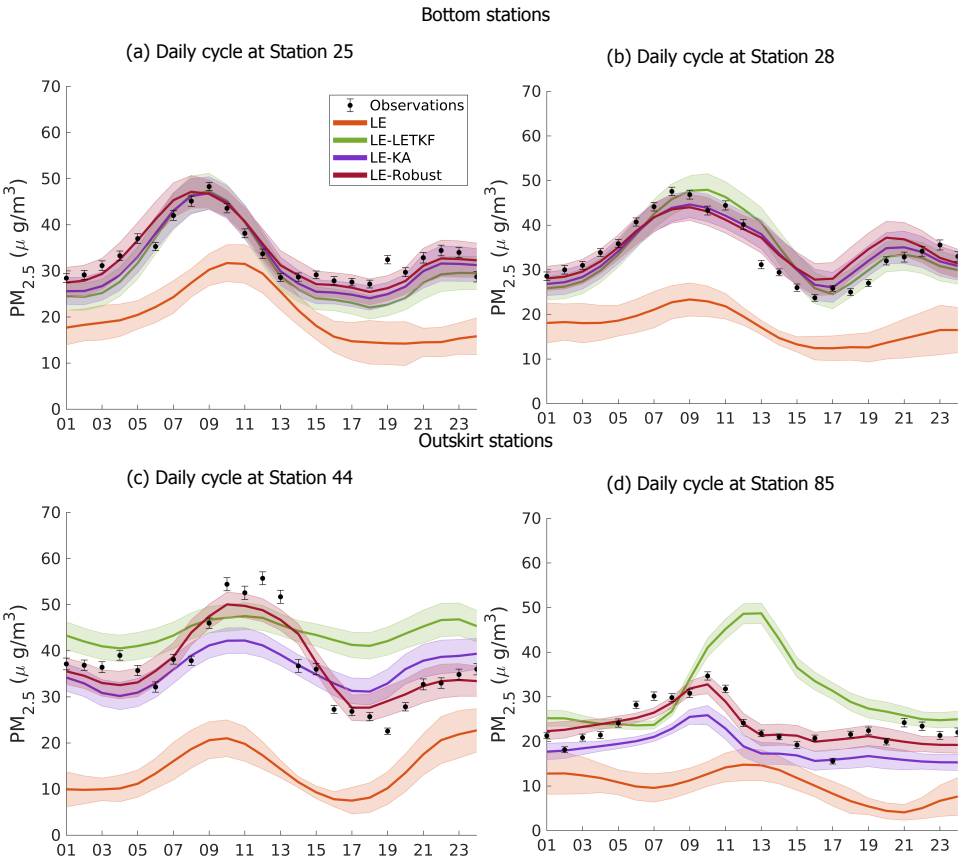


Figure 7.4: Daily cycle at different stations. The upper panel corresponds with stations located in the bottom of the valley. The bottom panel corresponds with stations located in the outskirts of the city.

daily cycle and the observations. The difference in magnitude between the LE-Robust and LE-KA simulations can be explained by the fact that the robust methods tend to put more weight in the observations when there is high uncertainty in the background (Luo and Hoteit, 2011), such as the case in this application. Finally, the shrinkage-based simulations tend to follow the diurnal variability which suggests that the $\mathbf{T}_{KA}$ matrix could limit the influence of observations from areas with a different temporal profile.

7.4.2. Spatial distribution

To better understand the influence of the target matrix $\mathbf{T}_{KA}$ on shrinkage-based methods, it is important to analyze the spatial distribution of the concentrations over the valley. Figure 7.5 shows a three-dimensional representation of the average value of $PM_{2.5}$ over March 9. In these graphs, values less than 5 μg/m^3 are omitted. The (averaged) observed values are shown using the same color bar for all the

validation stations by a circle and a star for the BS and OS stations respectively. The LE simulation has a spatial pattern similar to the observations, with the highest

