Data assimilation in a LOTOS-EUROS chemical transport model for Colombia using satellite measurements

Andrés Yarce Botero

Tesis Doctorado

Prof.dr.ir. A.W. Heemink Technische Universiteit Delft, promotor

Prof.dr. O.L. Quintero Montoya Universidad EAFIT, Colombia, promotor

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

DOCTORADO EN INGENIERÍA MATEMÁTICA

MEDELLÍN

2024

Data assimilation in a LOTOS-EUROS chemical transport model for Colombia using satellite measurements

Data assimilation in a LOTOS-EUROS chemical transport model for Colombia using satellite measurements

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
woensdag 7 februari 2024 om 12:30 uur

door

Andrés YARCE BOTERO

Master of Science Applied Physics,
Universidad EAFIT, Medellín, Colombia
born in Medellín, 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, Colombia, promotor
Onafhankelijke leden:	
Prof.dr.ir H.X. Lin	Technische Universiteit Delft
Prof.dr.ir. M. Verlaan	Technische Universiteit Delft
Prof.dr. R.G. Hanea	University of Stavanger and Equinor
Prof.dr E. Montilla	Universidad EAFIT
Prof.dr.ir M.B van Gijzen	Technische Universiteit Delft, reserve member
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, Satellite data assimilation, Low-cost in situ measurements

Author email: a.yarcebotero@tudelft.nl, ayarceb@eafit.edu.co

ISBN 978-90-834024-2-0

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

Dedicada a Dolly

Contents

Summary

When considering air quality, notably in South America, it seems that we are falling behind more developed regions in exacerbating the issue. This shortfall serves not just as observation, but as a warning, as air quality problems here are rapidly escalating. Nevertheless, by examining how other countries have addressed similar issues, we can prepare ourselves to tackle our own challenges. In this thesis we demonstrate how utilizing Data Assimilation DA we can reduce the uncertainty in some model uncertain parameters in an air quality model such as the LOTOS-EUROS Chemical Transport Model (CTM).

CTMs are critical for representing reality through numerical simulations of concentrations of atmospheric constituents. These models incorporate various processes, including emissions, transportation, chemical reactions, and deposition. It is imperative to use accurate models as they enable us to understand atmospheric processes better and develop effective solutions to environmental problems, more in regions with scarce measurenments. The LOTOS-EUROS model is employed, whereby the portrayal of reality accounts for uncertainty from multiple sources. To enhance model output, it is crucial to maximise the representativeness of the input information.

From a measurement perspective in Colombia, there is an evident scarcity of ground-level equipment to monitor air quality in a comprehensive manner. The principal urban areas are monitored but extensive regions remain unobserved. This is precisely where satellite data, together with cost-effective sensors, prove advantageous by offering a more comprehensive range. Satellite air quality data has become increasingly available and its temporal and spatial resolution improves. However, cloud coverage, particularly around the Andean mountains, often obstructs satellite observations. This dissertation uses TROPOMI satellite-derived NO_2 concentrations as the primary data source for assessing air quality in tropical regions. Furthermore, this thesis involved the development of a customised electronic hardware device specifically designed to collect in-situ measurements in a mountainous region, to compare models and perform remote data assimilation experiments. Data assimilation (DA) methods can be divided into two main categories: sequential and variational methods. Sequential methods introduce observations progressively as they become available. On the contrary, variational methods adopt a wider perspective by assimilating observations over a predetermined time frame and refining model accuracy through optimizing a cost function.

The 4D Var method is a noteworthy variational method that finds application in atmospheric sciences. The method employs an adjoint model that is a crucial component in enabling the optimization process through the computation of gradients that are vital for minimizing the cost function. The implementation of adjoint models, however, poses significant challenges, involving complex coding and main-

tenance requirements. These challenges are more pronounced when working in the area of Chemical Transport Model (CTM), where the goal is to significantly improve the physical modelling system based on an adjoint model that is not always available.

To overcome the hurdles related to adjoint models, this study explored adjoint-free data assimilation techniques. Adjoint-free methods, such as the 4DEnVar and the Local Ensemble Kalman Filter (LEnKF), employ ensemble propagation within the model to estimate variables, presenting a practical substitute. The research explored parameters that modulate emission model uncertainties as a means to reduce CTM-related uncertainties, given the focus on emissions as a significant contributing factor. Conducting experiments in various urban and rural locations in Colombia enabled a more nuanced comprehension of emission parameters.

The innovative use of ensemble-based data assimilation techniques, including the 4DEnVar and LEnKF, along with the incorporation of satellite observations, has substantially enhanced the refinement of emission parameters. The combination of chemical transport models (CTMs), satellite data, low-cost sensors, and data assimilation (DA) has led to significant progress in measuring atmospheric pollutants and forecasting emissions in Colombia. The integration of the LOTOS-EUROS model, improvements in satellite data processing, and alignment of sensors has substantially enhanced the region's atmospheric chemistry modelling capabilities in the region. Moreover, the implemented data assimilation techniques have proven effective in improving the precision of air quality models, strengthening the correlation between model projections and real-world observations.

Samenvatting

Bij het bestuderen van de luchtkwaliteit in Zuid-Amerika, lijkt het erop dat we daar achterblijven bij meer ontwikkelde regio's in het aanpakken van problemen. Dit is zorgwekkend omdat de problemen met de luchtkwaliteit hier snel toenemen. Desalniettemin kunnen we, door te kijken hoe andere landen soortgelijke problemen hebben aangepakt, leren hoe de uitdagingen in Zuid-Amerika aangegaan kunnen worden.

In deze dissertatie demonstreren we hoe het gebruik van Data Assimilatie (DA) ons kan helpen de onzekerheid in sommige modelparameters van het Chemical Transport Model (CTM) LOTOS-EUROS voor de Colombiaanse regio verminderd kan worden. CTM's bevatten diverse processen, waaronder emissies, transport, chemische reacties en depositie. Het is van essentieel belang om hierbij nauwkeurige modelparameters te gebruiken, om de atmosferische processen beter te begrijpen en om effectieve oplossingen voor milieuproblemen te ontwikkelen, vooral in regio's met weinig meetinformatie. Het LOTOS-EUROS model dat wordt gebruikt, houdt rekening met onzekerheid uit meerdere bronnen. Om de modeloutput te verbeteren, is het vooral cruciaal om de nauwkeurigheid van de emissie parameters in het model te verbeteren.

Vanuit het perspectief van metingen is er in Colombia een duidelijk gebrek aan grond apparatuur om de luchtkwaliteit op een uitgebreide schaal te monitoren. De belangrijkste stedelijke gebieden worden goed gemonitord, maar in veel uitgestrekte regio's zijn nu weinig meetgegevens beschikbaar. Dit is precies waar satellietgegevens en nieuwe, goedkope sensoren, een belangrijke rol kunnen spelen. Satellietgegevens van luchtkwaliteit worden steeds toegankelijker en verbeteren de temporele en ruimtelijke resolutie van de metingen. Echter, bewolking, met name rond de Andes, belemmert vaak de beschikbaarheid van satellietwaarnemingen. Deze dissertatie gebruikt TROPOMI-satelliet NO_2 concentraties als de primaire gegevensbron voor het beoordelen van de luchtkwaliteit. Daarnaast is ook op maat gemaakte elektronische meetapparatuur ontwikkeld, specifiek ontworpen om in situ metingen te verzamelen in bergachtige regio´s.

Data assimilatie (DA) methoden kunnen worden onderverdeeld in twee hoofdcategorieën: Sequentiële en variationele methoden. Sequentiële methoden gebruiken waarnemingen naarmate ze beschikbaar komen. Daarentegen worden bij variationele methoden alle beschikbare waarnemingen in een vooraf bepaalde tijdspanne tegelijk geassimileerd om de modelnauwkeurigheid te verbeteren via het optimaliseren van een kostfunctie. De 4DVar-methode is een methode die veel wordt toegepast in de atmosferische wetenschappen. Deze methode maakt gebruik van een adjoint model dat een cruciaal onderdeel is om het optimalisatieproces te versnellen. Dit adjoint model wordt gebruikt voor het efficiënt berekenen van de gradiënt van de kostfunctie, die nodig is voor het minimaliseren van de kostfunctie.

De implementatie van adjoint modellen brengt echter aanzienlijke uitdagingen met zich mee, waaronder een zeer complexe codering.

Om de hindernissen gerelateerd aan de ontwikkeling van adjoint modellen te overwinnen, zijn in deze dissertatie adjoint-vrije data-assimilatietechnieken gebruikt. Adjoint-vrije methoden, zoals de 4DEnVar en het Local Ensemble Kalman Filter (LEnKF), maken gebruik van een ensemble van model simulatie en vermijden daarmee het gebruik van een gelineariseerd model en een adjoint model. Het onderzoek in deze dissertatie richtte zich vooral op het schatten van onzekere emissie parameters. Er zijn experimenten uitgevoerd in verschillende stedelijke en landelijke regio´s in Colombia. Deze toepassingen van ensemble gebaseerde data assimilatie technieken, samen met het gebruik van satellietwaarnemingen, hebben het begrip van de emissies aanzienlijk verbeterd.

De combinatie van chemische transportmodellen, satellietgegevens, goedkope sensoren en data-assimilatie (DA) heeft geleid tot aanzienlijke vooruitgang bij het voorspellen van luchtverontreiniging in Colombia. Er zijn ook aanzienlijke verbeteringen geboekt op het gebied van de verwerking van satellietgegevens en de opzetten van een netwerk van goedkope sensoren in enkele Colombiaanse regio's. Bovendien zijn de geïmplementeerde data-assimilatietechnieken effectief gebleken bij het verbeteren van de nauwkeurigheid van de luchtkwaliteitsmodellen.

1

Introduction

... From phlogistic matter, I say, because, as noted above, pure phlogiston combined with common air seems to constitute another kind of air, namely, "noxious air."

Daniel Rutherford 1772

The early motivation for this research was the long-range transport of pollutants over Colombia. This problem could silently harm the natural areas with the constant deposition of pollutants that travel considerable distances from the emission source. We are interested in understanding how Colombia's principal cities contribute to contaminant deposition in faraway places and, more critically, areas with vulnerable ecosystems. The deposition of atmospheric contaminants like nitrogen dioxide (NO_2) induces chemical imbalances in natural ecosystems, with potential subsequent severe impacts, such as biodiversity loss.

The first approximation for this study was using Chemical Transport Models (CTM). CTMs use differential equations to describe the transport phenomena of pollutants; in this thesis, we use the LOTOS-EUROS (Long Term Ozone Simulation- European Operational Smog model), a Dutch model that simulates the physical dynamics of trace gases and aerosols, including emissions, chemical reactions, transport, and deposition.

The second approximation was the use of satellite information due to the lack of on-site measurements within Colombian territory. Satellite data is valuable for monitoring pollutant concentrations, transport, and deposition.

Parts of this chapter have been published in A. Yarce-Botero et al. (2021). Medellín Air Quality Initiative (MAUI). DOI: 10.5772/intechopen.97571.
About the MAUI project: https://www.eafit.edu.co/investigacion/grupos/modelado-matematico/proyectos/cofinanciados/Paginas/data-assimilation-2017-2020.aspx

1

A group of instruments, from the numerous instruments in orbit around the Earth, are designed to collect remote data for atmospheric composition studies. Furthermore, the usefulness and availability of this data is gradually increasing.

The third approximation was once preliminary simulations were conducted with the model, and the satellite measurements were incorporated into it to improve the model simulations in a process known as Data Assimilation (DA). The CTMs models are enhanced by feeding available real-life data, which leads to a better output representation and predictability of the CTM used once measured information is incorporated. Adjoint-free Data Assimilation techniques were implemented to estimate emission and other parameters in this thesis. This Introduction discusses the motivation and methodology used in this work and the objectives proposed.

1.1. Motivation

Colombia is a relatively large country with a population of around 50 million people and an area of 1,141,748 square kilometers (440,831 square miles). Colombia's natural protected areas are organized under the National System of Protected Areas www.parquesnacionales.gov.co: 59 protected areas as of 2018, covering 169,545 km^2 (65,462 sq mi) and accounting for more than 14 percent of the country's land area. An example of Colombia's unique high mountain ecosystem is the *páramo*, which provides water for many of Colombia's residents and is protected under Colombian law. The high mountain ecosystem of the Paramo and the native plant species Frailejón are shown in Figure 1.1.

Figure 1.1: Left image of a Páramo. It is a unique high-altitude ecosystem confined to the Andes of Ecuador, Peru, and Colombia. Páramo is located above the timberline (3000 m) and below the snowline (5000 m) and receives an average of 200 cm of precipitation annually. The right image shows one of the plants there presented, the Frailejones (*Espeletia hartwegiana*). Observers have described páramo as "grassland and bushes islands surrounded by a cloud forest sea."

Anthropogenic activities such as transport and thermoelectric generation are significant sources of reactive nitrogen (N_r) into the atmosphere [1], the component of primary interest during this work. NO_2 is a gas emitted by anthropogenic and natural sources as part of the family of nitrogen oxides NO_x($NO+NO_2$). NO_2 is emitted from anthropogenic (industrial activity, transport, and biomass burning) and natural (NO_x soil emissions and lighting) sources. Photochemical reactions of Nitrogen Oxide NO_x and Amonnia NH_3 create secondary inorganic aerosols [2] that can be transported over large distances [3]. Long-distance transport of secondary inorganic aerosols accounts for more than 8% of the reactive nitrogen flow on the planet in terrestrial ecosystems and constitutes a significant source of N_r in addition to the ocean [4].
Atmospheric transport and deposition are described as mechanisms that induce the flux of gases and particles to the land surface due to meteorological, chemical, and biological phenomena [5] and are responsible for the change in the balance concentration of different soil components and alter the dynamics of remote ecosystems [6, 7]. Deposition of atmospheric N_r alters oligotrophic ecosystems [2, 8], affecting

1

the distribution of communities of species [9–13] and the ecosystem's stability [14]. In Chemical Transport Models (CTMs), non-linear and stiff chemical interactions occur at periods that are often significantly shorter than transport time scales, making accurate modeling of the environment challenging [15] but necessary. Identifying natural and agricultural areas that may be excessively exposed to NO_2 deposition allows an estimation of the detrimental impact of these ecosystems. Unfortunately, most of the Colombian territory has no pollutant measurement networks. Hence, resorting to alternative sources of information, such as numerical models and satellite observations, is crucial. Furthermore, the constant interrelation between simulations and measured satellite data could eventually lead to an accurate description and prediction of the transport and deposition of the pollutants threatening fragile, large ecosystems in Colombia. The models have parameters that are modeled with a range of uncertainty because of the difficulty of mathematically representing some phenomena, making them inaccurate.

To reduce uncertainty and incorporate data in large-scale mathematical models, the project named *Data Assimilation Schemes in Colombian Geodynamics - Cooperative Research Plan Between Universidad EAFIT in Colombia and TUDelft in the Netherlands* was set in 2017. Several universities and state institutions' efforts aligned in search of the Colombian atmosphere's modeling dynamics in the framework of this project. **Medellin Air Quality Initiative** (MAUI) was created to understand this problem and suggest sensible decisions through scientific information. This project brought together local, regional, and international experts on subjects related to air quality and its impact on human and ecosystem health to establish a knowledge network to identify knowledge gaps to contribute to a deeper understanding of the local and regional scale of the impact of air pollution, motivating research collaborations, and inform the policymakers. [1]

Modeling the geodynamics of Colombia with a CTM requires considering the region's complex atmospheric dynamics. Colombia's unique location at the Intertropical Convergence Zone (ITCZ) confluence makes this phenomenon a primary driver of the transition between wet and dry seasons. Its rugged terrain, determined by the Andes mountain range system, creates a challenging scenario for a congruent representation of the spatio-temporal chemical distribution of the different atmospheric constituents. Quintero Montoya (2020) stated that the distinctive topography of the terrain gives rise to atmospheric conditions that degrade air quality in particular regions, notably in valleys, particularly during the shift from dry to rainy seasons. The unique geographical features in the tropical Andean region pose remarkable challenges for simulating the dynamics of atmospheric chemistry.

The uncertainty in parameters such as emissions and deposition velocities limits the accuracy of the model simulation of depositions. When the evolution of the states deviates from real-life expectations, for example, when it fails to describe how the concentration and deposition of different components of interest due to the transport of those from another area, we usually attribute this fault to the system parameters because we commonly do not discuss the physics or chemistry

[1] https://www.eafit.edu.co/investigacion/grupos/modelado-matematico/proyectos/cofinanciados/Paginas/data-assimilation-2017-2020.aspx

involved, for this reason, at the optimization step, we can ask: what parameters cause the difference between our behavior and the one we should have to match the best initial condition that makes the model perform the optimum evolution of the states of interest?

CTMs rely on environmental authorities to update emission inventories consistently. While this practice is standard in Europe and the United States, it often faces hurdles in Latin America. Here, the relevant authorities often struggle to provide timely updates on emission inventories, largely due to a mix of ethnographic and political factors. When relying on European emission inventories, it is essential to reassess both biogenic and anthropogenic emissions and to re-evaluate the current national inventory. However, the rapid urbanization of cities in Latin America, coupled with inadequate land use planning and a low number of air quality networks, means that these inventories are often outdated. In such scenarios, alternative methods such as top-down emission inventories - which estimate emissions based on satellite observations - become invaluable. When combined with remote sensing information and CTMs, these methods can provide a more accurate emission constraint.

The uncertainties in the emissions inventories lead to problems of over- or underestimation of the CTM prediction, which could be corrected by providing the model with the closest possible emissions parameterization. The assimilation of data from the high, medium, and low-cost and other ground sensor networks in Colombia has previously shown how the model representation of dynamics is improved in various orders of magnitude [16, 17] in cities such as Medellín, specifically for particulate matter pm2.5 and pm10 pollutants. In this case, the emission parametrization of other pollutants, such as nitrogen, has not been previously estimated, specifically from a top-down estimation perspective, by incorporating satellite information from satellites dedicated to air pollution monitoring into the CTM.