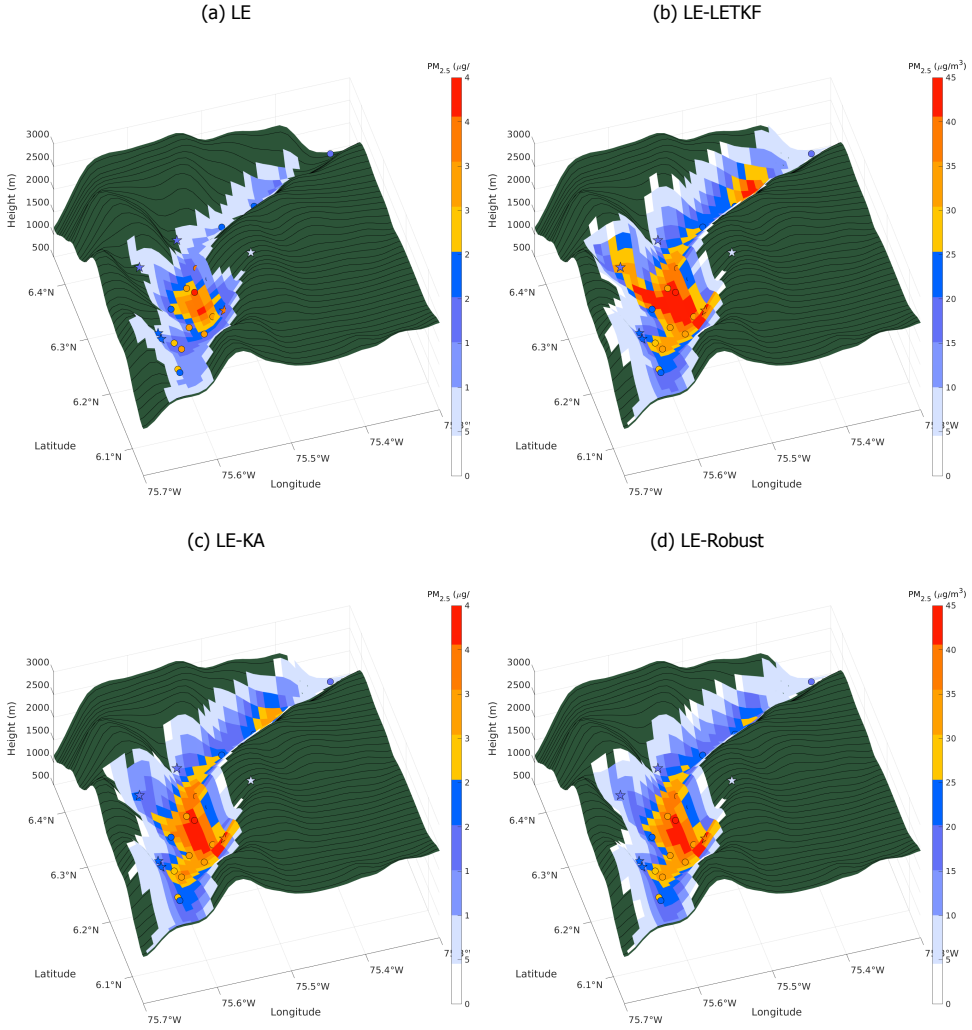


Figure 7.5: 3D maps of concentrations averaged over March 9 for different simulations. The values less than 5 $\mu g/m^3$ are omitted. The circles correspond with BS stations, and the stars corresponds with OS stations.

concentrations in the center and south part of the Medellín city (see Figure 7.1 for reference). In general, the concentrations are higher in the bottom part of the valley, where most of the population and industry facilities are located. This characteristic is well captured by the LE simulation. Nevertheless, the LE simulation tends to underestimate the concentration along the valley and the hills. The three

DA simulations are able to correct the concentration bias in the bottom part of the valley. The LE-LETKF assimilation increases the concentrations in the hills to values higher than the observations. In the station 85, located on the west slope of the valley (see Figure 7.1 for reference), the concentrations simulated by LE-LETKF are almost everywhere higher than the observed. This is because the concentrations in the west hill are influenced by observations located in the lower part of the valley, characterized by high concentrations. Those observations influence the grid cells located on the hill, generating values that do not correspond to the validation station. Both shrinkage-based simulations match better with the observations on the hills. In the case of the station 85, both methods have the same range of values as the observed concentrations. The use of the $\mathbf{T}_{KA}$ matrix limits the influence of the observations located in the bottom of the valley on the grid cells at the slopes. As can be seen in Figure 7.3 (d), the influence of the observations is limited by horizontal and vertical distance, representing better the dynamics in the valley. A particular situation is observed at station 94 (see Figure 7.1 for reference), located on the top of the east slope. Although the observed values are in the range of 5-10 μg/m^3, all the simulation, even the DA simulations, show values under 5 μg/m^3 (not plotted in Figure 7.5). The underestimation can be explained by an absence of emissions in the emission inventory (emission uncertainties), and the limited number of observations in that part of the domain.

7.4.3. Forecast evaluation

A fundamental prerequisite for a simulation and assimilation method of air quality to be valuable for a decision-making process is that it can predict the concentrations a few days in advance. Figure 7.6 shows examples of forecasts from 12 March 16:00 to 15 March 16:00. As was mentioned previously, the forecast runs are using the emission correction factors estimated between 10 March 16:00 and 11 March 16:00. The LE simulation persistently underestimates the concentrations, as observed in the assimilation window's results. In the BS stations, the three assimilation methods initiate a forecast that is quite close to the observations on the first day and remains with an acceptable similarity in the following two forecast days. As shown in the previous evaluations, the concentrations in the assimilation window are very similar for the three methods in the lower part of the valley. Thus, also the estimated emission correction factors are similar, leading to rather small differences between the forecasts. However, in the OS stations, the LE-LETKF forecasts show magnitudes and a temporal behavior that is different from the observations. This discrepancy in the values suggests an incorrect estimation of the emission correction factors on the slopes of the valley by LE-LETKF. The forecasts generated by the shrinkage-based methods are more similar to the observations. The LE-KA and LE-Robust show a good forecasting skill for the OS stations, with temporal behavior and magnitudes close to those observed for the first and second forecast days.

To be valuable for the public, a forecast should correctly warn for elevated air pollution events. The portion of true negatives, true positives, false negatives, and false positives regarding with the prediction of warning-triggering episodes (AQI in orange, red, or purple levels, see Table 4.1) is summarized by the confusion

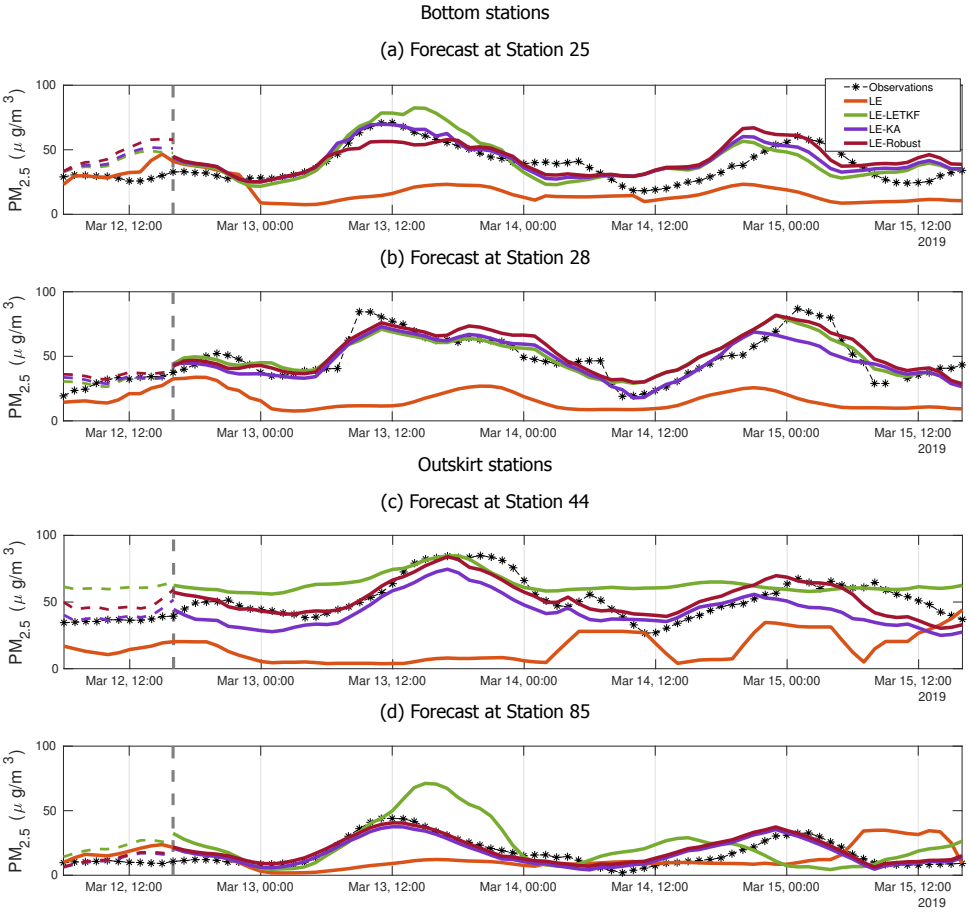


Figure 7.6: Forecast from 12 March 16:00 to 15 March 16:00 at different stations. The grey vertical dashed line represents the end of the assimilation window and the beginning of the forecast window.

matrix (Kohavi and Provost, 1998). Figure 7.7 shows the confusion matrices for LE-LETKF, LE-KA, and LE-Robust assimilations and forecasts. In the assimilation or forecast window, the LE simulation did not give an alert at any station; for that reason, we do not provide its confusion matrix. Data assimilation simulations have a ratio between true negatives and true positives equal to or greater than 90%Of the ˙20 alarms registered in the assimilation window, 18 correspond to BS stations. In the forecast window, the forecast skill of the three models was lower than in the assimilation window. From the 10 actually observed alerts in the forecast period, the DA simulations could replicate 8. A higher proportion of false-positive alerts was reported by the LE-LETKF, documenting nine false alerts more than the shrinkage-based approaches. The high amount of false-positive alerts is due to the overestimation of the LE-LETKF concentration in the OS stations where the additional alerts were recorded incorrectly. In general, the LE-KA and LE-Robust

simulations had better alert forecast performance than the LE-LETKF simulation.