This work uses data assimilation (DA) techniques to integrate large-scale numerical models, satellite information, and data from different sources.DA techniques play a fundamental role in improving the system states that represent meteorological or chemical atmospheric conditions. This allows us to incorporate different real observations into the models, taking advantage of these two sources of information: observations and mathematical models. Due to the lack of in-situ measurements over much of Colombia, satellite information is key to understanding the faithful chemical representation of the atmosphere.

Real-life measurements in combination with a CTM can be used to describe the release of tracers and aerosols from various sources (anthropogenic, biogenic, marine, airborne dust, fires). In the emission module of the LOTOS-EUROS CTM (explained in detail in the next subsection), these sources are considered either as areas or point sources. The input data for the model are spatially gridded temporal values from average emissions databases. Time factors are defined per source category for the month of the year, the day within the week, and the hour within the day. Figure (1.2) shows an example of a city's daily anthropogenic emission factor.

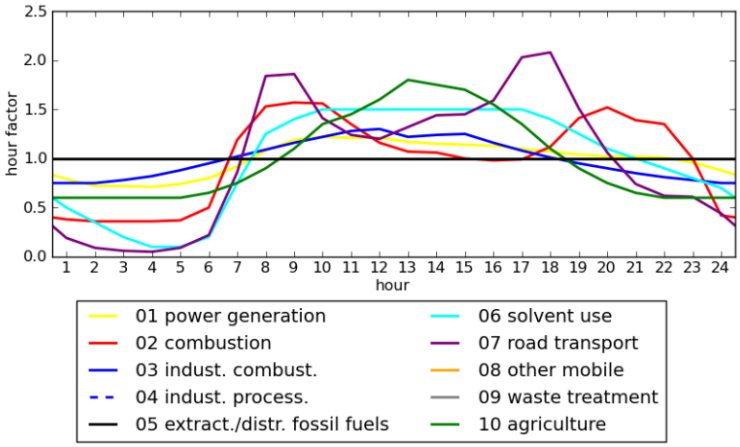


Figure 1.2: Daily cycle profiles of the emission factors applied to anthropogenic sectors for a typical city as presented in [18]. Different color curves correspond to different sources of pollutants, as indicated. Combustion and road transport have two pronunciated peaks corresponding to the busiest rush hours in the day when people are going and coming from house to work

Once emitted, the pollutant propagates to different places driven by the wind currents, then becomes dry/wet and deposited over areas that can be reached depending on its lifetime in the atmosphere and the magnitude and direction of the wind fields. We can improve the accuracy of modeling the long-range transport of pollutants from cities to natural areas in the country by adjusting the emission factors for specific points of interest using various data assimilation techniques. There is no way to measure these parameters directly from satellite observation, so methods are needed to estimate these indirectly. Through assimilation techniques and establishing relationships between what was emitted and observed, it was possible to determine parameters that suggest the correct emissions so that the model could approach the observed quantities. These results are presented in Chapters 4 of this thesis. In the following subsection, some of the mathematical details of the CTM used in this thesis are presented in detail.

1.2. The LOTOS-EUROS Chemical Transport Model

We aim to simulate the transport and deposition of pollutants with an accurate and robust numerical model. Modeling natural phenomena mathematically is a formidable challenge, as the intricacies of nature's complexity create inherent limitations on the representativeness attainable through a group of equations. In the particular case of transport and deposition of pollutants, one needs to consider the atmospheric chemistry concentration and its production/loss through the chemical reactions and depositions. CTMs receive real-life information from land use, emissions inventories, meteorology, orography, and atmospheric chemical boundary conditions to solve a numerical equation in discretized domains (parcels) to cal-

culate the concentration of the species of interest in a 3D space georeferenced that evolves in time (as represented in Figure 1.3). LOTOS-EUROS is the CTM[18, 19] used in this thesis to simulate Colombia's emissions, transport, concentrations, and depositions of pollutants (in particular NO_2 in Colombia).

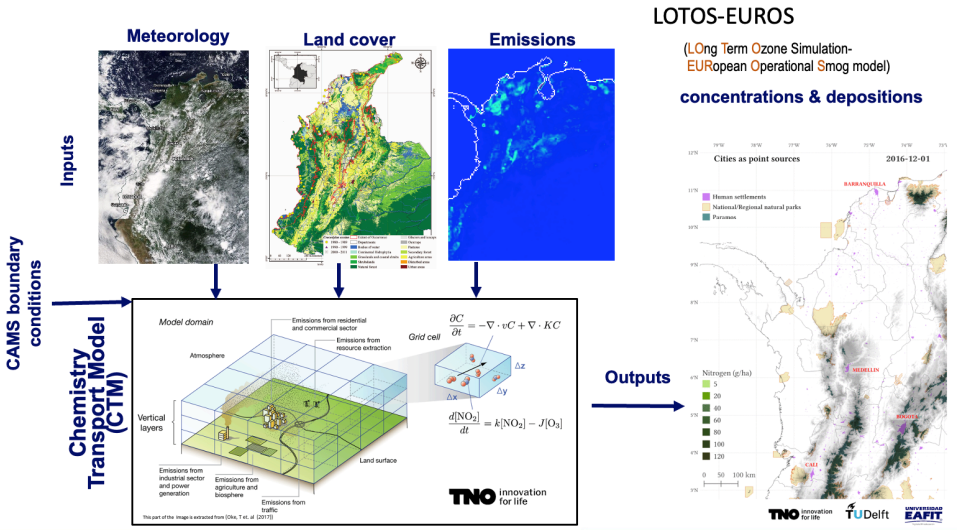


Figure 1.3: Schematic flow of the Chemical Transport Model LOTOS-EUROS implemented in Colombia. Input information (Chemical concentration boundary conditions, meteorology fields, land use maps and emission inventories) to solve the numerical transport propagation and chemical reactions to produce different gas concentration outputs.

LOTOS-EUROS (Long Term Ozone Simulation- European Operational Smog model) [20] is a large-scale, 3-D chemistry transport model simulating lower troposphere air pollution. Through the iterative numerical solution of Eq. (1.1), the model solves the concentration of the different components considered in the domain enclosed by each parcel.

$$\frac{\partial C_k}{\partial t} + U\frac{\partial C_k}{\partial x} + V\frac{\partial C_k}{\partial y} + W\frac{\partial C_k}{\partial z} = \frac{\partial}{\partial z}\left(K_z\frac{\partial C_k}{\partial z}\right) + E_k + R_k + Q_k - D_k - W_k \quad (1.1)$$

Here, C_k is the concentration of each pollutant involved in the model; U,V, and W are the large-scale wind components in West-East, North-South, and vertical directions, respectively; K_z is the vertical turbulent diffusion coefficient; Q_k represents the entrainment and detrainment of pollutants in the airflow due to variations in atmospheric layer height; R_k gives the amount of material produced or destroyed as a result of chemical reactions; E_k is the contribution by emissions to the pollutant concentration, and D_k and W_k are concentration loss terms due to processes

of dry and wet deposition respectively [18]. The LOTOS-EUROS central equation comprises different operators, each with different modeling components of the pollutant's behavior.
First, the transport operator consists of advection dynamics in three dimensions: horizontal and vertical diffusion and entrainment. Horizontal advection is described by horizontal winds (U, V) that are input from the model. Second, the chemistry operator models everything related to the production and consumption of components by different chemical reactions in the atmosphere. As a result of the complexity of LOTOS-EUROS, handling a complete mechanism of chemical reactions is unmanageable. To avoid this problem, LOTOS-EUROS can use one of two simplified reaction mechanisms: the Carbon Bond Mechanism CBMIV scheme, which consists of 81 reactions [20], and the aerosol mechanism, ISORROPIA parameterization [21]. Third, the dry deposition operator is parameterized following the mechanism known as the resistance approach [22]. Fourth, the process of the wet deposition operator is represented by the removal of gases from the below cloud [23].

The LOTOS-EUROS model can be driven with different meteorological inputs. This work is driven using ECMWF meteorology and is constrained by boundary conditions from the global MACC IFS/MOZART forecasts [24]. The anthropogenic emissions are prescribed following the EDGAR10 emission database for 2012 https://edgar.jrc.ec.europa.eu/. Biogenic emissions are calculated using the MEGAN model https://bai.ess.uci.edu/megan. The MACC global fire assimilation system [25] is used hourly to account for occasional fire events. The sea salt emissions are parameterized following source formulations for coarse [26] and fine [27] aerosol modes.
Due to the vast geographical diversity in the study area, the LE for Colombia was not correctly represented for its default model configurations. More representative input conditions were needed for the LOTOS-EUROS CTM. The LOTOS-EUROS model was updated for the surface fields that provide boundary conditions to the simulation parcels from the ground, the land use, and the topographic input.

CTMs simulate the atmospheric concentrations of chemical species and the exchange of components with the surface, even if, in most cases, these are imperfect representations of reality. With land use and topography update, vulnerable natural areas were identified by quantifying atmospheric deposition pollutants from the CTM simulations that can be more representative. There is uncertainty in their values due to the difficulty in obtaining this information and the fact that the elaboration processes are not in real-time, so the information is worked offline and with delays in emission values. Additional simulations were performed for the model configuration using point sources of pollutant emissions from the country's principal cities. This experiment consisted of controlled emissions from cells in those cities to identify the extent of spatial deposition of certain long-transport pollutants from the main Colombian cities.

This thesis is concentrated on Nitrogen dioxide (NO_2) because of the potential impact that its deposition may represent on ecosystems. Their role in acidification, eutrophication, and their impacts on the ecosystem and water quality are well doc-

umented [28–32]. The deposition problem will be revisited in Chapter 2, where different simulation experiments are shown. In the next section, different satellites used for data collection to measure pollutant concentrations are presented, with particular emphasis on the one used in this thesis (TROPOMI).

1.3. Satellites to monitor atmospheric composition

Different Low Earth Orbit (LEO) satellites have been carrying sensors to monitor NO_2 atmospheric concentrations since 1964, see Figure (1.4). From Paul J. Crutzen's idea of having spectrometers onboard satellites, an open field in the study of atmospheric chemical composition was created. Chronologically appeared the Global Ozone Monitoring Experiment (GOME, (1996-2003), GOME2A (2007-), GOME2B (2013-)) [33], [34], the Scanning Imaging Absorption Spectrometer for Atmospheric Cartography (SCIAMACHY, 2002e2012) [35], and the Ozone Monitoring Instrument (OMI) [36]. Each instrument generation has improved different data characteristics in the measured spectra, solving ratios and signal-to-noise problems and increasing spatial and temporal resolution.

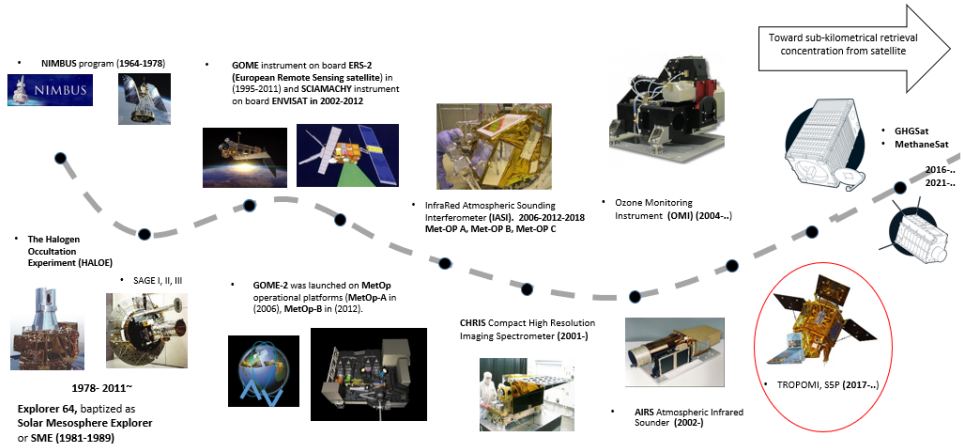


Figure 1.4: Timeline of the remote sensing instruments in orbit for measuring atmospheric composition. The TROPOMI instrument is in the red circle because it is the data of NO_2 used in this thesis.

TROPOMI is the instrument used in this thesis as the source of observed NO_2 for assimilation with the LOTOS-EUROS model. In Chapter 3, more general information about this satellite and the process of retrieving data from it is presented.

The NitroSat is a satellite mission that is in the planning phase. It could be the continuation of orbital spectrometers with improved resolution (500x500 m^2) [37]. Another mission coming is the TANGO mission, two CubeSat form factor satellites in development from SRON [38] to measure CO and NO_2. On the other hand, the deployment of a geostationary sensor observing network for air quality-related species is the next decade's milestone: the TEMPO (Tropospheric Emissions: Mon-

itoring of Pollution [39]), Sentinel-4 [40], and GEMS (Geostationary Environment Monitoring Spectrometer [41]) measurements will be directed to continuously observe North America, Europe, and East Asia, respectively. As a result, low-altitude polar-orbiting sensors will be required to examine the atmospheric composition of tropical and southern regions because the geostationary sensor observing network cannot monitor those regions. The following section presents the theoretical background of Data Assimilation (DA) as a tool to improve this imperfect representation using satellite data and other measurements.

1.4. Data Assimilation

Data assimilation (DA) is a process in which observations of the actual system are incorporated into the model states using the residuals (differences between the model states and the observations) as small as possible [42]. DA is used in meteorological areas to produce a regular, physically consistent representation of the atmosphere from a heterogeneous collection of in situ and remote instruments that imperfectly and irregularly cover the domain of interest in space and time [43].
DA methods for CTMs are inspired mainly by meteorological DA experiences [44]. Many successful applications have demonstrated the benefits of assimilation for CTMs, either to produce reanalysis fields and forecasts or to focus on improving the accuracy of model inputs (initial conditions, boundary conditions, emissions) [45]. A common characteristic of these applications is that in regional air-quality simulations, the influence of initial conditions quickly fades over time, as emissions and lateral boundary conditions primarily determine the model fields [46].
In Colombia, DA methods with CTMs have not been implemented at the regional scale; in the local scale, the case studio of particulate matter in Medellin [47] constitutes the init of data assimilation activities in this territory. The in-depth knowledge of atmospheric dynamics that these techniques provide is essential because it provides the region with accurate forecasting capabilities, accurate simulation of pollutant transport trends, and estimation of process dynamics such as emissions and deposition, which are useful for understanding the impact of long-range transport of pollutants on natural ecosystems.
Variational and sequential methods constitute the two primary classes of DA techniques, with variational methods centering on minimizing a cost function within a designated time window to encompass the available observations. This cost or objective function is generally depicted as the sum of squared discrepancies between the data and the corresponding model values. The sequential methods, on the other hand, provide the best representation of reality by combining both sources of information each time the observations are available. The Ensemble Kalman Filter (EnKF) is the name of the DA technique that can be used when the model is nonlinear [48]. From the Variational method, there are two methods for incorporating observations in one specific time or a time window: 3DVar and 4DVar. The 3DVar method uses a static, flow-independent, background error covariance that is often spatially homogeneous and anisotropic [49]. The 4DVar method allows fitting the model forecast trajectory to observations distributed over a time window to provide more accurate model state estimations that are also more consistent with the

prediction model [50].

A cornerstone of the 4D-Var method is the adjoint model, which is utilized to compute gradients required for the optimization process that minimizes the cost function. In the 4D-Var framework, data assimilation is morphed into an optimization issue resolved using gradient-based optimization methods, where the 4D-Var gradient is derived by enforcing the adjoint model with observation increments. The formulation of the adjoint model necessitates significant developmental effort, underscoring its complexity and the intricacy of the variational approach to data assimilation [51]. Through adjoint methods, the adjoint model assists in incorporating measured observations into a dynamical system model, aiming for accurate estimations of the system's current and future state variables. The optimal estimates, minimizing a variational principle, are achieved using adjoint methods, with model equations acting as stringent constraints on the problem [52]. The adjoint method is recognized for its capability to adjust unknown control parameters based on the provided data, including model initial conditions or other model parameters and inputs[53]. Through various scenarios and sets of observations, the 4-D variational approach demonstrates its efficiency in extracting information from the model's dynamics in conjunction with the information contained in the observations, thereby enhancing the accuracy and consistency of model state estimations [54]. The applicability and centrality of adjoint operators are evident in numerous operational data assimilation systems, notably in numerical weather prediction, and are progressively gaining traction in oceanography, underlining the broad-based utility and the evolving adoption of adjoint methods for data assimilation [55]

The use of the adjoint perspective in data assimilation is a potent tool; however, it comes with certain limitations. The main drawback lies in the intricate and resource-demanding nature of constructing and maintaining adjoint models, which requires a profound grasp of mathematical fundamentals and considerable developmental effort. This intricacy also encompasses the creation of adjoint code, which is a nuanced task fraught with potential pitfalls. Furthermore, the established techniques require thorough validation for the precise and efficient functioning of the adjoint models, which could consume significant time and resources. Furthermore, the methodology mandates a methodical mistake covariance model to correctly present the system's uncertainties, posing a formidable challenge to obtain. The inflexibility of the adjoint model's structure can occasionally limit its ability to handle non-linear dynamics or integrate new data sources, which could impede the development of data assimilation tactics. While the adjoint perspective is vital in enhancing model state estimations, its challenges often require a combination of skill, time, and computational resources that may not always be accessible.

An essential difference between Variational and EnKF is how the results are delivered. Variational provides an assimilated result in the form of an optimal model evaluation; EnKF provides the result in terms of mean and covariance [42]. Both methods are suitable for online forecast applications; for offline applications such as parameter estimation, the Variational method is often favored. Compared to the Variational methods, the EnKF is generally quite simple to implement since only the forward model must be used, and there is no need for an adjoint model [56].