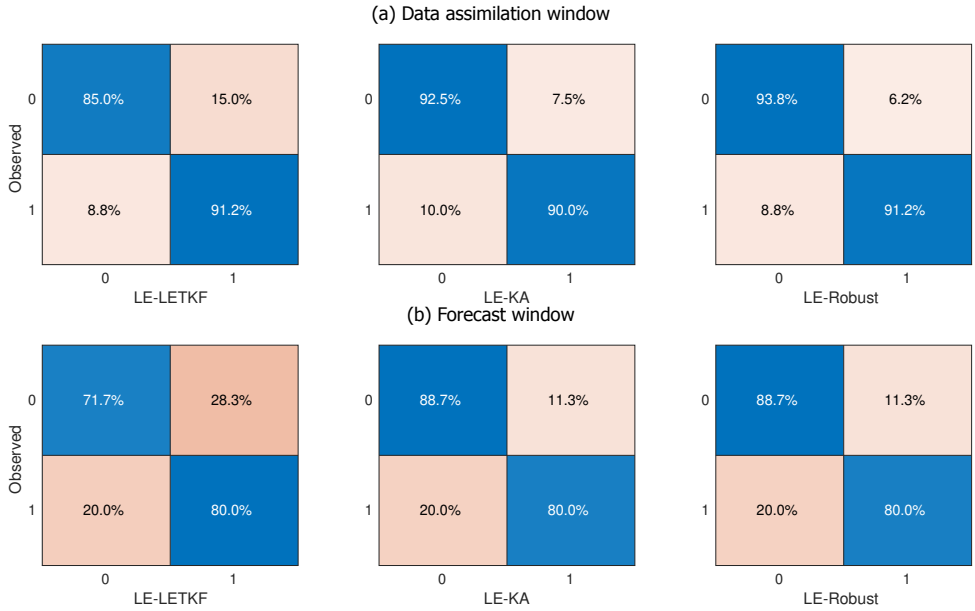


Figure 7.7: Comparison of confusion matrices for the data assimilation and forecast window depending on warning or no warning per station. The values are calculated across all the days of the corresponding window. The value of 0 corresponds with no warning, the value of 1 corresponds with a warning. For the LE simulation, there are no warnings in the data assimilation window nor forecast windows.

7.5. Discussion and comments

The simulations of the LE model showed an improvement when using the meteorology generated by the WRF model. Previous studies already suggested the need for meteorological fields at a higher resolution to correctly represent the dynamics and transport of pollutants in the Aburrá Valley (Lopez-Restrepo *et al.*, 2020) (chapters 3, 4). Simulation without data assimilation and using WRF meteorology (LE simulation) shows an improvement compared to implementations using the lower resolution ECMWF meteorology. An underestimation of $PM_{2.5}$ concentrations is strongly reduced (although still present) and an increment in the correlation is observed. It is important to continue evaluating the model's performance with different configurations of the WRF model specifically to reproduce the dominant dynamics of pollutant transport in inhabited valleys (Henao *et al.*, 2020; Rendón *et al.*, 2020). Additionally, it is necessary to carry out a more exhaustive evaluation of the model's vertical resolution given the new possibilities offered by the coupling with the WRF model. Finally, a reduction in the meteorology's uncertainty will improve the estimation of the emissions using data assimilation and could help to create more accurate emission inventories.

The data assimilation considerably improves the simulations by the model. With

each of the three assimilation methods, smaller differences and higher similarities between the simulated and observed concentrations were found, as shown in Table 7.3. The standard metrics that are used to compare the various algorithms showed an improvement compared to previous EnKF implementations assimilating the same observations (Lopez-restrepo *et al.*, 2021) (Chapter 4). This improvement is due to the better background obtained using WRF meteorology and the impact of the localization schemes present in the DA algorithms. Using the new assimilation schemes, the spatial distribution of concentrations within the valley is better resolved. Using a target covariance matrix to adapt the covariances computed from the ensemble results in betterrepresentation of the actual covariance structure. The target covariance matrix limits the influence of observations located in the lower part of the valley on the grid cells located in the hills of the valley, and vice versa. This makes it possible to separate the different regimes and avoids incorrect corrections in concentrations, as could occur with the standard LEKTF method. The forecast experiments also suggest a better estimate of the emission correction factors when shrinkage methods are employed. As a result, the forecasts of dangerous pollution levels is improved in all the stations (see Figure 7.7). These results encourage further improvement of these types of methods and to incorporate more and more prior knowledge in the covariance estimation. Possible new directions include dynamic target matrices dependent on the weather or on patterns in public behaviour.

Both shrinkage-based methods EnKF-KA and EnTLHF-KA, showed lower error statistics than the standard LETKF. The use of the shrinkage estimator and the incorporation of orography information through the $\mathbf{T}_{KA}$ matrix allows both methods to achieve satisfactory results with a relatively low number of ensemble members (25). Previous experiments in toy models (Lorenz96 and 2D advection-diffusion model) and real pseudo applications (SPEEDY model) have shown that the shrinkage-based family of methods can improve data assimilation when the size of the ensemble is small (Lopez-Restrepo *et al.*, 2021; Nino-Ruiz and Sandu, 2015) (Chapter5), supported by our results in a real high-dimensional application. This capability is important given the computational difficulty involved in generating many simulations of highly complex models. Althoughthe overall performance of both methods is similar, the robust method achieves better results, especially in stations on the slopes of the valley. The EnTLHF-KA algorithm tends to put more weight on the observations than the EnKF-KA in the analysis step due to the adaptive inflation term that is present. Additionally, the robust methods do not require a completely correct characterization of the observation representation errors or the uncertainties of the model (Luo and Hoteit, 2011). This characteristic benefits the EnTLHF-KA in our application, given the lack of precise information on the modeling system's uncertainties, e.g., emissions inventory, meteorology, composition, and reaction schemes.

Although the methods presented in this work were tested in a specific setting, their formulation is quite general and could be be used in other applications (Lopez-Restrepo *et al.*, 2021) (chapters 5 and 6). The basic concept of both EnKF-KA and EnTLHF-KA is to incorporate information or prior system knowledge that is not cap-

tured by the model directly in the data assimilation. In our case, for example, this principle works as a modification to the well-known concept of distance-based location. Several works have followed this line, mainly in history matching applications (Soares *et al.*, 2018; Lacerda *et al.*, 2021) but with a different approach. We believe that EnKF-KA and EnTLHF-KA possess sufficiently interesting characteristics to be applied and tested in areas other than that shown in this work.

7.6. Conclusions

We presented a data assimilation application using the shrinkage-based methods EnKF-KA and the robust EnTLHF-KA, with the chemical transport model LOTOS-EUROS over a densely populated valley. Both proposed methods outperform the standard LETKF, especially in places with complex orography. Incorporating the orography characteristics in the data assimilation through a target matrix, limits the influence of observations in grid cells that are far away in vertical distance. The final result can be understood as a localization scheme that does not depend only on the horizontal distance, but also on the change in orography. The robustness of the EnTLHF-KA allows to have a high similarity between the simulated and observed $PM_{2.5}$ concentrations, even with a small ensemble size and an incomplete representation of the system uncertainties. The model's forecasting capabilities are also improved, achieving a good representation of the concentrations on the first forecast day, being acceptable until the third day. After assimilation, the model is an accurate tool for forecasting alerts for high levels of air pollution.

References

G. Fu, F. Prata, H. Xiang Lin, A. Heemink, A. Segers, and S. Lu, *Data assimilation for volcanic ash plumes using a satellite observational operator: A case study on the 2010 Eyjafjallajökull volcanic eruption,* Atmospheric Chemistry and Physics **17**, 1187 (2017).

S. Lu, H. X. Lin, A. W. Heemink, G. Fu, and A. J. Segers, *Estimation of Volcanic Ash Emissions Using Trajectory-Based 4D-Var Data Assimilation,* Monthly Weather Review **144**, 575 (2016).

J. Jin, H. X. Lin, A. Heemink, and A. Segers, *Spatially varying parameter estimation for dust emissions using reduced-tangent-linearization 4DVar,* Atmospheric Environment **187**, 358 (2018).

S. Lopez-Restrepo, A. Yarce, N. Pinel, O. Quintero, A. Segers, and A. Heemink, *Forecasting PM10 and PM2.5 in the Aburrá Valley (Medellín, Colombia) via EnKF based data assimilation,* Atmospheric Environment **232**, 117507 (2020).

W. C. Skamarock, J. B. Klemp, J. Dudhi, D. O. Gill, D. M. Barker, M. G. Duda, X.-Y. Huang, W. Wang, and J. G. Powers, *A Description of the Advanced Research WRF Version 3*, Tech. Rep. (University Corporation for Atmospheric Research, 2008).

C. Misenis and Y. Zhang, *An examination of sensitivity of WRF/Chem predictions to physical parameterizations, horizontal grid spacing, and nesting options,* Atmospheric Research **97**, 315 (2010), arXiv:9809069v1 [gr-qc] .

D. Carvalho, A. Rocha, M. Gómez-Gesteira, and C. Santos, *A sensitivity study of the WRF model in wind simulation for an area of high wind energy,* Environmental Modelling and Software **33**, 23 (2012).

P. Tuccella, G. Curci, G. Visconti, B. Bessagnet, L. Menut, and R. J. Park, *Modeling of gas and aerosol with WRF/Chem over Europe: Evaluation and sensitivity study,* Journal of Geophysical Research Atmospheres **117**, 1 (2012).

X. M. Hu, P. M. Klein, and M. Xue, *Evaluation of the updated YSU planetary boundary layer scheme within WRF for wind resource and air quality assessments,* Journal of Geophysical Research Atmospheres **118**, 10490 (2013).

M. E. Dillon, Y. G. Skabar, J. Ruiz, E. Kalnay, E. A. Collini, P. Echevarría, M. Saucedo, T. Miyoshi, and M. Kunii, *Application of the WRF-LETKF Data Assimilation System over Southern South America: Sensitivity to Model Physics,* Weather and Forecasting **31**, 217 (2016).

A. Kumar, R. Jiménez, L. C. Belalcázar, and N. Y. Rojas, *Application of WRF-Chem Model to Simulate PM10 Concentration over Bogotá,* Aerosol and Air Quality Research **16**, 1206 (2016).

J. J. Henao, J. F. Mejía, A. M. Rendón, and J. F. Salazar, *Sub-kilometer dispersion simulation of a CO tracer for an inter-Andean urban valley,* Atmospheric Pollution Research **11**, 0 (2020).

C. D. Hoyos, L. Herrera-Mejía, N. Roldán-Henao, and A. Isaza, *Effects of fireworks on particulate matter concentration in a narrow valley: the case of the medellín metropolitan area,* Environmental Monitoring and Assessment **192**, 6 (2019).

S. Lopez-Restrepo, E. D. Nino-Ruis, A. Yarce, O. L. Quintero, N. Pinel, A. Segers, and A. W. Heemink, *An Efficient Ensemble Kalman Filter Implementation Via Shrinkage Covariance Matrix Estimation: Exploiting Prior Knowledge,* Computational Geosciences **25**, 985–1003 (2021).

E. Ott, B. R. Hunt, I. Szunyogh, A. V. Zimin, E. Kostelich, M. Corazza, E. Kalnay, D. Patil, and J. A. Yorke, *A local ensemble Kalman filter for atmospheric data assimilation,* Tellus **56**, 415 (2004).