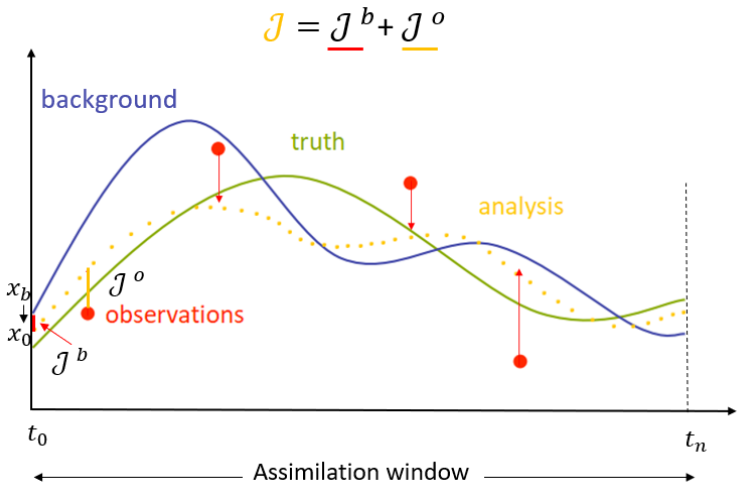


Figure 1.5: Time evolution of the model to present the parts that compose the cost function $\mathcal{J}$ in the variational assimilation method. This cost functions $\mathcal{J}$ is established by the distance from the background initial condition (blue curve) $\mathcal{J}^b$ and the distance of the model values from observations (red disks) $\mathcal{J}^o$. These two contributions encompass the analyzed cost function (dotted yellow curve). The green curve corresponds to the truth, which is one model propagation the modeler assumed as the real value where synthetic observations are sampled in a twin experiment procedure.

Variational DA methods are based on minimizing a cost function $\mathcal{J}$ to find the optimal initial condition that minimizes the distance between the background and the initial condition and the observations and the analysis. Each cycle is repeated at each analysis step for a window of time, searching for an appropriate initial condition that produces the best functional analysis. Figure 1.5 shows one assimilation window where observation helps to calculate a new initial condition state x_0 to produce a new trajectory known as the analysis. This process can also be used to calculate parameters, as we use it to estimate the emission parameters in this thesis. Variational assimilation is based on optimal control theory derived from the calculus of variations. The advantage of variational methods is that they address the entire phenomenology and provide an optimal solution consistent with the model dynamics, the assimilation window based on a prior state (the background or forecast), the observations, and the prescribed Gaussian uncertainty statistics for the background and observations.[57]. In contrast, the sequential approach is progressively updated by reconciling the state with the uncertainties defined for the simulated state and the observations.[48, 58, 59].Since no randomness is involved in the state's evolution, variational approaches assume that the underlying model is deterministic. The parameter estimation problem is treated as an optimization problem using variational techniques, where a cost function $\mathcal{J}$ penalizes the difference between model simulations and observations and the variation of the

parameters from their initial value. The vector that minimizes $\mathcal{J}$ is the best estimate of the parameter vector.

A state space is a mathematical depiction of a physical system comprising input, output, and state variables interconnected through first-order discrete equations. It serves as a mathematical model wherein state variables determine the values of output variables. Let the element vectors $\mathbf{x}_t \in \mathbb{R}^{n\times 1}$ represent a state of the model at time t where n is the number of states. The state $\mathbf{x}_t$ is found by propagating $\mathbf{x}_{t-1}$ forward by one time step by the model operator $\mathcal{M}_{t-1,t}$:

$$\mathbf{x}_t = \mathcal{M}_{t-1,t}(\mathbf{x}_{t-1}) + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \mathbf{Q}_t), \tag{1.2}$$

With $\mathbf{x}_0$ as the initial condition and where η_t is an *n*-element model uncertainty that is normally distributed with an error covariance matrix $\mathbf{Q}_t$. Parameters control the temporal evolution of the system and can be included in the state $\mathbf{x}_0$. This thesis focuses on emission parameters. Given the parameters, the evolution of the state vector over time can be calculated. In data assimilation, this evolution is called the trajectory of $\mathbf{x}_t$.

$$\mathbf{y}_t = \mathbf{H}_t\mathbf{x}_t + \nu_t, \qquad \nu_t \sim \mathcal{N}(0, \mathbf{R}_t) \tag{1.3}$$

The set of observations $\mathbf{y}_t$ made at time *t* are related through the observation operator $\mathbf{H}_t$ from the model space $\mathbf{x}_t$ to the observation space $\mathbf{y}_t \in \mathbb{R}^{m\times 1}$ where m is the number of observations. The linear observation operator is denoted $\mathbf{H}_t$ and ν_t is an *m*-element uncertainty that denotes the observation representation error that is Gaussian distributed with mean 0 and covariance matrix $\mathbf{R}_t$.

In 4DVar we consider a deterministic system with $\mathbf{Q}_t = 0$, known as strong constraint 4DVar with no model uncertainty. The 4DVar problem can be posed as a cost function $\mathcal{J}$, which is a function represented by the sum of terms or distances between components of the system space.

$$\mathcal{J}(\mathbf{x}_0) = \underbrace{\frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^T\mathbf{P}_0^{-1}(\mathbf{x}_0 - \mathbf{x}^b)}_{\mathcal{J}^b} + \overbrace{\frac{1}{2}\sum_{t=0}^{T}(\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t)^T\mathbf{R}_t^{-1}(\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t)}^{\mathcal{J}^o}, \tag{1.4}$$

The first term is known as the background term $\mathcal{J}^b$, and it measures the difference between the first guess $\mathbf{x}^b$ and the estimated state $\mathbf{x}_0$. This difference is calculated with the norm $\mathbf{P}_0$ (prior information covariance). In this case, we have the cost function for the initial condition $\mathbf{x}_0$. The second term,$\mathcal{J}^o$, measures the deviation between the measured states $\mathbf{y}_t$ and $\mathbf{H}_t\mathbf{x}_t$ weighted with the observation-error covariance matrix $\mathbf{R}_t$ over the assimilation window.

The four-dimensional variational method or the 4DVar method allows the fitting of the model forecast trajectory to observations distributed over the assimilation window $[t = 0, .., t = T]$ to provide more accurate model state estimations that are also more consistent with the prediction model [50]. The minimization of this cost

function requires the adjoint model operator, which transports the sensitivity fields to estimate the initial conditions and can be interpreted as a backward propagation of the model from one moment t to the initial time 0. One of the disadvantages of the method is that implementing the adjoint model is often very complicated for nonlinear models. Especially implementing the linearized model of a complex, existing numerical model can be very hard and depends on the model version evolution for which the adjoint model must be updated too, making the adjoint effort expensive [60]. Only for a very structured model code can automatic differentiation tools be used to generate the linearized model code and the adjoint model code by a computer. This thesis uses adjoint-free methods for different data simulation implementations in Colombia.

The EnKF method, depicted in Figure 1.6 is the other DA method used in this thesis that does not use an adjoint of the model [61, 62]. It is known as sequential, meaning that the information is continuously ingested once presented. The EnKF approach updates the model state sequentially based on each new observation, incorporating it into the state estimate from the previous assimilation step. Sequential methods typically consist of two main steps: the forecast step, which propagates the model forward in time, and the analysis step, which adjusts the model state based on the new observation and propagates the new model simulation or analysis trajectory, which is not always the outcome of a model simulation. The physical relationship between emissions and observed NO_2 concentrations is important for estimating emission parameters. Such a physical relationship fundamentally entails mass conservation, which the variational technique preserves.

In ensemble-based methods, an ensemble of N model realizations and n states is generated :

$$\mathbf{X}_t^f = [\mathbf{x}_t^{f[1]}, \mathbf{x}_t^{f[2]}, .., \mathbf{x}_t^{f[N]}] \in \mathbb{R}^{n \times N} , \tag{1.5}$$

Which correspond to the different realizations of the model generated by perturbing the model uncertainty assumed to be normally distributed:

$$\eta_t \sim \mathcal{N}(0, \mathbf{Q}_t) . \tag{1.6}$$

The mean $\overline{\mathbf{x}}_t = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_t^{f(i)}$ and the covariance background error matrix is:

$$\mathbf{B}_t^f = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_t^{f(i)} - \overline{\mathbf{x}}_t)(\mathbf{x}_t^{f(i)} - \overline{\mathbf{x}}_t)^T \in \mathbb{R}^{n \times n} . \tag{1.7}$$

The EnKF analysis step is applied to each member of the ensemble update:

$$\mathbf{x}_t^{a(i)} = \mathbf{x}_t^{f(i)} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t \mathbf{x}_t + \nu_t^i), \tag{1.8}$$

with the optimal gain $\mathbf{K}_t$ defined by the expression:

$$\mathbf{K}_t = \mathbf{B}_t^f \mathbf{H}_t^T [\mathbf{H}_t \mathbf{B}_t^f \mathbf{H}_t^T + \mathbf{R}_t]^{-1} , \tag{1.9}$$

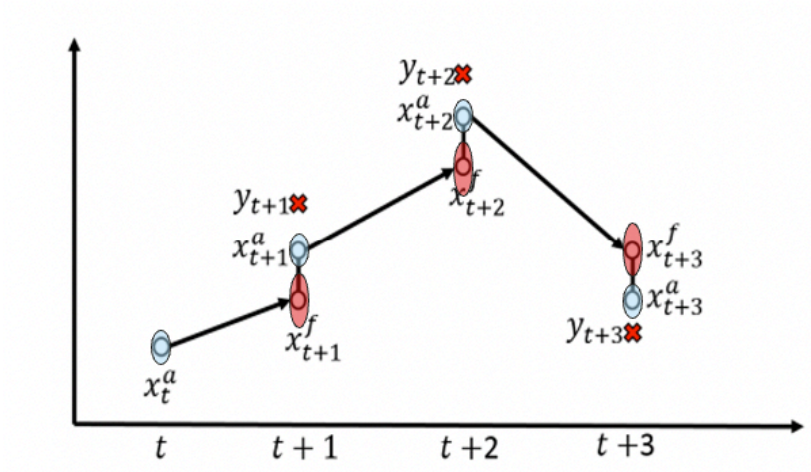


Figure 1.6: A prediction is issued from the analysis until the next observations at $t+1$; At time $t+1$, a new analysis is performed by combining the forecast and the observations at $t+1$. The process is repeated from the analysis at $t+1$, and a new forecast until $t+2$ is issued, and a new analysis is performed using the observation at $t+2$.In blue and red, the uncertainty of the analysis and forecast ensemble members are depicted

The matrix $\mathbf{B}_t^f$ is not computed and $\mathbf{K}_t$ is calculated directly from the model ensemble. Since Geir Evensen's innovative introduction in 1994 [61], the Ensemble Kalman Filter (EnKF) has occupied a pivotal position in geophysical data assimilation. Based on Monte Carlo techniques, the EnKF approach involves propagating an ensemble of model states over time, integrating observational data at designated intervals. This paradigm provides a probabilistic framework for state estimation, crucial for capturing inherent uncertainties in such complex systems [63]. Numerous variations of the EnKF have been developed over the years, a testament to its versatility and wide applicability. Over the years, the foundational principles of the EnKF have been adapted and extended, leading to a suite of variations tailored to different challenges. Other noteworthy variations include the Deterministic Ensemble Kalman Filter (DEnKF) [64], which offers a deterministic perspective on ensemble updates, and the Ensemble Transform Kalman Filter (ETKF), which transforms the ensemble members directly in the observation space, and the LEnTKF that has garnered attention [65]. Yet another variant, the Stochastic EnKF, incorporates model and observational uncertainties more explicitly by introducing stochastic perturbations. The proliferation of these versions underscores Evensen's original concept's adaptability and foundational strength.

Both DA perspectives reach similar accuracy levels if long assimilation DA windows are used for 4DVar [66] and can be used for parameter estimation by augmenting the state vector. Regarding implementation, 4DVar depends on the tangent linear and adjoint model to solve the minimization procedure, making this

method less easy to implement due to the common unavailability of this adjoint model for a highly complex model. This is a considerable drawback of variational DA because adjoints are prohibitive for large-scale models ($\sim 10^6 - 10^9$ state elements) and very expensive to maintain [44]. Hybrid approaches tackle this problem, aiming to take advantage of the characteristics of each method [59, 67]. Ensemble-based approaches avoid the construction of tangent linear and adjoint representations of the forecast model (e.g., [68]), which is interesting for low-budget operational scenarios using numerical models. This thesis will first focus on the implementation of an adjoint free variational DA (4DEnVar) and the sequential implementation of EnKF, which also does not require the adjoint model.

1.4.1. Covariance localization

Due to the approximation of the state-space covariance by a finite number of ensemble members, spurious correlations between state items are unavoidable. These erroneous connections can be eliminated using a technique known as localization [69]. Covariance localization [70] is the localization method employed in this study. The covariance localization, also known as the Schur localization, focuses on the forecast error covariance matrix, removing correlations in error covariances after a specific distance [70, 71]. The localization is accomplished by a pointwise multiplication known as a Schur product and is denoted by $\circ$, where i and j are the row and column matrix members:

$$[\mathbf{f} \circ \mathbf{B}_t^f]_{i,j} = [\mathbf{B}_t^f]_{i,j} \cdot [\mathbf{f}]_{i,j} \tag{1.10}$$

The Schur product says that if $\mathbf{f}$ and $\mathbf{B}_t^f$ are positive semidefinite, then the Schur product, $\mathbf{f} \circ \mathbf{B}_t^f$, is positive semi-definite too. A cutoff function $G(r/\rho)$ to fill $\mathbf{f}$, where r is the Euclidean distance between two state members and ρ is a length scale called the localization radius [72]. The cutoff function used in this thesis has the following form:

$$[\mathbf{f}]_{i,j} = \exp(-0.5 \cdot (r_{i,j}/\rho)^2) \tag{1.11}$$

This results in a matrix structure with a band of non-zero elements surrounding the main diagonal, with ones on the diagonal and zeros at the localization radius distance from the diagonal. The covariance localization method attempts to eliminate misleading correlations in error covariances over a long distance, thereby improving background error covariance estimates.

This localized covariance matrix $\mathbf{f} \circ \mathbf{B}_t^f$ is used in the EnKF analysis and in the generation of the posterior ensemble of perturbations as a replacement for $\mathbf{B}_t^f$:

$$\mathbf{K}_t = (\mathbf{f} \circ \mathbf{B}_t^f)\mathbf{H}_t^T[\mathbf{H}_t(\mathbf{f} \circ \mathbf{B}_t^f)\mathbf{H}_t^T + \mathbf{R}_t]^{-1}\ , \tag{1.12}$$

1.5. Objectives of the research

The objectives of this thesis are as follows:

Development of a CTM for the Colombia region using the LOTOS-EUROS modeling system.

The main goal is to create a Chemical Transport Model (CTM) for the Colombia area utilizing the LOTOS-EUROS modeling system. It is important to collect additional representative data for the CTM to work, specifically focusing on land use and topography updates. Simulations shall be conducted, considering both point sources and pollutant emissions from major cities across Colombia. The objective is to identify natural areas that are vulnerable to atmospheric pollutant deposition. The research shall involve quantifying pollutant levels per unit area in selected natural protected areas, thereby seeking a clearer understanding of the spatial range of atmospheric deposition pollutants. Further simulations will be conducted to assess the spatial extent of deposition areas, aiming for a more detailed comprehension of pollutant dispersion throughout the region.

Evaluate the Potential Benefits of Satellite Measurements for Enhancing Emission Data within the LOTOS-EUROS Model in the Colombian Region.

This study aims to evaluate how satellite measurements can enhance emission data within the LOTOS-EUROS model in Colombia. Since there is incomplete data coverage across the Colombian territory and insufficient air quality networks in major cities, utilizing remote satellite data is imperative. With a history of satellites devoted to studying atmospheric composition and the progressively enhanced spatial and temporal resolution of orbits, the primary objective is to incorporate this data into the model. The process of data assimilation, which involves a mathematical method of reconciling model representation with measurements, will be utilized. However, the integration of satellite products is complex and requires sufficient calibration to accurately align the observed state with the model, as shown through the applied methodology.

Evaluate the Impact of a Low-Cost Sensor Network on Improving Emission Data in the LOTOS-EUROS Model for the Colombia Region.

The objective is to assess the effect of a newly developed low-cost sensor network on improving emission data in the LOTOS-EUROS model, with a focus on Colombia. The customized hardware will act as both a ground node and a gateway in air quality networks, as Chapter 3 will explain. This chapter will detail the construction, testing, and challenges of creating a device suitable for conducting on-site air quality measurements and long-distance data transmission. This project aims to address the data scarcity issue in Colombia by providing new data sources for the Data Assimilation of LOTOS-EUROS. A significant milestone in advancing this thesis, this venture will present opportunities for tackling prevalent data scarcity issues in the region.

To use adjoint-free data assimilation techniques to estimate uncertain parameters within the LOTOS-EUROS Model.

The aim is to estimate uncertain parameters within the LOTOS-EUROS model by employing adjoint-free data assimilation methods, as this system has no adjoint model. Chapter 3 and Chapter 4 will showcase the implementation results of two methods that avoid using the adjoint model through an ensemble.The LOTOS-

EUROS model will estimate NOx emissions in a region covering Colombia, Venezuela, and Ecuador, using TROPOMI NO2 observations. The Ensemble Kalman Filter (EnKF) and 4D-Ensemble-Variational (4DEnVar) methods will be employed to achieve this. Avoiding linearized and adjoint model implementations, EnKF and 4DEnVar will employ an array of model simulations to approximate the background covariance matrix, non-linear model, and observation operator with the goal of achieving a more precise estimation of uncertain parameters within the model.

1.6. Aim and structure of this thesis

In this thesis, we integrate satellite data into the LOTOS EUROS Chemical Transport Model via data assimilation in Colombia. This results in a substantial improvement in the predictive accuracy of LOTOS EUROS for this area outlined in Chapter 2 where we detail the modeling phase, including updating model characteristics such as land use and topography. In Chapter 3, the observational section is discussed, outlining the implemented satellite information procedure and the creation of electronic hardware for measuring pollutants, and in Chapter 4, we present the adjoint free data assimilation method in which we show how this technique works for estimating emission parameter over the territory domain. Finally, in the last section 5.3, a recommendation for the estimation with the data assimilation technique of other parameters that introduce uncertainty to the system is presented.