S. Shin, J. S. Kang, and Y. Jo, *The Local Ensemble Transform Kalman Filter (LETKF) with a Global NWP Model on the Cubed Sphere,* Pure and Applied Geophysics **173**, 2555 (2016).

E. D. Nino-Ruiz and A. Sandu, *Ensemble kalman filter implementations based on shrinkage covariance matrix estimation,* Ocean Dynamics **65**, 1423 (2015).

A. M. Rendón, J. F. Salazar, and V. Wirth, *Daytime air pollution transport mechanisms in stable atmospheres of narrow versus wide urban valleys,* Environmental Fluid Mechanics **20**, 1101 (2020).

G. Gaspari and S. E. Cohn, *Construction of correlation functions in two and three dimensions,* Quarterly Journal of the Royal Meteorological Society **125**, 723 (1999).

N. J. Higham, *Computing a nearest symmetric positive semidefinite matrix,* Linear Algebra and Its Applications **103**, 103 (1988).

S. Lopez-restrepo, A. Yarce, N. Pinel, and A. W. Heemink, *Urban Air Quality Modeling Using Low-Cost Sensor Network and Data Assimilation in the Aburrá Valley , Colombia,* Atmosphere **12**, 1 (2021).

C. Mogollón-sotelo, L. Belalcazar, and S. Vidal, *A support vector machine model to forecast ground-level* $pm_{2.5}$ *in a highly populated city with a complex terrain,* Air Quality, Atmosphere & Health **14**, 399 (2020).

EPA, *Meteorological Monitoring Guidance for Regulatory Modeling Applications*, Tech. Rep. (U.S. ENVIRONMENTAL PROTECTION AGENCY, 2000).

J. W. Boylan and A. G. Russell, *Pm and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models,* Atmospheric Environment **40**, 4946 (2006), special issue on Model Evaluation: Evaluation of Urban and Regional Eulerian Air Quality Models.

X. Luo and I. Hoteit, *Robust Ensemble Filtering and Its Relation to Covariance Inflation in the Ensemble Kalman Filter,* Monthly Weather Review **139**, 3938 (2011), arXiv:1108.0158 .

R. Kohavi and F. Provost, *Applications of Machine Learning and the Knowledge,* Applications of Machine Learning and the Knowledge in Machine Learning **30**, 349 (1998).

R. V. Soares, C. Maschio, and D. J. Schiozer, *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 (2018).

J. M. Lacerda, A. A. Emerick, and A. P. Pires, *Using a machine learning proxy for localization in ensemble data assimilation,* Computational Geosciences , 11 (2021).

8

Conclusion

8.1. Discussion of research questions

This thesis further advanced ensemble-based DA techniques by elaborating upon four research questions proposed in Chapter 1. The thesis concludes the answers to the research questions by addressing specific points.

RQ1: How to integrate all the possible information captured in the Ensemble-based DA process (state value, parameters value and dynamic, etc.) into the forecast simulation of systems with high uncertainty?

The LE model showed a low forecast skill for the particulate matter (PM) dynamics in the Aburrá valley without DA. DA increase the accuracy of the forecast primarily for two reasons: first, if the simulation is initialized with an assimilated field value, the initial conditions at the beginning of the forecast window are closer to reality than the model alone can provide; second, it is possible to use improved parameters in the forecast window. In Chapter 3, different inheritance systems were tested to incorporate correction factors for emission parameters to improve the forecasts. The study concludes that given the cyclical nature with well-marked patterns in the particulate matter concentrations, using the latest estimated hourly profile of the emission correction factors resulted in the best results. The second conclusion is that for the case studied in this thesis, the inheritance of emission parameters has a more significant impact on the quality of the forecast than the initial condition of the forecast window. This is because the behavior of PM is governed mainly by emissions rather than initial concentrations. The impact of the estimated parameters against the initial condition in the forecast window is clearly shown in Chapter 3, where the forecast generated by the default inheritance method (initial condition of the forecast window from the last value estimated, but without using the estimated emission factors) presents the worst performance, converging with time to the forecast generated by the model without assimilation. It should be

noted that this behavior is specific to the application in question. For applications other than the one shown here, it is essential to conduct an exhaustive study to determine the impact of particular parameters and initial conditions.

RQ2: Can low-cost monitoring networks assimilated into a CTM be a more accessible alternative to standard air quality monitoring systems?

In Chapter 4 the accuracy of the Aburrá valley's low-cost sensor network is evaluated against the standard and official $PM_{2.5}$ monitoring network. Additionally, we tested the LE performance when low-cost observations are assimilated. The low-cost network evaluation concludes that low-cost devices are feasible to represent the temporal behavior and the magnitude of nearby official stations. Although low-cost observations do not provide high spatial representativeness such as official observations due to their locations and nature, the possibility of deploying a much denser network can make up for these deficiencies. The large number of low-cost sensors that can be deployed allows for more detailed spatial monitoring of concentrations than that obtained by an official monitoring network. In terms of DA, the results when assimilating this network outperform even those obtained when the official monitoring network is assimilated as shown in Chapter 4. We conclude that with the current advances in low-cost sensors, it is possible to use low-cost networks and DA to model and predict air quality in urban areas.