References

[1] D. Fowler, M. Coyle, U. Skiba, M. A. Sutton, J. N. Cape, S. Reis, L. J. Sheppard, A. Jenkins, B. Grizzetti, J. N. Galloway, *et al.*, *The global nitrogen cycle in the twenty-first century,* Philosophical Transactions of the Royal Society B: Biological Sciences **368**, 20130164 (2013).

[2] J. W. Erisman, A. Vermeulen, A. Hensen, C. Flechard, U. Dämmgen, D. Fowler, M. Sutton, L. Grünhage, and J. P. Tuovinen, *Monitoring and modelling of biosphere/atmosphere exchange of gases and aerosols in Europe,* Environmental Pollution **133**, 403 (2005).

[3] Y. Jia, G. Yu, Y. Gao, N. He, Q. Wang, C. Jiao, and Y. Zuo, *Global inorganic nitrogen dry deposition inferred from ground- and space-based measurements.* Scientific Reports **6**, 19810 (2016).

[4] R. A. Duce, J. LaRoche, K. Altieri, K. R. Arrigo, A. R. Baker, D. G. Capone, S. Cornell, F. Dentener, J. Galloway, R. S. Ganeshram, R. J. Geider, T. Jickells, M. M. Kuypers, R. Langlois, P. S. Liss, S. M. Liu, J. J. Middelburg, C. M. Moore, S. Nickovic, A. Oschlies, T. Pedersen, J. Prospero, R. Schlitzer, S. Seitzinger, L. L. Sorensen, M. Uematsu, O. Ulloa, M. Voss, B. Ward, and L. Zamora, *Impacts of atmospheric anthropogenic nitrogen on the open ocean,* Science **320**, 893 (2008).

[5] M. Giardina and P. Buffa, *A new approach for modeling dry deposition velocity of particles,* Atmospheric Environment **180**, 11 (2018).

[6] S. M. Bard, *Global transport of anthropogenic contaminants and the consequences for the arctic marine ecosystem,* Marine Pollution Bulletin **38**, 356 (1999).

[7] D. Fowler, J. N. Cape, M. Coyle, C. Flechard, J. Kuylenstierna, K. Hicks, D. Derwent, C. Johnson, and D. Stevenson, *The global exposure of forests to air pollutants,* Water, Air, and Soil Pollution **116**, 5 (1999).

[8] J. W. Erisman, J. N. Galloway, S. Seitzinger, A. Bleeker, N. B. Dise, A. M. R. Petrescu, A. M. Leach, and W. de Vries, *Consequences of human modification of the global nitrogen cycle.* Philosophical transactions of the Royal Society of London: Series B, Biological Sciences **368**, 20130116 (2013).

[9] R. Bobbink, K. Hicks, J. Galloway, T. Spranger, R. Alkemade, M. Ashmore, M. Bustamante, S. Cinderby, E. Davidson, F. Dentener, B. Emmett, J.-W. Erisman, M. Fenn, F. Gilliam, a. Nordin, L. Pardo, and W. De Vries, *Global assessment of nitrogen deposition effects on terrestrial plant diversity: a synthesis.* Ecological applications : a publication of the Ecological Society of America **20**, 30 (2010), arXiv:9403058 [gr-qc] .

[10] E. C. Farrer and K. N. Suding, *Teasing apart plant community responses to N enrichment: the roles of resource limitation, competition and soil microbes,* Ecology letters **19**, 1287 (2016).

1

[11] L. C. Maskell, S. M. Smart, J. M. Bullock, K. Thompson, and C. J. Stevens, *Nitrogen deposition causes widespread loss of species richness in British habitats,* Global Change Biology **16**, 671 (2010).

[12] S. M. Simkin, E. B. Allen, W. D. Bowman, C. M. Clark, J. Belnap, M. L. Brooks, B. S. Cade, S. L. Collins, L. H. Geiser, F. S. Gilliam, S. E. Jovan, L. H. Pardo, B. K. Schulz, C. J. Stevens, K. N. Suding, H. L. Throop, and D. M. Waller, *Conditional vulnerability of plant diversity to atmospheric nitrogen deposition across the United States,* Proceedings of the National Academy of Sciences of the United States of America **113**, 4086 (2016).

[13] C. J. Stevens, N. B. Dise, J. O. Mountford, and D. J. Gowing, *Impact of Nitrogen Deposition Grasslands,* Science **303**, 1876 (2004).

[14] S. E. Koerner, M. L. Avolio, K. J. La Pierre, K. R. Wilcox, M. D. Smith, and S. L. Collins, *Nutrient additions cause divergence of tallgrass prairie plant communities resulting in loss of ecosystem stability,* Journal of Ecology **104**, 1478 (2016).

[15] A. Sandu, E. M. Constantinescu, W. Liao, G. R. Carmichael, T. Chai, J. H. Seinfeld, and D. Dăescu, *Ensemble–based data assimilation for atmospheric chemical transport models,* in *International Conference on Computational Science* (Springer, 2005) pp. 648–655.

[16] 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).

[17] S. Lopez-Restrepo, A. Yarce, N. Pinel, O. L. 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**, 91 (2021).

[18] 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, *LOTOS-EUROS v2.0 Reference Guide. TNO 2016 R10,* TNO Innovation for life (2016), https://airqualitymodeling.tno.nl/publish/pages/3175/lotos-euros$_user - guide_v2 - 2 - 002.pdf$.

[19] 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).

[20] M. Schaap, R. M. A. Timmermans, M. Roemer, G. A. C. Boersen, P. J. H. Builtjes, F. J. Sauter, G. J. M. Velders, and J. P. Beck, *The LOTOS-EUROS*

model: description, validation and latest developments, International Journal of Environment and Pollution **32**, 270 (2008).

[21] C. Fountoukis and A. Nenes, *ISORROPIA II : a computationally efficient thermodynamic equilibrium model for* $K^+ - Ca^{2+} - Mg^{2+} - NH_4^+ - SO_4^{2-} - NO_3^- - CL^- - H_2O$ *aerosols,* Atmospheric Chemistry and Physics **7**, 4639 (2007).

[22] R. J. Wichink Kruit, M. Schaap, F. J. Sauter, M. C. Van Zanten, and W. A. Van Pul, *Modeling the distribution of ammonia across Europe including bi-directional surface-atmosphere exchange,* Biogeosciences **9**, 5261 (2012).

[23] M. Schaap, M. van Loon, H. M. ten Brink, F. J. Dentener, and P. J. H. Builtjes, *Secondary inorganic aerosol simulations for Europe with special attention to nitrate,* Atmospheric Chemistry and Physics **4**, 857 (2004).

[24] J. Flemming, a. Inness, H. Flentje, V. Huijnen, P. Moinat, M. G. Schultz, and O. Stein, *Coupling global chemistry transport models to ECMWF's integrated forecast system,* Geoscientific Model Development Discussions **2**, 763 (2009).

[25] J. W. Kaiser, A. Heil, M. O. Andreae, A. Benedetti, N. Chubarova, L. Jones, J. J. Morcrette, M. Razinger, M. G. Schultz, M. Suttie, and G. R. Van Der Werf, *Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power,* Biogeosciences **9**, 527 (2012).

[26] E. C. Monahan, D. E. Spiel, and K. L. Davidson, *A Model of Marine Aerosol Generation Via Whitecaps and Wave Disruption,* Oceanographic Sciences Library SPRINGER **2**, 167 (1986).

[27] E. Mårtensson, E. Nilsson, G. de Leeuw, L. Cohen, and H.-C. Hansson, *Laboratory simulations and parameterization of the primary marine aerosol production,* Journal of Geophysical Research: Atmospheres **108**, AAC 15 (2003).

[28] J. W. Erisman and M. Schaap, *The need for ammonia abatement with respect to secondary PM reductions in Europe,* Environmental Pollution **129**, 159 (2004).

[29] J. N. Galloway, F. J. Dentener, D. G. Capone, E. W. Boyer, R. W. Howarth, S. P. Seitzinger, G. P. Asner, C. C. Cleveland, P. A. Green, E. A. Holland, D. M. Karl, A. F. Michaels, J. H. Porter, A. R. Townsend, and C. J. Vörösmarty, *Nitrogen cycles: Past, present, and future,* Biogeochemistry **70**, 153 (2004), arXiv:arXiv:1011.1669v3 .

[30] M. A. Sutton, J. W. Erisman, F. Dentener, and D. Möller, *Ammonia in the environment: From ancient times to the present,* Environmental Pollution **156**, 583 (2008).

[31] M. Van Damme, L. Clarisse, E. Dammers, X. Liu, J. B. Nowak, C. Clerbaux, C. R. Flechard, C. Galy-Lacaux, W. Xu, J. A. Neuman, Y. S. Tang, M. A. Sutton, J. W. Erisman, and P. F. Coheur, *Towards validation of ammonia measurements from the IASI satellite,* Atmospheric Measurement Techniques **8**, 1575 (2015).

[32] J. N. Galloway, A. M. Leach, A. Bleeker, and J. W. Erisman, *A chronology of human understanding of the nitrogen cycle,* Philosophical Transactions of the Royal Society B: Biological Sciences **368**, 20130120 (2013).

[33] J. P. Burrows, M. Weber, M. Buchwitz, V. Rozanov, A. Ladstätter-Weißenmayer, A. Richter, R. DeBeek, R. Hoogen, K. Bramstedt, K.-U. Eichmann, *et al.*, *The global ozone monitoring experiment (gome): Mission concept and first scientific results,* Journal of the Atmospheric Sciences **56**, 151 (1999).

[34] L. Flynn, C. Long, X. Wu, R. Evans, C. Beck, I. Petropavlovskikh, G. McConville, W. Yu, Z. Zhang, J. Niu, *et al.*, *Performance of the ozone mapping and profiler suite (omps) products,* Journal of Geophysical Research: Atmospheres **119**, 6181 (2014).

[35] H. Bovensmann, J. Burrows, M. Buchwitz, J. Frerick, S. Noël, V. Rozanov, K. Chance, and A. Goede, *Sciamachy: Mission objectives and measurement modes,* Journal of the atmospheric sciences **56**, 127 (1999).

[36] P. F. Levelt, G. H. van den Oord, M. R. Dobber, A. Malkki, H. Visser, J. de Vries, P. Stammes, J. O. Lundell, and H. Saari, *The ozone monitoring instrument,* .

[37] P.-F. Coheur, P. Levelt, L. Clarisse, M. Van Damme, H. Eskes, P. Veefkind, C. Clerbaux, F. Dentener, J. W. Erisman, M. Schaap, *et al.*, *Nitrosat, a satellite mission concept for mapping reactive nitrogen at the landscape scale,* in *EGU General Assembly Conference Abstracts* (2021) pp. EGU21–9932.

[38] V. Neumeie, S. Venturini, A. Davitt, H. V. Madiraju, and B. Ryan, *Ghg monitoring from space: A mapping of capabilities across public, private, and hybrid satellite missions,* (2021).

[39] P. Zoogman, X. Liu, R. Suleiman, W. Pennington, D. Flittner, J. Al-Saadi, B. Hilton, D. Nicks, M. Newchurch, J. Carr, S. Janz, M. Andraschko, A. Arola, B. Baker, B. Canova, C. Chan Miller, R. Cohen, J. Davis, M. Dussault, D. Edwards, J. Fishman, A. Ghulam, G. González Abad, M. Grutter, J. Herman, J. Houck, D. Jacob, J. Joiner, B. Kerridge, J. Kim, N. Krotkov, L. Lamsal, C. Li, A. Lindfors, R. Martin, C. McElroy, C. McLinden, V. Natraj, D. Neil, C. Nowlan, E. O'Sullivan, P. Palmer, R. Pierce, M. Pippin, A. Saiz-Lopez, R. Spurr, J. Szykman, O. Torres, J. Veefkind, B. Veihelmann, H. Wang, J. Wang, and K. Chance, *Tropospheric emissions: Monitoring of pollution (tempo),* Journal of Quantitative Spectroscopy and Radiative Transfer **186**, 17 (2017), satellite Remote Sensing and Spectroscopy: Joint ACE-Odin Meeting, October 2015.

[40] S. Gulde, M. Kolm, D. Smith, R. Maurer, G. B. Courrèges-Lacoste, M. Sallusti, and G. Bagnasco, *Sentinel 4: A geostationary imaging uvn spectrometer for air quality monitoring: Status of design, performance and development,* in *International Conference on Space Optics—ICSO 2014*, Vol. 10563 (International Society for Optics and Photonics, 2017) p. 1056341.

[41] J. Kim, U. Jeong, M.-H. Ahn, J. H. Kim, R. J. Park, H. Lee, C. H. Song, Y.-S. Choi, K.-H. Lee, J.-M. Yoo, *et al.*, *New era of air quality monitoring from space: Geostationary environment monitoring spectrometer (gems),* Bulletin of the American Meteorological Society **101**, E1 (2020).

[42] A. Segers, *Data assimilation in atmospheric chemistry models using Kalman filtering* (Delft University Press, The Netherlands, 2002).

[43] R. Daley, *Atmospheric data analysis,* Cambridge University Press (2009).

[44] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen, *Data assimilation in the geosciences: An overview of methods, issues, and perspectives,* Wiley Interdisciplinary Reviews: Climate Change **9**, e535 (2018).

[45] M. Bocquet, H. Elbern, H. Eskes, M. Hirtl, R. Žabkar, G. Carmichael, J. Flemming, A. Inness, M. Pagowski, J. Pérez Camaño, *et al.*, *Data assimilation in atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models,* Atmospheric chemistry and physics **15**, 5325 (2015).

[46] E. M. Constantinescu, A. Sandu, T. Chai, and G. R. Carmichael, *Ensemble-based chemical data assimilation. i: General approach,* Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography **133**, 1229 (2007).

[47] 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).

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

[49] S. Lu, *observations to estimate volcanic ash emissions*, Ph.D. thesis (2017).

[50] C. Liu, M. Xue, C. Liu, and M. Xue, *Relationships among Four-Dimensional Hybrid Ensemble–Variational Data Assimilation Algorithms with Full and Approximate Ensemble Covariance Localization,* Monthly Weather Review **144**, 591 (2016).

[51] K. Singh and A. Sandu, *Variational chemical data assimilation with approximate adjoints,* Computers & Geosciences **40**, 10 (2012).

[52] A. K. Griffith and N. K. Nichols, *Adjoint methods in data assimilation for estimating model error,* Flow, turbulence and combustion **65**, 469 (2000).

[53] M. Altaf, M. El Gharamti, A. W. Heemink, and I. Hoteit, *A reduced adjoint approach to variational data assimilation,* Computer Methods in Applied Mechanics and Engineering **254**, 1 (2013).

1

[54] J.-N. Thepaut and P. Courtier, *Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model,* Quarterly Journal of the Royal Meteorological Society **117**, 1225 (1991).

[55] A. M. Moore, *Adjoint data assimilation methods,* Operational Oceanography in the 21st century , 351 (2011).

[56] M. Verlaan and J. Sumihar, *Observation impact methods for storm surge forecasting systems,* Ocean Dynamics **66**, 221–241 (2016).

[57] M. Asch, M. Bocquet, and M. Nodet, *Data assimilation, methods, algorithms, and applications* (Society for Industrial and Applied Mathematics, 2016).

[58] 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).

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

[60] M. U. Altaf, M. El Gharamti, A. W. Heemink, and I. Hoteit, *A reduced adjoint approach to variational data assimilation,* Computer Methods in Applied Mechanics and Engineering **254**, 1 (2013).

[61] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics,* Journal of Geophysical Research: Oceans **99**, 10143 (1994).

[62] G. Evensen and P. J. Van Leeuwen, *An ensemble kalman smoother for nonlinear dynamics,* Monthly Weather Review **128**, 1852 (2000).

[63] P. J. Van Leeuwen and G. Evensen, *Data assimilation and inverse methods in terms of a probabilistic formulation,* Monthly weather review **124**, 2898 (1996).

[64] P. Sakov and P. R. Oke, *A deterministic formulation of the ensemble kalman filter: an alternative to ensemble square root filters,* Tellus A: Dynamic Meteorology and Oceanography **60**, 361 (2008).

[65] 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).

[66] E. Kalnay, H. LI, T. Miyoshi, S.-C. Yang, and J. Ballabrera-Poy, *4-d-var or ensemble kalman filter?* Tellus A **59**, 758 (2007), https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0870.2007.00261.x .

[67] A. C. Lorenc and M. Jardak, *A comparison of hybrid variational data assimilation methods for global nwp,* Quarterly Journal of the Royal Meteorological Society **144**, 2748 (2018).

[68] E. Emili, S. Gürol, and D. Cariolle, *Accounting for model error in air quality forecasts: an application of 4denvar to the assimilation of atmospheric composition using qg-chem 1.0,* Geoscientific Model Development **9**, 3933 (2016).

[69] 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).

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

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

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

2

Chemical Transport Models to study the trace gasses reactions and dynamics in the atmosphere

Unless there is a global catastrophe — a meteorite impact, a world war, or a pandemic — mankind will remain a major environmental force for many millennia. A daunting task lies ahead for scientists and engineers to guide society toward environmentally sustainable management during the era of the Anthropocene.

Paul J. Crutzen

Chemical Transport Models (CTMs) allow the simulation of the physical dynamics of trace gasses and aerosols, including emission, chemical reactions, transport, and deposition processes. The deposition of atmospheric contaminants like nitrogen dioxide (NO_2) induces chemical fluxes in natural ecosystems, with severe consequences like biodiversity loss. Depositions are sensitive to changes in the land cover maps used in the model.

In this Chapter, we show how the update on topography plays a role in the vertical disposition of the planet's boundary layer for the LOTOS-EUROS CTM

Parts of this Chapter have been published in A. Yarce-Botero et al. (2021). Medellín Air Quality Initiative (MAUI). DOI: 10.5772/intechopen.97571.
More about the MAUI project: https://www.eafit.edu.co/investigacion/grupos/modelado-matematico/proyectos/cofinanciados/Paginas/data-assimilation-2017-2020.aspx

for a regional atmospheric CTM implementation in Northwestern South America. Land use input information and topography input for this simulation area are updated. The impact of these updates in the model was analyzed to identify potentially vulnerable natural areas by quantifying atmospheric deposition pollutants per area in the natural protected zones of interest.

Additional simulations for the updated scenario using point sources were performed to assess the spatial extent of the deposition areas due to the pollutant emissions of Colombia's principal cities. Implementing this model in Colombia provides a fundamental understanding of the main dynamics of pollutants on this territory. It allowed us to define an appropriate set of model parameterizations and detect emissions factors as one of the uncertain parameters to be updated by incorporating measured concentrations via data assimilation, as shown in Chapter 3

2.1. Introduction

Colombia has a territory characterized by a rough topography that introduces complex meteorological dynamics. These dynamics, in consonance with the land use characteristics, drive the transport and deposition of pollutants. The deposition is described as the mechanism that induces a flux of gasses and particles to the land surface due to meteorological, chemical, and biological phenomena [1]. Through atmospheric transport and deposition, emitted pollutants can alter the dynamics of remote ecosystems when they generate an imbalance of the soil and plant concentrations([2, 3]).

Anthropogenic (human-originated) activities are significant sources of N_r (reactive nitrogen) to the atmosphere [4]. The subsequent photochemical reactions of nitrogen oxides NO_x and Amonnia NH_3 create secondary inorganic aerosols [5] that can be transported over large distances [6]. The long-distance transport of secondary inorganic aerosol accounts for over 8% of the planet's reactive nitrogen flow in terrestrial ecosystems [4], and constituted a significant source of N_r to the oceans [7]. Deposition of atmospheric N_r alters oligotrophic ecosystems [5, 8], affecting the distribution of communities of species [9–13] and ecosystem stability [14]. Nitrogen Dioxide NO_2 is a gas emitted by anthropogenic and natural sources as part of the family of the nitrogen oxides NO_x($NO+NO_2$) emitted from anthropogenic (industrial activity, transport, and biomass burning) and natural (NO_x soil emissions and lighting) sources. Once in the atmosphere, NO_2 could be transported over long distances before it is chemically transformed or deposited on the surface. From this process of deposition, this trace gas contributes to the reactive nitrogen in the soil, water, and vegetation, leading to eutrophication (overabundance of nutrients in water bodies, primarily nitrogen and phosphorus) and loss of biodiversity [15]. Additionally, model studies indicate that lighting in the tropical troposphere may significantly influence the NO_x budget [16–18]. Agricultural activities and livestock feedlot operations are the primary causes of atmospheric NH_3, followed by wood-burning (including forest fires) and fossil fuel combustion to a lesser degree. Acute exposures to NH_3 near its origin (4-5 km) can lead to substantial foliage damage to the plant, growth, and productivity decreases [19]. Results of this Chapter have been cited in [20], where the ecosystem impacts from different sources were reviewed and updated for Colombia.

This Chapter focuses on the study of the atmospheric transport and deposition of N_r, primarily (NO_2), with additional experiments with ammonia (NH_3), over Northwest South America, and in particular over Colombia using Chemical Transport Models (CTMs). Due to the lack of observations in this territory, mathematical models are essential for understanding the spatial and temporal dynamics of atmospheric gasses over Colombia. CTMs simulate dynamic concentrations of trace gases and aerosols influenced by emission, chemical reactions, transport, and deposition. All these processes are not devoid of uncertainty.

The use of CTMs in the Tropical Andean Region (including NW South America) has been recently reviewed ([21]), highlighting the difficulties involved in modeling atmospheric chemistry and transport in tropical regions with abrupt topography. Other applications with CTMs over Colombia have included the use of another

model, the WRF-Chem CTM, for evaluating the dynamics behind high particulate matter (pm) pm_{10} and $pm_{2.5}$ episodes under different meteorological conditions for the city of Bogotá [22]. Another case is in the exploration of methane (CH_4) emission sources with the use of satellite data, for example, from landfills [23, 24], and the dispersion of CO within the Aburrá Valley [25]. In terms of the pollutant NO_2, surface NO_2 concentrations for Colombia in 2007 were estimated from simulations of the global model GEOS-Chem CTM (resolution 2.5° x 2°) and Ozone Monitoring Instrument (OMI) NO_2 column measurements, resulting in the identification of biomass burning as a significant source of atmospheric NO_2 [26]. Furthermore, Barten *et al.* use the WRF-Chem model and OMI-generated satellite NO_x sources and sinks in Colombia for 2014 in a coarse resolution of 1° x 1° resolution identifying lighting as one of the main contributors to the total nitrogen emission budget. In this Chapter, a qualitative and quantitative analysis of the spatial and temporal patterns of deposition of (NO_2) and (NH_3) over the Colombian domain using the LOTOS-EUROS CTM is presented.

This Chapter updates the LOTOS-EUROS model land use and other input characteristics to simulate nitrogen deposition in Colombia.

2.1.1. Updating the LOTOS-EUROS Chemical Transport Model in a new territory

The CTM model used was the LOTOS-EUROS (Long Term Ozone Simulation- European Operational Smog model) v2.0.001 to simulate in three dimensions and time, the atmospheric species of the lower layers of the troposphere over Colombia. The LOTOS-EUROS model was created to study the atmospheric distribution of photo-oxidants in the 1990s by the Dutch Organization for Applied Scientific Research (TNO) https://www.tno.nl/ and the Environmental Assessment Agency of the Dutch National Institute for Public Health and the Environment (RIVM/MNP)https://www.rivm.nl/. TNO, RIVM, and the Royal Dutch Meteorological Institute (KNMI) https://www.knmi.nl/ utilized LOTOS-EUROS version 1.0, which was developed by combining the two models, LOTOS and EUROS, in 2004. In 2016, the open-source 2.0 version of the model was published to expand the user base (https://ci.tno.nl/gitlab/lotos-euros). In the LOTOS-EUROS, the species' dynamics are regulated by processes of chemical reactions, diffusion, drag, dry and wet depositions, emissions, and advection. Numerous studies have used the model simulating trace gas and aerosol concentrations [28–30]. Nowadays, the model is used for daily operational air quality forecasts over the Europe domain [31], and more focused over the Netherlands [32], and China [33], and it is used for daily forecasts of dust concentrations over North Africa as part of the SDS-WAS service [34]. Recently, it has been used to forecast the dynamics of particulate matter (PM) in the Aburrá Valley in Colombia [35], improving the emission inventory of this specific pollutant through the implementation of Ensemble Kalman Filter (EnKF) or Local Ensemble Kalman Filter (LEnKF), which are the data assimilation techniques that are implemented by default in the model for estimating emission factors.

Among the various removal mechanisms of pollutants in the atmosphere, dry

deposition is an important one that comprises a large part of total deposition [36]. The deposition of gases is a diffusion flux process driven by the difference in concentrations inside and outside the vegetation tissues. Deposition can occur directly into the soil but is much stronger on vegetation surfaces. The primary surfaces for deposition are the leaves, where gasses and particles can be absorbed via the stomata or the cuticle [37], though outer bark surfaces can also contribute. In LOTOS-EUROS, the resistance approach [38] is used to simulate gas-phase deposition following the **DEPAC** (**DEP**osition of **A**cidyfing **C**ompounds) formulation [39]. The DEPAC module considers three pathways: deposition through the stomata, deposition through the external leaf surface, and deposition through the soil to calculate the deposition velocities for each one of the categories. The model considers the components exchanged between the atmosphere's lower layers and surfaces for estimating deposition. For aerosol deposition, a uniform structure is used for all land use categories in the analysis ([40]) with an explicit dependence on aerosol size [38]. The simulation of atmospheric species' deposition is challenging depending on whether the phenomena are modeled in the close, local, or long-field range. Larger particles dominate a near-field deposition. Local field events occur in the plume portion dominated by the wind-driven trajectory. The peak-to-mean concentration ratios are much smaller than closer to the source, resulting in more uniform deposition patterns. In long-range fields, the larger particles have been removed, so smaller particles are the primary concern of the modeling [41].

The application of the LOTOS-EUROS model in South America has substantial challenges due to processes that must be better described in the CTM to improve its performance. For example, the tropical meteorological dynamics like the bi-annual transit of the Inter-Tropical Convergence Zone and el Niño, a climate pattern that leads to an unusual warming of surface waters in the eastern Pacific Ocean [42]. Other sources of uncertainty are the abrupt topography, the diversity of ecosystems, and the representation of biogenic emissions, which substantially contribute to the uncertainty when applying a CTM in the region [21]. Previous studies in South American regions of similar characteristics were limited by using relatively coarse spatial resolutions [26, 27]. In this Chapter, we investigate the effect on the CTM pollutant deposition predictions of input modifications in land use and topography.

In addition, point source simulations were conducted to understand the spatial deposition dynamics of N_r emanating from the main cities of the country (Bogotá, Medellin, Barranquilla, Cali). This work contributes to evaluating the performance of the LOTOS-EUROS CTM in NW South America. The evaluation of large-scale models in the tropical Andean zone appears to be challenging because this zone is not often entirely covered by a sufficient sensor network, so CTM is often compared to the scarce measurements, and satellite information is increasingly valued to be used for this purpose.

2.2. Methods

2.2.1. Domains and model set up

The two nested domains used for the simulations to provide higher-resolution boundary conditions for the area of interest are shown in Figure 2.1. The first domain (D1) stretched from the west coast of Nicaragua to the Caribbean Dutch Islands and Eastern Venezuela, with a horizontal resolution of 0.27° (about 28 km). The inner domain (DCol) covered the continental territory of Colombia, with a horizontal resolution of 0.09°. All deposition analyses were conducted in the inner domain.

A table with the main characteristics of the simulation run is shown in Table 2.1.

Period	1-January-2016 to 31-December-2016
Meteorology	ECMWF; Temp. res.: 3h; Horiz. res.: 0.07° x 0.07°
Initial and boundary conditions	LOTOS-EUROS (D1). Temp. res.: 1h. Horiz. res.: 0.09° x 0.09°
Anthropogenic emissions	EDGAR v4.2. Spat. res.: 0.1° x 0.1°
Biogenic emissions	MEGAN Spat. res.: 0.1° x 0.1°
Fire emissions	MACC/CAMS GFAS Spat. res.: 0.1° x 0.1°
Landuse	GLC2000. Spat. res.: 0.1° x 0.1°
Orography	GMTED2010. Spat. res: 0.002° x 0.002°
Domain 1 (D1) Lat x Lon	[-8.5°,18°] x [-84°,-60°]
Domain Colombia (DCol) Lat x Lon	[-4.55°,13.27°] x [-79.80°,-65.94°]

Table 2.1: LOTOS-EUROS configuration settings for the simulations in this Chapter.

The default vertical layer definition for LOTOS-EUROS is obtained from ECMWF meteorological data, with a resolution of 0.07°(≈ 7 km). LOTOS-EUROS can receive meteorological fields in Cartesian and non-Cartesian projections and has three different option schemes to process this information: mixed layer, hybrid, and meteorological levels driven [43–45]. The first implementation was made with the mixed-layer scheme. In this scheme, the model follows a mixed-layer level with a 25 m surface layer, then a mixed layer at the top at the boundary layer height taken from the meteorological input (ECMWF), and two reservoir layers of at least 500 m at the top. When the mixing layer is thick over mountainous terrain in the tropics, where elevations can regularly reach more than 3500 m, the model's top is extended to accommodate the minimum thickness of the reservoir layers.

An updated elevation model used for the region was obtained from the Global Multi resolution Terrain Elevation Data (GMTED 2010) [46], with a resolution of 0.002°(≈ 220 m).

A detailed high-resolution and up-to-date anthropogenic emissions inventory for the time of the simulations was not available for this region. Therefore, emissions were taken from EDGAR inventory (Emission Database for Global Atmospheric Research) 4.2 for 2008. However, previous studies have shown a significant gap in knowledge for the Colombian territory in the EDGAR inventory (e.g., [47–49]), this database was at the time the only one accessible with all the species required to operate the model in the selected domain for the time of the simulations. Biogenic emissions were taken from the MEGAN 2.1 model. The MACC/CAMS GFAS global fire assimilation system from [50] was used with a time resolution of 1 hour to account

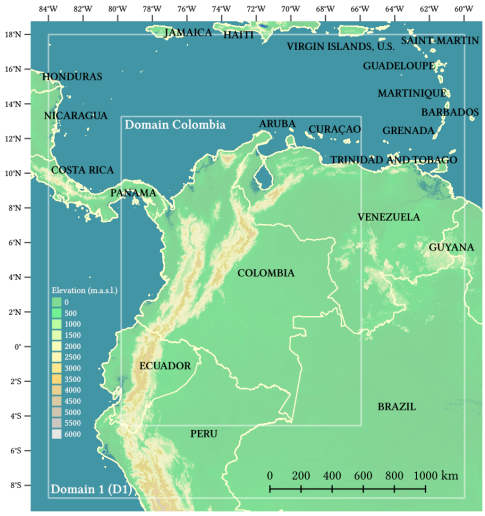


Figure 2.1: Nested domain model configuration selected for the LOTOS-EUROS simulations over North West South America Domain (D1) (0.27°) and Domain Colombia (0.09°).

for the occasional fire events. The chemical mechanism used was the Carbon Bond mechanism 5 (CB05), and the sea salt emissions were parameterized according to formulations for the fine and coarse aerosol modes from [51] and [52]. Once the simulation domains were set and the characterization for the emission was chosen, we proceeded to characterize the land cover and land use data.

2.2.2. Land Cover/Land Use data

LOTOS-EUROS requires a 2D map with deposition properties per grid cell to model the deposition dynamics. Land use characteristics are relevant for the CTM deposition dynamics because they define the terrain roughness and canopy altitude of each category that determine the velocity of the component, depending on the vegetation type[53].

The default land use/land cover (LU/LC) input data for LOTOS-EUROS were derived from the Global Land Cover (GLC2000) project [54]. GLC2000 includes 23 categories consistent with the FAO (Land Cover Classification System of the Food and Agriculture Organization) classification [55]. For South America, mapping these categories at spatial resolutions of 0.01°(approx. 1 km) was done in [56], with contributions from some regional experts based on multi-resolution satellite data. The GLC project was updated with the LU/LC data from the 2009 Land Cover Climate Change Initiative (CCI) dataset [57]. CCI has 38 categories with a horizontal resolution of 0.003°x 0.003°(300 m x 300 m). Figure 2.2 compares the default and updated LU/LC models for Aburrá Valley. The mapping from CCI to GLC considered the similar morphological characteristics between categories and the aseasonality in this tropical region. The mapping from GLC to DEPAC is the standard scheme constructed for LOTOS EUROS. The model defines each grid cell as the fraction

Figure 2.2: Comparison between (A) the LOTOS-EUROS's original land cover model (Global Land Cover, with a resolution of 0.01°x 0.01°(1 km x 1 km).) and (B) the updated land cover scheme (Land cover from the Climate Change Initiative, with resolution 0.003°x 0.003°(300 m x 300 m)). Real-color Landsat cloudless imagery from the date 2019-09-03 is included in (C) to reference the artificial surfaces from the city infrastructure. The political boundaries marked with white lines correspond to the municipality of Medellín and the other nine municipalities that constitute the Aburrá Valley Metropolitan Area conurbation.

covered by each of the LU/LC classes used by the DEPAC module and calculates the deposition of each fraction. The mapping of the 39 (CCI) and 23 (GLC) LU/LC categories to the nine classes of the DEPAC deposition model is illustrated in Figure 2.3. The descriptions of each category are presented in Table 2.2.

Matlab and Python scripts are available from the author upon reasonable request.

2.2.3. Simulations description

Simulations were carried out using the following configurations for the experiments: *Experiment 1 (Exp-1)*, default elevation model and default LU/LC data; *Experiment 2 (Exp-2)*, default elevation model and updated LU/LC data; *Experiment 3 (Exp-3)*, updated elevation model and default LU/LC data; and *Experiment 4 (Exp-4)*, updated elevation model and updated LU/LC data. For each of these four experiments, the total deposition (wet and dry) of nitrogen dioxide (NO_2) and ammonia (NH_3) was calculated for the year 2016. Zonal analysis for estimating annual nitrogen deposition in protected areas and Páramo ecosystems was conducted in QGIS 3.10 (https://qgis.org), using official Colombian cartography obtained from the National Geographic Information Systems (SIGOT, https://sigot.igac.gov.co/). Simulations were conducted in the APOLO cluster (https://www.eafit.edu.co/apolo), and data was post-processed (analyzed) using MATLAB and Python software.