RQ3: How can a covariance localization scheme that uses direct knowledge of the system, for instance, a very complex topography, improve the performance of an Ensemble-based Data Assimilation method?

In Chapter 5 we proposed a method that allows incorporating previous knowledge or information of the system that is not well represented by the model, directly in the assimilation of data. EnKF-KA (Knowledge Aided) uses a shrinkage-based covariance estimator and a target matrix to guide the error covariance matrix's final structure. In Chapter 7, we implemented the EnKF-KA together with the LE to assimilate the low-cost observation network of the Aburrá valley. In this application, we configured the target matrix to represent the physical barrier in the transport of particulate matter that is present due to the valley's complex topography and that current meteorological fields do not capture very well. The target matrix used limits the influence of observations located in the lower part of the valley on the grid cells located in the hills of the valley, and vice versa. This makes it possible to separate the different regimes and avoid unrealistic corrections in concentrations, as could occur with standard localization techniques. Using a target covariance matrix to guide the assimilation results in better use of the available observations. The forecast experiments also suggest a better estimate of the emission correction factors when the shrinkage-based methods are employed.

RQ4: How does the performance of robust estimators compared to the EnKF under a scenario of high uncertainty sources like emissions, mete-

orology and observations?

We developed a robust version of the EnKF-KA based on the concept of the H_∞ filter in Chapter 6. The EnTLHF-KA uses an adaptive inflation factor dependent on the covariance matrix, which increases the method's robustness in the face of uncertainty in the observations, the model, and a low number of ensembles. The EnTLHF-KA also uses the shrinkage-based estimator to incorporate prior knowledge and information in the assimilation process. The proposed robust method was compared to existing methods under both known and controlled conditions using the Lorenz-96 model and outperformed them. The EnTLHF-KA has lower errors compared to the other approaches. When the number of ensemble members (10) is small, the shrinkage-based estimator approximates the background covariance matrix better. The non-Gaussian shrinkage estimator, in combination with the adaptive inflation factor, provides a higher robustness in applications where the ensemble has a non-Gaussian signature. Finally, the EnTLHF-KA was tested in the LE case study over the Aburrá valley in Chapter 7. The robust method was the most effective among the tested options. The EnTLHF-KA put more weight on the observations in the analysis step due to the presence of an adaptive inflation term. The robust methods do not require that the measurement errors or the model uncertainties be completely characterized. This is advantageous in our application, given the lack of information regarding the modeling system parameters such as the emissions inventory, meteorology, composition, and reaction schemes.

8.2. Outlook

In this thesis, we showed how, with DA, the performance and forecasting capacity of the LE model over the Aburrá valley could be considerably improved. Additionally, techniques were presented to improve DA by incorporating knowledge of the system and increasing the robustness. Some questions arose from the results, motivating future research.

All the experiments showed considered emissions as the only source of uncertainty. This assumption is based on the significant impact that emissions have on PM concentrations and the complexity of uncertainty in meteorological fields. The changes made in the modeling system incorporating the higher resolution meteorology of the WRF model showed a considerable improvement in the model's performance. To further improve the accuracy of particulate matter simulations, especially in applications that require high resolution, uncertainties in meteorology should be considered. This would allow incorporating the meteorological fields that have a high impact on the dynamics of pollutants in the DA system.

The use of observations from a low-cost network in the DA system made it possible to increase the spatial detail in the emission correction factor estimation. This capacity is of great relevance in applications such as the one shown here, in which a high resolution is necessary to capture the complexity of the domain. The low-cost network used in this study would actually a higher spatial and temporal resolution than that used by our application. With low-cost sensors, it is possible to deploy a network that monitors the particulate material's behavior at a spatial

resolution of even streets in an urban area and a sampling period of minutes. These characteristics make low-cost networks an exciting observation source for DA or validation of exposure models. Finally, the LE model results at high resolution (1 km $\times$ 1 km) can be used for exposure models at street level as initial or boundary condition, or as initial input to be spatially dis-aggregated following high resolution information.

We used a static target matrix in both shrinkage-based proposed algorithms. This target matrix's objective was to represent the complex topography of the valley and use it to guide the covariance estimation. Although for this specific application, an improvement was demonstrated when using this static matrix, the proposed methods do not present restrictions on the target matrix's temporal behavior. The use of dynamic objective matrices would allow, for example, to incorporate some observed wind patterns or changes in the influence of the observations based on meteorology. Additionally, future research could continue by developing methods that create target matrices using a more significant number of information sources. This would allow to refine the assimilation system and increase the understanding of the state covariance.

Acknowledgments

This thesis is the outcome of supports and contributions from many people who deserve a special mention. In the first place, I would like to thank my supervisors Prof. Dr. Ir. O.Lucia Quintero and Prof. Dr. Ir. Arnold W. Heemink, for the help, guidance, and support they offered me during my Ph.D. research, it was a great honor to work together. I would like to thank my co-supervisor Dr. Nicolas Pinel, for his advice and our fruitful discussions. I am in outstanding debt to Dr. Arjo Segers for all his help in my research's essential steps. My gratitude to Prof. Dr. Ir. Elias Nino-Ruiz for share with me his great ideas about data assimilation. I would like to thank my colleagues and friends Andres Yarce and John Hinestroza for all the support in the hard times and be always open for a conversation.

Working on my Ph.D. project was a enormous experience, therefore I would like to thank the people surrounding me, all my colleagues from the Department of Applied Mathematics and the Mathematical Modeling Group of EAFIT University, for the constructive seminar and help with the essential small details.

Last but certainly not least, I would like to thank my wife Angely for being the greatest pillar in my life; all this is yours as well as mine. To my family, my parents, and brother, all the efforts are for you. To my aunt, Raquel, and my Uncle Alfonso for all the support and contribution with the opportunity to start and complete my studies.

Santiago Lopez Restrepo
Delft, 2021

Curriculum Vitæ

Santiago LOPEZ RESTREPO

Santiago Lopez Restrepo was born on November 15, 1992, in Bello, Colombia. He obtained his Control Engineering BSc. and Chemical Engineering MSc. degree at the National University of Colombia in March 2015 and September 2017, respectively. In the same year, he started the Mathematical Engineering Ph.D. program at EAFIT University and the Ph.D. in the Institute of Applied Mathematics (DIAM), Delft University of Technology, sponsored by the Colombia Científica scholarship. His Ph.D. research focuses are theory and applications of data assimilation in atmospheric modeling and air quality. During his Ph.D., he worked as a lecturer of dynamical systems and mathematical modeling at EAFIT University. Additionally, he did a research stay in Universidad del Norte. From March 2021, Santiago is postdoctoral research in the Earth Science department at Vrije Universiteit Amsterdam.

List of Publications

In this thesis

1. **Santiago Lopez-Restrepo**, Andres Yarce, Nicolas Pinel, O.L Quintero, Arjo Segers, and A.W. Heemink. (2020). *Forecasting PM_{10} and $PM_{2.5}$ in the Aburrá Valley (Medellín, Colombia) via EnKF based Data Assimilation*, Atmospheric Environment **232**, 117507.

2. **Santiago Lopez-Restrepo**, Elias D Nino-Ruiz, Luis Guzman-Reyes, Andres Yarce, Nicolas Pinel, O.L Quintero, Arjo Segers, and A.W. Heemink. (2021). *An Efficient Ensemble Kalman Filter Implementation Via Shrinkage Covariance Matrix Estimation: Exploiting Prior Knowledge*, Computational Geosciences **25**, 985.

3. **Santiago Lopez-Restrepo**, Andres Yarce, Nicolas Pinel, O.L Quintero, Arjo Segers, and A.W. Heemink. (2021). *Urban Air Quality Modeling Using Low-Cost Sensor Network and Data Assimilation in the Aburrá Valley, Colombia*, Atmosphere **12**, 91.

4. **Santiago Lopez-Restrepo**, Andres Yarce, Nicolas Pinel, O.L Quintero, ..., and A.W. Heemink. (2021). *Data assimilation as a tool to improve Chemical Transport Models performance in developing countries* in IntechOpen *Air Quality*, ISBN 978-1-83968-786-0

5. **Santiago Lopez-Restrepo**, Elias D Nino-Ruiz, Luis Guzman-Reyes, Andres Yarce, Nicolas Pinel, O.L Quintero, Arjo Segers, and A.W. Heemink. (2020). *A Robust Ensemble-based Data Assimilation Method using Shrinkage Estimator and Adaptive Inflation*, Geophysical Research Letters (under review).

6. **Santiago Lopez-Restrepo**, Andres Yarce, Nicolas Pinel, O.L Quintero, Arjo Segers, and A.W. Heemink. (2020). *Using a robust data assimilation method to improve $PM_{2.5}$ modelling in the Aburrá Valley*, Atmospheric Environment (to be submitted).

Side Work and Conferences

1. Elias D Nino-Ruiz, Alfonso Mancilla-Herrera, **Santiago Lopez-Restrepo**, O.L Quintero. (2020). *A Maximum Likelihood Ensemble Filter via a Modified Cholesky Decomposition for Non-Gaussian Data Assimilation*, Sensors **20**, 877.

2. **Santiago Lopez-Restrepo**, Andres Yarce, Nicolas Pinel, O.L Quintero, Arjo Segers, and A.W. Heemink. (2017). *Challenges and opportunities for Open LOTOS-EUROS model to reproduce the Dynamics for Tropical Andes Domain*, 3th CMAS South America Conference, Vitoria-Brazil .

3. **Santiago Lopez-Restrepo**, Andres Yarce, Nicolas Pinel, O.L Quintero, Arjo Segers, and A.W. Heemink. (2019). *Lotos-Euros data assimilation for improving forecast of PM_{10} and $PM_{2.5}$ in the Aburrá Valley*, EGU General Assembly Conference 2019, Vienna-Austria.

4. **Santiago Lopez-Restrepo**, Andres Yarce, Nicolas Pinel, O.L Quintero, Arjo Segers, and A.W. Heemink. (2019). *Prediction of pollutant dynamics in the Aburrá Valley by data assimilation from the LOTOS-EUROS model*, Congreso Colombiano y Conferencia Internacional de Calidad del Aire y Salud Pública CASAP 2019, Barranquilla-Colombia, (in Spanish).