Number	Land Cover Climate Change Initiative (CCI) label	Code CCI
0	No data	ND
10	Cropland, rainfed	CRA
11	Herbaceous cover	HBC
12	Tree or shrub cover	TSC
20	Cropland, irrigated or post-flooding	CIP
30	Mosaic cropland (>50%) / natural vegetation (tree, shrub, herbaceous cover) (<50%)	MC/ NV
40	Mosaic natural vegetation (tree, shrub, herbaceous cover) (>50%) / cropland (<50%)	MNV
50	Tree cover, broadleaved, evergreen, closed to open (>15%)	TBECO
60	Tree cover, broadleaved, deciduous, closed to open (>15%)	TBDCO
61	Tree cover, broadleaved, deciduous, closed (>40%)	TBDC
62	Tree cover, broadleaved, deciduous, open (15-40%)	TBDO
70	Tree cover, needle-leaved, evergreen, closed to open (>15%)	TNECO
71	Tree cover, needle-leaved, evergreen, closed (>40%)	TNEC
72	Tree cover, needle-leaved, evergreen, open (15-40%)	TNEO
80	Tree cover, needle-leaved, deciduous, closed to open (>15%)	TNDCO
81	Tree cover, needle-leaved, deciduous, closed (>40%)	TNDC
82	Tree cover, needle-leaved, deciduous, open (15-40%)	TNDO
90	Tree cover, mixed leaf type (broadleaved and needle-leaved)	TMBN
100	Mosaic tree and shrub (>50%) / herbaceous cover (<50%)	MTS
110	Mosaic herbaceous cover (>50%) / tree and shrub (<50%)	MHC
120	Shrubland	SRU
121	Shrubland evergreen	SRE
122	Shrubland deciduous	SHD
130	Grassland	GRA
140	Lichens and mosses	LIM
150	Sparse vegetation (tree, shrub, herbaceous cover) (<15%)	SPV
151	Sparse tree (<15%)	SPT
152	Sparse shrub (<15%)	SPS
153	Sparse herbaceous cover (<15%)	SHC
160	Tree cover, flooded, fresh or brackish water	TFFB
170	Tree cover, flooded, saline water	TFS
180	Shrub or herbaceous cover, flooded, fresh/saline/brackish water	SHCF
190	Urban areas	URB
200	Bare areas	BAR
201	Consolidated bare areas	CBA
202	Unconsolidated bare areas	UBA
210	Water bodies	WAB
220	Permanent snow and ice	SNI

Number	Global Land Cover (GLC) label	Code CORINE
1	Tree Cover, broadleaved, evergreen	TBE
2	Tree Cover, broadleved, deciduous, closed	TBDC
3	Tree Cover, broadleaved, deciduous, open	TBDO
4	Tree Cover, needle-leaved, evergreen	TNE
5	Tree Cover, needle-leaved, deciduous	TND
6	Tree Cover, mixed leaf type	TML
7	Tree Cover, regularly flooded, freshwater	TRF
8	Tree Cover, regularly flooded, saline water	TRS
9	Mosaic: Tree Cover / Other natural vegetation	MTO
10	Tree Cover, burnt	TCB
11	Shrub Cover, closed-open, evergreen	SCOE
12	Shrub Cover, closed-open, deciduous	SCCD
13	Herbaceous Cover, closed-open	HCCO
14	Sparse herbaceous or sparse shrub cover	SHSS
15	Regularly flooded shrub and/or herbaceous cover	RFSH
16	Cultivated and managed areas	CMA
17	Mosaic: Cropland / Tree Cover / Other natural vege	MCTO
18	Mosaic: Cropland / Shrub and/or grass cover	CSG
19	Bare areas	BAR
20	Water bodies	WAB
21	Snow an Ice	SNI
22	Artificial surfaces and associated areas	URB
23	No Data	ND

Number	LOTOS-EUROS fractional categories	LE-fract-code
1	Arable	ARA
2	Permanent crops	CRP
3	Grass	GRS
4	Coniferous forest	CNF
5	Deciduous forest	DEC
6	Other	OTH
7	Water	WAT
8	Urban	NO2
9	Desert	DSR

Table 2.2: Land Use/ Land Cover categories and acronyms for the two data sources used in this study and for the DEPAC module. The table in the right bottom part contains the LOTOS-EUROS fractional categories

2.2.4. Fate of urban contaminants experiments

The grid cells that house the centroids of the urban areas of Bogotá, Medellín, Cali, and Barranquilla were assumed to be artificial point sources of emissions to explore the fate of the atmospheric nitrogenous species emitted from the main Colombian cities. The simulations were conducted with the updated elevation model and the updated LU/LC scheme detailed in the last section for ten days at four different times of the year: March 1-10, June 1-10, September 1-10, and December 1-10. After a 2-day model spin-up, the point source was activated from 08:00-18:00 of day 3 of the simulation, emitting a total of 1000 kg/hour NO_2, which is the amount of daily NO_2 emissions reported for Medellín [58]. The artificial emissions were monitored for seven additional days, during which time all the emitted species had either been deposited or transformed. Similar simulations were conducted without the point source's activation to estimate the background deposition values for each grid cell.

2.3. Results

2.3.1. Influence of the elevation model update

LOTOS-EUROS interpolates the input elevation data within each grid cell according to the simulation's resolution (Fig. 2.4) (0.12°,0.09,0.06,0.12). The input elevation model can change variables such as the temperature profiles of the vertical layers. The effect of an updated elevation model depends on the desired resolution of the simulation grid; it means it is not as beneficial to go to the updated orography if the resolution is higher than the input information, but for getting specific details on the structure of the terrain, a more detailed topography impacts also in the structures

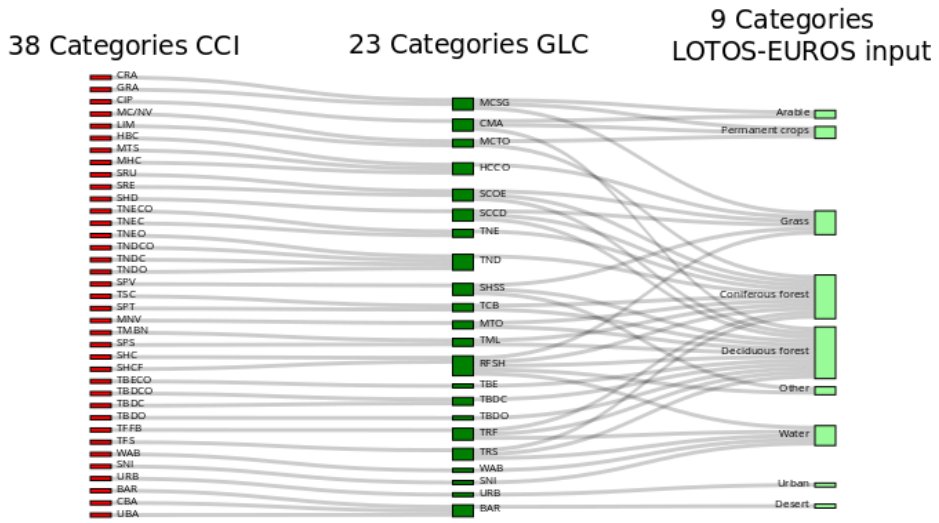


Figure 2.3: Climate Change Initiative (CCI) Land Cover data categories mapped to the Global Land Cover (GLC). The GLC categories are, in turn, mapped to the model's equivalencies. The category's codes are as in Table 2.2.

in the interpolated grid.

The most significant temperature changes occurred in the upper layers above the rugged terrain, reaching differences of up to 5°C degrees in the top layers. Figure 2.5 compares the default (left image) and the updated elevation model (right image). The inset zoomed in around the Aburrá Valley (Medellín and neighboring municipalities) demonstrates the absence of the valley in the default elevation model with the update to the new topography. This improved representativity has impacts on the definition of the low boundary information.

Figure 2.6 shows a transverse cut at a latitude of 6.6°North, corresponding to the Aburrá Valley's location. In the left panels, the transversal cut is depicted over the topography input. The center panels show the temperature interpolated for the simulation at a horizontal resolution of 0.09°x 0.09°, illustrating the impact of the change of input elevation information for the Aburrá Valley. The right image shows a difference between the two temperature fields, where the more noticeable changes were in the upper level.

2.3.2. Conversion of land use categories

Figure 2.7 compares the fractional coverage changes of the 23 categories of the GLC2000 land use when using the CCI land use data and the mapping presented in Figure 2.3. Image (A) compares the total area represented by each category over the entire study domain. Image (B) shows the fractional bias error calculations between these scenarios for all the simulation domains, illustrating the fractional change for each category. LOTOS-EUROS allocates a fractional usage per grid of

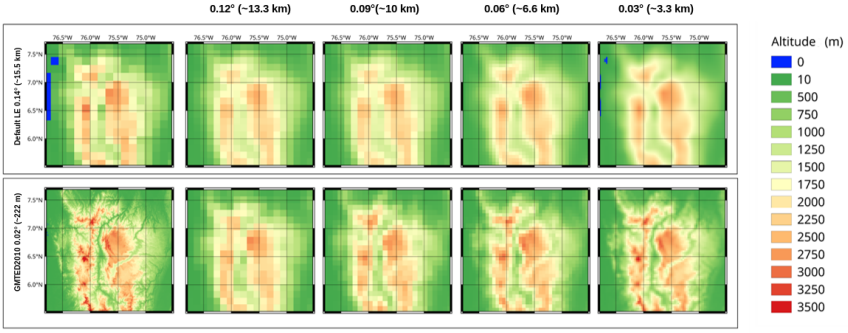


Figure 2.4: Comparison of the two topography interpolated to four different grid resolutions, in the upper left panel, the default topography for the LOTOS-EUROS and the left panel below the GMTED2010. Colorscale depicts the altitude in meters.

the land surface to each of the nine categories used by DEPAC to generate the corresponding deposition estimates.

Figure 2.8 corresponds to the land use fractions assumed by the model for both default land use GLC2000 (a) and the updated scenario CCI (b), the default. These fractions are the spatial percentage cover categories for which the model performs the simulation calculations. It shows the mean fractional bias of the updated CCI land cover model relative to the default GLC LU/LC data. Changes were noticeable for categories like grasslands, where the deposition decreased with the updated land-use configuration. For categories such as deciduous forest, arable, coniferous forest, and permanent crops, primarily located in the country's eastern region, the simulations showed an increase in depositions related to the change in the fraction percentage of these categories.

The Amazon region (SE) presented minimal changes in deposition between the two LU/LC scenarios, primarily due to the minor changes in LU/LC between the two data sources. The most significant changes related to deposition in deciduous and coniferous forests were found along the Andean cordilleras, which were higher in the updated LU/LC scheme.

The pictures in Fig. 2.9 show the comparison between the different scenarios for the NO_2 and NH_3 accumulated dry deposition for the year 2016 simulated with the LOTOS-EUROS CTM. The upper left corner of each situation corresponds to *Exp-1*, and the lower right corner corresponds to the *Exp-4* conditions. The top right images correspond to the MFB comparison between *Exp-1* and *Exp-2* and the lower left to the same MFB comparison between *Exp-1* and *Exp-3* elucidating the impact on the total deposition for 2016 due to the change in land use and orography scenarios respectively. MFB comparisons allowed us to highlight some regions (A, B, C, and D) of interest due to drastic changes that overestimate or underestimate deposition values concerning the initial condition given in the *Exp-1*.

Figure 2.8 shows the land use comparison for the respective chemical compound

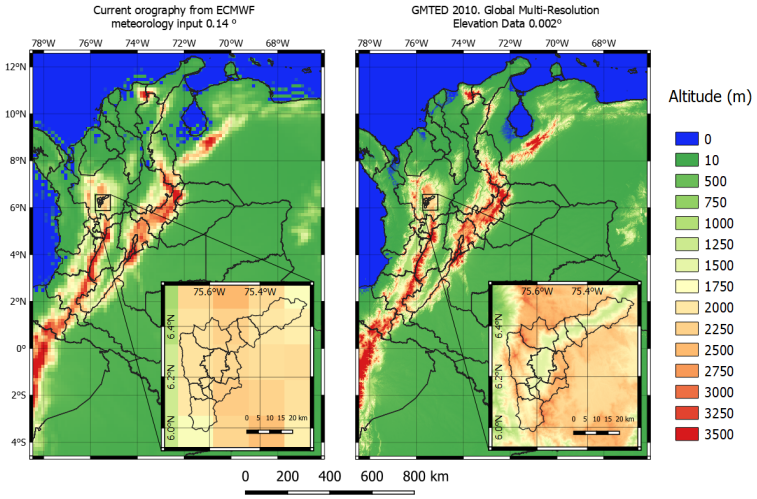


Figure 2.5: Comparison between the default elevation model for LOTOS-EUROS (A) and the updated elevation model derived from the GMTED 2010 data (B). The insets illustrate the differences in elevation representation for the Aburrá Valley.

simulation deposition output. For zone A, the CCI water bodies category is more abundant than the GLC. The regularly flooded shrub and herbaceous cover are predominant in the bodies of internal waters. The three cover categories and the urban areas are noticeable in the zones with an MFB image overestimated. On the other hand, the bare areas and the three cover and regularly flooded, cultivated, and managed areas and mosaic cropland are more abundant in the old land cover. For zone C, the urban areas are more relevant in updating land use. The land use located at the left of the principal metropolitan area of the images corresponding to cultivated and managed areas seems to promote the deposition of the updated condition significantly. Moreover, in Figure 2.10 E, we can see that the herbaceous cover closed-open was prevalent in the old land use and changed to mosaic cropland, shrub, and grass cover in the new one. In the MFB image, overestimating respect to the Exp-1 conditions is seen in how the dominant land use was changed.

2.4. Point sources experiment

The LOTOS-EUROS emission module describes the discharge of tracers and aerosols from various sources (anthropogenic, biogenic, marine, airborne dust, fires) that can be configured to define emissions in specific point sources to simulate scenarios. Figure 2.11 shows the simulation of the total deposition (dry and wet) of nitrogen considering the emission of the four main cities of Colombia. Cities were assumed to be point emission sources for a regional domain, determining this city's influence area.

Figure (2.11, A) shows the contours generated with an increment of 5 g/ha

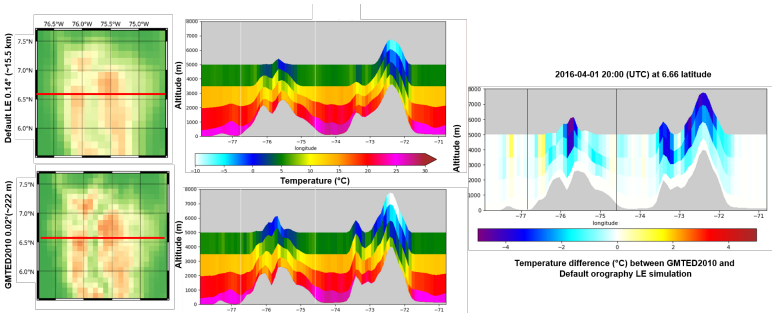


Figure 2.6: Left panels: Topography and transversal cut comparison at 6.6°N for vertical temperature profile comparison. Right panel: the difference between the vertical cuts.

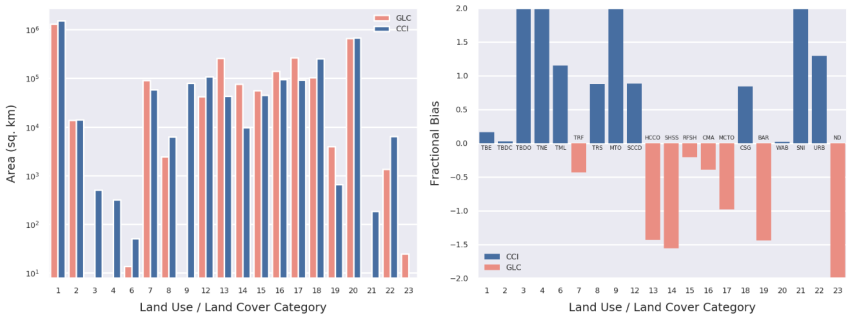


Figure 2.7: a) Comparison of the area between categories from both land uses b) Fractional Bias comparison.

between the contour level and a bias correction of +2 g/ha to avoid negative values that appear as numerical noise from the rest of the reference run minus the punctual perturbed emission run. The more rounded contour delimits the impact zone for each of the cities.

Identifying some wind direction trends for the four different times of 2016 (1-9 days of March, June, September, and December) is possible. The influence of Barranquilla on faraway zones is perceivable due to the proximity of this city to the Caribbean coast, where intense wind conditions exist, and flat topography that drives the transport dynamics far away. Most of the depositions from Barranquilla go to the ocean direction and, at other times, to the southwest of the city, reaching the other cities' deposition areas. For the other cities, the impact area is more limited but has higher deposition values due to the roughness of the mountainous terrain and the lower magnitude of the wind patterns presented. It is also interesting to see that the deposition from Bogotá could reach other cities, such as Medellín, at some time of the year. From the simulations conducted, it's possible to extract some trends for the wind directions.

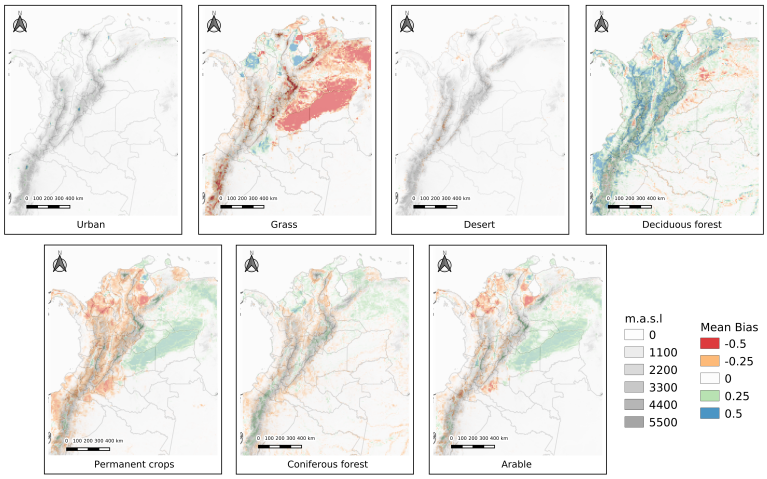


Figure 2.8: Only 7 of 9 categories are shown here. The other two are the categories water and other.

Figure (2.11, B) shows a simulation of what happens with Medellín for simulation resolution of 0.03 °x 0.03 °, conceiving the point source concentrated in a grid cell in the middle of the metropolitan area. The simulation results indicate that the northwest-west area of the Aburrá Valley (Medellín and Metropolitan Area) is the most affected, which could be seen in how páramos ecosystems located at points A and B in the map receive nitrogen that was emitted from the cities. The mosaic in the right part of Fig. (2.11, C) is a detailed simulation, conceiving the point source concentrated now in a grid cell in the middle of the metropolitan Rionegro area. During some parts of the year, it is crucial to notice that Rionegro influences Medellín's emissions far away depositions. The transport dynamic of contaminants should be studied in more detail because of the increasing Rionegro area for industrial and urban settings.

After the first implementation of the model with the updating of land use, deposition analysis was performed. The following Figure shows a quantitative analysis of the páramos ecosystems (Fig. 2.12, a)) with the natural protected areas (Fig. 2.12, b)) delimitation, via the LOTOS-EUROS NO_2 model output and the total deposition of those areas are receiving. Cartographic data for ecosystems and protected areas were obtained from the Colombian System for Environmental Data (Sistema de Información Ambiental de Colombia, SIAC; http://www.siac.gov.co/).

The Páramo de Las Baldías was the paramo zone with greater exposure to atmospheric pollutants. It receives 14 kg/(ha yr), above the standard critical load of Nr, and 60 kg/(ha yr) of ozone. The páramos of western Antioquia receive 5-6 kg/(ha yr), while the Sonsón Páramos are identified as those with the lowest exposure to atmospheric pollutants(2.2-2.6 kg/(ha yr) Nr; 35-40 kg/(ha yr) ozone). The ecosystems of the west of the Caribbean coast receive the highest ozone load in the country (50-100 kg / (ha yr)).

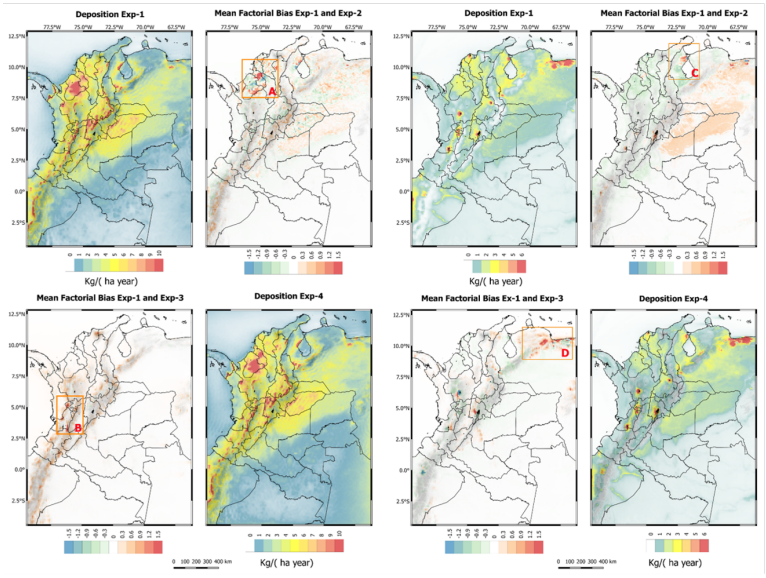


Figure 2.9: Comparison between the experimental scenarios with the updating of orography and land cover for two chemical compounds a) the Ammonia NH_3, b) the Nitrogen dioxide (NO_2).

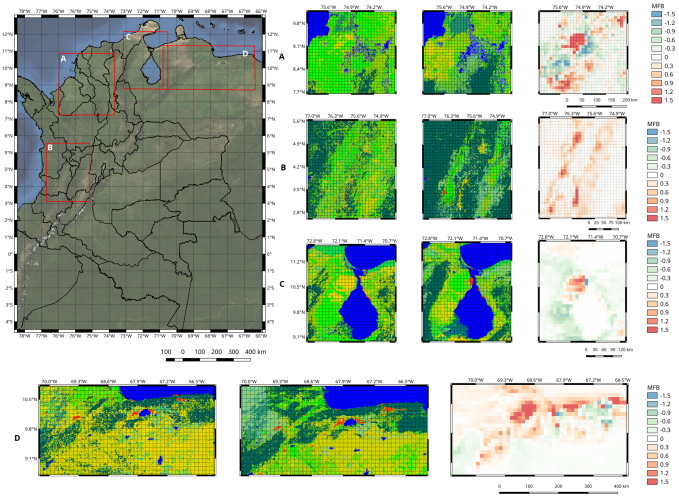


Figure 2.10: Detailed comparison in four regions A, B, C, D for the changes in land use for the simulations performed with the LOTOS-EUROS for this chapter. The order in the graphics are the left the GLC, the middle the CCI and in the right the mean fractional bias between the nitrogen concentration outputs of the LOTOS-EUROS simulating with this two input information.

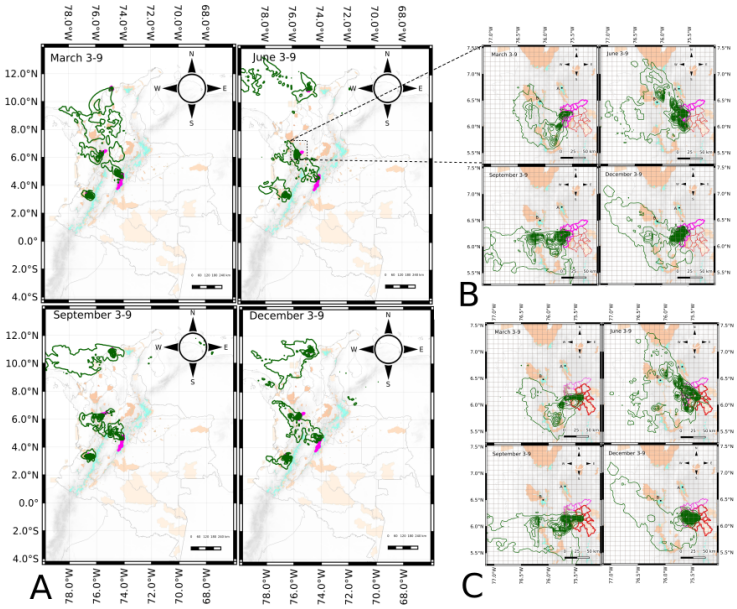


Figure 2.11: Point source contour experiment in 2016 at different times of the year for A) the 4 biggest cities in the country (pink lines) and contour limitation of the influence area from the deposition plots. B) Higher resolution for Medellin and C) Higher resolution for Rionegro denoting the impact in terms of transport and deposition.

Figure 2.12 shows a histogram quantification over the natural protected areas and Paáaramo ecosystems to identify the critical deposition load of NO_2 in ecosystems and how many sites are at higher risk for the transport of pollutants. The critical load in a tropical ecosystem is not very well determined, and, probably, due to the high biodiversity index in this region, the critical load should be lower.

Using LOTOS-EUROS deposition simulations, we calculated the average annual total nitrogen deposition in Colombia in 2016, estimated to be 4.2 kg/(ha yr) with min and max values of (0.8-18.9) kg/(ha yr). The average annual ozone deposition was 36.0 with min and max values of (7.7-109.4) kg/(ha yr), respectively. Within the year, the lowest ozone deposition has been estimated for May and July, while the maximum occurred in June and December-January. Updated protected areas and ecosystem cartography were superposed on simulation results. Weighted average annual total nitrogen deposition in national protected areas and Páramos (a critical ecosystem for the water cycle in Colombia) has been estimated to be 2.61 kg/(ha yr) with min and max values of (0.0-18.5) kg/(ha yr) and 4.11 kg/(ha yr) and min and max values of (2.2-14.2) kg/(ha yr), respectively. Total nitrogen deposition at or above 5 kg/(ha yr) occurs in 60% of declared protected areas and over 13% of the páramo ecosystems. Figure 2.14 present a weighted average annual total ozone deposition in nationally protected areas and páramos has been estimated at

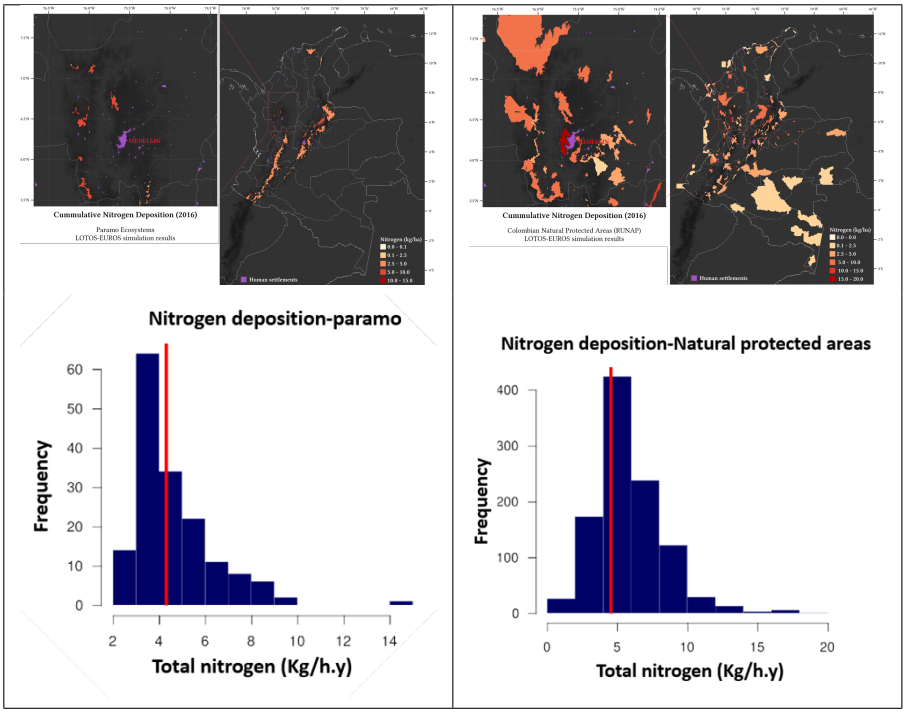


Figure 2.12: Simulated calculation of annual (2016) deposited nitrogen budget for Paramo ecosystem (left image) and natural protected areas (right image). Based on the model output, the histograms below quantify the total nitrogen deposition in the Paramo region and other national protected areas. The red line in each histogram refers to this region's theoretical, critical nitrogen load, which shows that a considerable area is vulnerable nowadays to transport and deposition.

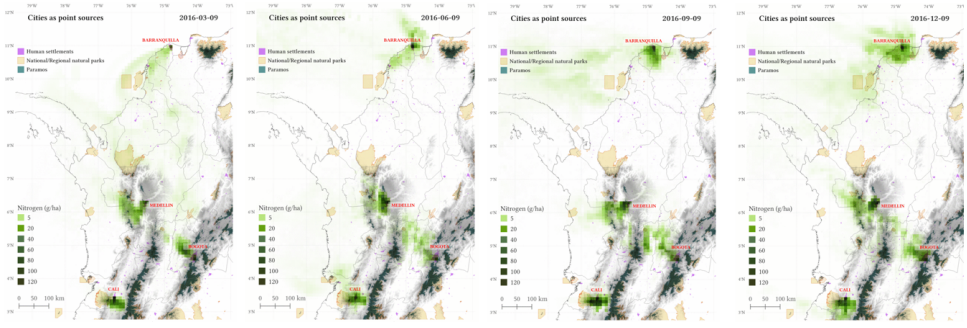


Figure 2.13: Point source experiment over the main cities in Colombia for moments in the year. Nitrogen deposition patterns for different seasonal trends.

19.8 kg/(ha yr) with min and max values (0.0-99.3) kg/(ha yr) and 24.8 kg/(ha yr) with min and max (13.2-62.4) kg/(ha yr), respectively.

Figure 2.14: Simulated calculation of annual deposited Ozone budget for natural protected areas and paramo ecosystem.Weighted average annual total ozone deposition (Kg/(ha year)) in nationally protected areas and páramos and zoom over the east region over the Andean mountain corridor

2.5. Summary and Discussion

This work's main contribution is assessing the LOTOS-EUROS CTM Model's performance in the Tropical Andes Domain, focusing the analysis on deposition simulations. Determining the relevance of the Land Use Land Cover and Topography as model inputs and paying attention to specific points of the study zone were certain factors driving the deposition process and were some of the results of this work. Future improvements of the simulation method could consider cities as pollutant sources with, at first, shapes closer to the emission inventory.

Although LOTOS-EUROS is terrain-following, the chosen domain simulations are sensitive to the orography map only if the simulation is also high resolution. We also know that the EDGAR inventory does not appropriately represent some component emission dynamics. For the time of the experiments, this was the inventory we had available for the study domain. In later works (see Chapter 4), an updating of the inventory employing parameter estimation using data assimilation corresponding to a top-down reconstruction methodology [59].

Deposition depends on land-use changes. It was essential to analyze the area of interest's temporal behavior to identify zones with over and sub-estimation compared with the early reference. Natural protected areas with notable changes in deposition between the default and updated input data were identified, highlighting

the importance of using accurate and up-to-date land cover data in the simulation model. Vulnerable areas such as natural protected areas and páramos ecosystems may require more than a local conservation effort to preserve their ecological functions.

Simulations with point sources identified the transport patterns in the territory. They showed the regional influence of the major cities based on qualitative and quantitative results to understand the dynamics of emission and deposition of contaminants for the main cities of Colombia, which consists of an attractive information supply to begin understanding the transport of atmospheric pollutants in this territory. The atmospheric transport and deposition of contaminants present ecosystem risk factors that require an evaluation of impacts directly in the field based on the reported results and their inclusion in conservation strategies.

References

[1] M. Giardina and P. Buffa, *A new approach for modeling dry deposition velocity of particles,* Atmospheric Environment **180**, 11 (2018).

[2] S. M. Bard, *Global transport of anthropogenic contaminants and the consequences for the arctic marine ecosystem,* Marine Pollution Bulletin **38**, 356 (1999).

[3] D. Fowler, J. N. Cape, M. Coyle, C. Flechard, J. Kuylenstierna, K. Hicks, D. Derwent, C. Johnson, and D. Stevenson, *The global exposure of forests to air pollutants,* Water, Air, and Soil Pollution **116**, 5 (1999).

[4] D. Fowler, M. Coyle, U. Skiba, M. A. Sutton, J. N. Cape, S. Reis, L. J. Sheppard, A. Jenkins, B. Grizzetti, J. N. Galloway, *et al.*, *The global nitrogen cycle in the twenty-first century,* Philosophical Transactions of the Royal Society B: Biological Sciences **368**, 20130164 (2013).

[5] J. W. Erisman, A. Vermeulen, A. Hensen, C. Flechard, U. Dämmgen, D. Fowler, M. Sutton, L. Grünhage, and J. P. Tuovinen, *Monitoring and modelling of biosphere/atmosphere exchange of gases and aerosols in Europe,* Environmental Pollution **133**, 403 (2005).

[6] Y. Jia, G. Yu, Y. Gao, N. He, Q. Wang, C. Jiao, and Y. Zuo, *Global inorganic nitrogen dry deposition inferred from ground- and space-based measurements.* Scientific Reports **6**, 19810 (2016).

[7] R. A. Duce, J. LaRoche, K. Altieri, K. R. Arrigo, A. R. Baker, D. G. Capone, S. Cornell, F. Dentener, J. Galloway, R. S. Ganeshram, R. J. Geider, T. Jickells, M. M. Kuypers, R. Langlois, P. S. Liss, S. M. Liu, J. J. Middelburg, C. M. Moore, S. Nickovic, A. Oschlies, T. Pedersen, J. Prospero, R. Schlitzer, S. Seitzinger, L. L. Sorensen, M. Uematsu, O. Ulloa, M. Voss, B. Ward, and L. Zamora, *Impacts of atmospheric anthropogenic nitrogen on the open ocean,* Science **320**, 893 (2008).

[8] J. W. Erisman, J. N. Galloway, S. Seitzinger, A. Bleeker, N. B. Dise, A. M. R. Petrescu, A. M. Leach, and W. de Vries, *Consequences of human modification of the global nitrogen cycle.* Philosophical transactions of the Royal Society of London: Series B, Biological Sciences **368**, 20130116 (2013).

[9] R. Bobbink, K. Hicks, J. Galloway, T. Spranger, R. Alkemade, M. Ashmore, M. Bustamante, S. Cinderby, E. Davidson, F. Dentener, B. Emmett, J.-W. Erisman, M. Fenn, F. Gilliam, a. Nordin, L. Pardo, and W. De Vries, *Global assessment of nitrogen deposition effects on terrestrial plant diversity: a synthesis.* Ecological applications : a publication of the Ecological Society of America **20**, 30 (2010), arXiv:9403058 [gr-qc] .

[10] E. C. Farrer and K. N. Suding, *Teasing apart plant community responses to N enrichment: the roles of resource limitation, competition and soil microbes,* Ecology letters **19**, 1287 (2016).

[11] L. C. Maskell, S. M. Smart, J. M. Bullock, K. Thompson, and C. J. Stevens, *Nitrogen deposition causes widespread loss of species richness in British habitats,* Global Change Biology **16**, 671 (2010).

[12] S. M. Simkin, E. B. Allen, W. D. Bowman, C. M. Clark, J. Belnap, M. L. Brooks, B. S. Cade, S. L. Collins, L. H. Geiser, F. S. Gilliam, S. E. Jovan, L. H. Pardo, B. K. Schulz, C. J. Stevens, K. N. Suding, H. L. Throop, and D. M. Waller, *Conditional vulnerability of plant diversity to atmospheric nitrogen deposition across the United States,* Proceedings of the National Academy of Sciences of the United States of America **113**, 4086 (2016).

[13] C. J. Stevens, N. B. Dise, J. O. Mountford, and D. J. Gowing, *Impact of Nitrogen Deposition Grasslands,* Science **303**, 1876 (2004).

[14] S. E. Koerner, M. L. Avolio, K. J. La Pierre, K. R. Wilcox, M. D. Smith, and S. L. Collins, *Nutrient additions cause divergence of tallgrass prairie plant communities resulting in loss of ecosystem stability,* Journal of Ecology **104**, 1478 (2016).

[15] J. A. Geddes and R. V. Martin, *Global deposition of total reactive nitrogen oxides from 1996 to 2014 constrained with satellite observations of NO2 columns,* Atmospheric Chemistry and Physics **17**, 10071 (2017).

[16] M. Lawrence, W. Chameides, P. Kasibhatla, H. Levy, and W. Moxim, *Lightning and atmospheric chemistry: The rate of atmospheric no production, Handbook of Atmospheric Electrodynamics, Volume I*, , 189 (2017).

[17] P. H. Wine, *1997 atmospheric chemistry colloquium for emerging senior scientists,* (1998).

[18] X. Tie, R. Zhang, G. Brasseur, L. Emmons, and W. Lei, *Effects of lightning on reactive nitrogen and nitrogen reservoir species in the troposphere,* Journal of Geophysical Research: Atmospheres **106**, 3167 (2001).

[19] S. Krupa, *Effects of atmospheric ammonia (nh3) on terrestrial vegetation: a review,* Environmental pollution **124**, 179 (2003).

[20] C. F. Álvarez, L. M. Berrouet, M. E. Chaves, G. A. Corzo Mora, I. Gil, R. Gómez, A. González, V. González, R. Peñuela, W. Ramírez, *et al.*, *Evaluación nacional de biodiversidad y servicios ecosistémicos de colombia: resumen para tomadores de decisión,* (2023), http://hdl.handle.net/20.500.11761/35942.

[21] 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 (2020), 10.1007/s11356-020-08268-4.

[22] A. Casallas, N. Celis, C. Ferro, E. L. Barrera, C. Peña, J. Corredor, and M. B. Segura, *Validation of pm 10 and pm 2.5 early alert in bogotá, colombia, through*

the modeling software wrf-chem, Environmental Science and Pollution Research , 1 (2020).

[23] J. Verkaik and L. Ganzeveld, *Evaluation of Colombian Methane Emissions Combining WRF-Chem and TROPOMI* (Wageningen University and Research, Meteorology and air quality, MSc Thesis, 2019).

[24] J. D. Maasakkers, D. J. Varon, A. Elfarsdóttir, J. McKeever, D. Jervis, G. Mahapatra, S. Pandey, A. Lorente, T. Borsdorff, L. R. Foorthuis, *et al.*, *Using satellites to uncover large methane emissions from landfills,* Science advances **8**, eabn9683 (2022).

[25] 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**, 928 (2020).

[26] J. F. Grajales and A. Baquero-Bernal, *Inference of surface concentrations of nitrogen dioxide (no2) in colombia from tropospheric columns of the ozone measurement instrument (omi),* Atmósfera **27**, 193 (2014).

[27] J. G. M. Barten, L. N. Ganzeveld, A. J. Visser, R. Jiménez, and M. C. Krol, *Evaluation of nitrogen oxides sources and sinks and ozone production in colombia and surrounding areas,* Atmospheric Chemistry and Physics Discussions **2019**, 1 (2019).

[28] R. Timmermans, R. Kranenburg, A. Manders, C. Hendriks, A. Segers, E. Dammers, Q. Zhang, L. Wang, Z. Liu, L. Zeng, *et al.*, *Source apportionment of pm2. 5 across china using lotos-euros,* Atmospheric Environment **164**, 370 (2017).

[29] K. Petersen, G. P. Brasseur, I. Bouarar, J. Flemming, M. Gauss, F. Jiang, R. Kouznetsov, R. Kranenburg, B. Mijling, V.-H. Peuch, *et al.*, *Ensemble forecasts of air quality in eastern china-part 2: Evaluation of the marcopolo-panda prediction system, version 1,* Geoscientific Model Development **12**, 1241 (2019).

[30] T. Kociok, F. February, P. Grossmann, D. Sprung, A. van Eijk, and K. Stein, *Investigation of aerosols in south africa: Comparison of measurements and modeling,* in *Environmental Effects on Light Propagation and Adaptive Systems*, Vol. 10787 (International Society for Optics and Photonics, 2018) p. 1078704.

[31] V. Marécal, V. H. Peuch, C. Andersson, S. Andersson, J. Arteta, M. Beekmann, A. Benedictow, R. Bergstrem, B. Bessagnet, A. Cansado, F. Chiroux, A. Colette, A. Coman, R. L. Curier, H. A. C. D. Van Der Gon, A. Drouin, H. Elbern, E. Emili, R. J. Engelen, H. J. Eskes, G. Foret, E. Friese, M. Gauss, C. Giannaros, J. Guth, M. Joly, E. Jaumouill, B. Josse, N. Kadygrov, J. W. Kaiser, K. Krajsek, J. Kuenen, U. Kumar, N. Liora, E. Lopez, L. Malherbe, I. Martinez,

D. Melas, F. Meleux, L. Menut, P. Moinat, T. Morales, J. Parmentier, A. Piacentini, M. Plu, A. Poupkou, S. Queguiner, L. Robertson, L. Roul, M. Schaap, A. Segers, M. Sofiev, L. Tarasson, M. Thomas, R. Timmermans, Valdebenito, P. Van Velthoven, R. Van Versendaal, J. Vira, and A. Ung, *A regional air quality forecasting system over Europe: The MACC-II daily ensemble production,* Geoscientific Model Development **8**, 2777 (2015).

[32] C. Hendriks, R. Kranenburg, J. Kuenen, R. van Gijlswijk, R. Wichink Kruit, A. Segers, H. Denier van der Gon, and M. Schaap, *The origin of ambient particulate matter concentrations in the Netherlands,* Atmospheric Environment **69**, 289 (2013).

[33] G. P. Brasseur, Y. Xie, K. Petersen, I. Bouarar, J. Flemming, M. Gauss, F. Jiang, R. Kouznetsov, R. Kranenburg, B. Mijling, *et al.*, *Ensemble forecasts of air quality in eastern china-part 1: Model description and implementation of the marcopolo-panda prediction system, version 1,* Geoscientific Model Development **12**, 33 (2019).

[34] A. Dominguez-Rodriguez, N. Baez-Ferrer, S. Rodríguez, P. Avanzas, P. Abreu-Gonzalez, E. Terradellas, E. Cuevas, S. Basart, and E. Werner, *Saharan dust events in the dust belt-canary islands-and the observed association with in-hospital mortality of patients with heart failure,* Journal of Clinical Medicine **9**, 376 (2020).

[35] 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).

[36] S. C. Van Der Graaf, E. Dammers, M. Schaap, and J. W. Erisman, *Technical note: How are NH_3 dry deposition estimates affected by combining the LOTOS-EUROS model with IASI-NH_3 satellite observations?* Atmospheric Chemistry and Physics , 1 (2018).

[37] M. Wesely, *Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models,* Atmospheric Environment (1967) **23**, 1293 (1989).

[38] L. Zhang, S. Gong, J. Padro, and L. Barrie, *A size-segregated particle dry deposition scheme for an atmospheric aerosol module,* Atmospheric Environment **35**, 549 (2001).

[39] M. Van Zanten, F. Sauter, W. K. RJ, J. Van Jaarsveld, and W. Van Pul, *Description of the depac module: Dry deposition modelling with depac_gcn2010,* RIVM rapport 680180001 (2010), 680180001.

[40] 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, *LOTOS-EUROS v2.0 Reference Guide. TNO 2016 R10,* TNO Innovation

for life (2016), https://airqualitymodeling.tno.nl/publish/pages/3175/lotos-euros$_u ser - guide_v 2 - 2 - 002.pdf$.

[41] J. G. Droppo, *Improved formulations for air-surface exchanges related to national security needs: dry deposition models,* (2006).

[42] H. Wanner, J. Beer, J. Bütikofer, T. J. Crowley, U. Cubasch, J. Flückiger, H. Goosse, M. Grosjean, F. Joos, J. O. Kaplan, *et al.*, *Mid-to late holocene climate change: an overview,* Quaternary Science Reviews **27**, 1791 (2008).

[43] A. M. Manders, P. J. Builtjes, L. Curier, H. A. Denier van der Gon, C. Hendriks, S. Jonkers, R. Kranenburg, J. J. Kuenen, A. J. Segers, R. Timmermans, *et al.*, *Curriculum vitae of the lotos–euros (v2. 0) chemistry transport model,* Geoscientific Model Development **10**, 4145 (2017).

[44] A. Manders-Groot, A. Segers, S. Jonkers, M. Schaap, R. Timmermans, C. Hendriks, F. Sauter, R. W. Kruit, E. van der Swaluw, and H. Eskes, *Lotos-euros v2. 0 reference guide,* TNO report TNO2016 **10898** (2016).

[45] K. R. Segers A, Manders-Groot AMM, *Lotos-euros v2.2002 use guide,* TNO report TNO2019 (2021).

[46] 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).

[47] 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).

[48] 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).

[49] 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).

[50] J. W. Kaiser, A. Heil, M. O. Andreae, A. Benedetti, N. Chubarova, L. Jones, J. J. Morcrette, M. Razinger, M. G. Schultz, M. Suttie, and G. R. Van Der Werf, *Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power,* Biogeosciences **9**, 527 (2012).

[51] E. C. Monahan, D. E. Spiel, and K. L. Davidson, *A Model of Marine Aerosol Generation Via Whitecaps and Wave Disruption,* Oceanographic Sciences Library SPRINGER **2**, 167 (1986).

[52] E. Mårtensson, E. Nilsson, G. de Leeuw, L. Cohen, and H.-C. Hansson, *Laboratory simulations and parameterization of the primary marine aerosol production,* Journal of Geophysical Research: Atmospheres **108**, AAC 15 (2003).

[53] S. C. van der Graaf, R. Kranenburg, A. J. Segers, M. Schaap, and J. W. Erisman, *Satellite-derived leaf area index and roughness length information for surface–atmosphere exchange modelling: a case study for reactive nitrogen deposition in north-western europe using lotos-euros v2.0,* Geoscientific Model Development **13**, 2451 (2020).

[54] S. Fritz, E. Bartholome, A. Belward, A. Hartley, H.-J. Stibig, H. Eva, P. Mayaux, S. Bartalev, R. Latifovic, S. Kolmert, *et al.*, *Harmonisation, mosaicing and production of the global land cover 2000 database,* JRC report EUR **20849** (2003), https://publications.jrc.ec.europa.eu/repository/handle/JRC26168.

[55] A. Di Gregorio, *Land cover classification system: classification concepts and user manual: LCCS*, Vol. 2 (Food & Agriculture Org., 2005).

[56] H. Eva, E. Miranda, and C. Di Bella, *The Geographical Journal*, Vol. 148 (2002) p. 287.

[57] P. Defourny, I. Moreau, and S. e. a. Bontemps, *Land Cover CCI PRODUCT USER GUIDE VERSION 1.1*, Tech. Rep. (UCL, Gamma remote sensing, 2017).

[58] 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).

[59] 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).

3

Measurements, the input needed in data assimilation for improving models performance

'You see, but you do not observe. The distinction is clear. '
-A Scandal in Bohemia

Sherlock Holmes

Data assimilation is a mathematical technique that reconciles the mathematical modelling representation of reality with the measurement perspective. Measurements are essential in any Data Assimilation (DA) system. Satellite information is becoming increasingly valuable as new applications emerge. The increase in satellite data availability presents novel monitoring prospects, especially in areas where ground-based measurement systems are insufficient, lack operational network infrastructure, or are infeasible to deploy due to geographical constraints.

There exists a legacy of satellites dedicated to researching atmospheric composition, with an enhancement of spatio-temporal resolution over time. However, incorporating satellite products into data analysis is a complex task that requires appropriate data preparation to ensure the accurate comparison with the observed model states. This chapter outlines the methodology

Parts of this chapter have been published in - Yarce Botero, Andrés, et al. "Design and Implementation of a Low-Cost Air Quality Network for the Aburra Valley Surrounding Mountains." Pollutants 3.1 (2023): 150-165. https://doi.org/10.3390/pollutants3010012

used to handle satellite information, from data download to processing, for effective comparison with models and subsequent assimilation purposes.

This chapter outlines the design and manufacturing process of a low-cost device that can measure air quality variables in-situ and transmit data over long distances. The objective behind developing this device is to gather new information to integrate into the LOTOS-EUROS in this region. As there is a scarcity of data in Colombia, the development of new technologies to generate it has been emphasized throughout this thesis.

Using an adjoint-free technique, this chapter describes the process for implementing the Local Ensemble Kalman Filter to assimilate satellite data and the development of assimilating these low-cost measurements.

3.1. Introduction

Data Assimilation (DA) is a mathematical technique that reconciles the mathematical modeling representation of reality with the measurements, these applications will be presented in Chapter 4 of this thesis. In this Chapter, two measurements are described. One is the TROPOMI (TROPOspheric Monitoring Instrument [1]) which is on board the Sentinel 5-Precursor satellite, and the other is a low-cost, in-house developed sensor device used to monitor pollutants in Colombia, specifically in the Aburrá Valley.

3.1.1. Satellite information to study atmospheric composition

Colombia is demarcated by the Andes mountain range divided into three branches, with peaks up to 5750 m.a.s.l., a suitable territory for generating clouds. This terrain is highly rough with variable heights covered with rainforest, a highly appropriate place for cloud formations. The problem of cloudiness is crucial for the retrieval of pollutants. To understand the spatial distribution of the high quantity of available data for a period of time, an early study was made. Figure 3.1 shows the data comparison density for the Colombian territory vs. the Netherlands. This image was generated with the count of available pixels for each grid during a period. It can be seen how the prevailing clouds over the Andes mountains reduce the number of available observations. Cities on the Caribbean north coast of Colombia have more observations per month than inner cities. Magdalena River Valley and Orinoquia sector also show many available observations.

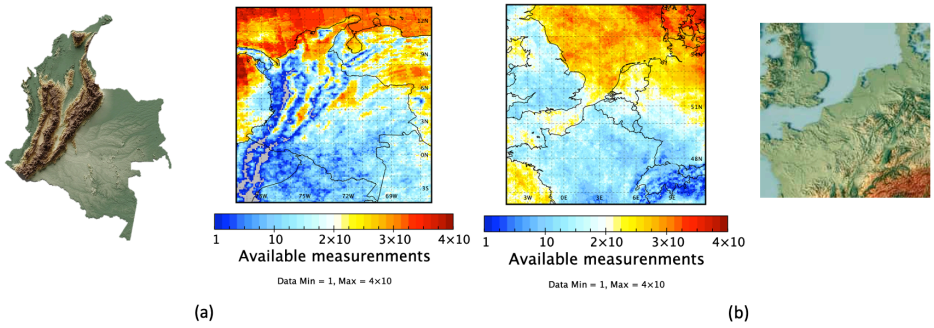


Figure 3.1: Comparison of available measurement for a month with the topographic surface of Colombia and the Netherlands. Satellite pixels are less available during a period, mainly over mountainous terrain, which is correlated to cloud formation.

Most of Colombia's air quality ground networks are located in the principal cities and urban areas, having most of the territory uncovered. The high-density information of NO_2 instruments, such as the TROPOspheric Monitoring Instrument TROPOMI, onboard the Copernicus Sentinel-5 Precursor satellite, increases the potential use of this data over this region. Sentinel-5 Precursor (S5P) is a low Earth polar orbit satellite. The scanner's polar orbit and wide coverage provide almost

daily global coverage. However, the high levels of cloudiness in the area continue to be a significant problem for retrieving pollutants. Satellite information is a valued resource for acquiring information; no matter how consistently good quality, satellite data is scarce, especially in regions with high cloudiness [2].

The volume of data is constantly increasing and improving in spatial and temporal resolution, particularly in instruments for atmospheric chemistry composition (see Chapter 1). This decade, 2020-2030, foresees almost complete planet monitoring with the geostationary network dedicated to air quality for the northern hemisphere. Thus, a cornerstone of state-of-the-art atmospheric modeling is incorporating real observations into simulations. In many applications, incorporating this data with numerical models induces parameter estimation capabilities that could overcome the complex tasks of developing emission inventories. Numerical model implementations are constantly improving in resolution thanks to the increasing availability of satellite observations and ground-based instruments.

3.1.2. Low-cost sensors

The densest network for measuring air pollutant concentrations in Colombia is in Medellín, where most sensors are located in the heavily polluted lower parts of the valley. Measuring stations in the higher elevations in the mountains surrounding the Aburrá Valley are unavailable due to the efforts in the city to get the best-measured representation in the populated areas. This limits our understanding of the valley's pollutant dynamics and hinders the effectiveness of data assimilation studies using Chemical Transport Models CTMs such as LOTOS-EUROS.

To address this measurement gap, we designed a network of low-cost sensors to be installed at altitudes above 2000 m.a.s.l. The network consists of custom-built, solar-powered, and remotely connected sensors. The locations were strategically selected using the LOTOS-EUROS model driven by various meteorological simulated fields to explore the effects of the valley wind representation on the transport of pollutants. The sensors transmit collected data to internet gateways for posterior analysis. Various tests to verify the critical characteristics of the equipment, such as long-range transmission modeling and experiments with an R score of 0.96 for the best propagation model, energy power system autonomy, and sensor calibration procedures, besides case exposure to dust and water experiments, to ensure IP certifications. An inter-calibration procedure was performed to characterize the sensors against reference sensors and describe the observation error to provide acceptable ranges for the data assimilation algorithm (<10% nominal).

This air quality network's design, installation, testing, and implementation, oriented toward data assimilation over the Aburrá Valley, constitute an initial experience of the system's operative capabilities. Our solution approach adds value by removing the disadvantages of low-cost devices and offers a viable solution from a developing country perspective, employing hardware explicitly designed for the situation. The next section details data collection from the TROMOPI satellite and the specifics of the low-cost sensor network designed.

3.1.3. Parametrization

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

$$\mathbf{c}_t = \mathcal{M}_{LE}(\mathbf{c}_{t-1}, \mathbf{e}_{t-1}) \tag{3.1}$$

The initial condition is $\mathbf{c}_0$, and the state vector $\mathbf{c}_t$ contains the concentrations of all trace gases and aerosols in each cell of the three-dimensional grid valid for time t, $\mathbf{e}_t$ is the nominal emission from the emission inventory. The LOTOS-EUROS model operator $\mathcal{M}_{LE}$ computes the state at time t from the concentrations at one timestep Δt before at time $t-1$ using the model input data. Note that in the following equations, some arguments of $\mathcal{M}_{LE}$ have been omitted to simplify notations.

A parametrization for representing the model uncertainty is required to implement the data assimilation algorithm. A significant source of uncertainty is the emissions that, in reality, may differ vastly from the inventory in both space and time. Therefore, the emissions utilized by the model operator are modeled as a stochastic process:

$$\hat{\mathbf{e}}_t = \mathbf{e}_t \cdot (\, 1 + \delta \mathbf{e}_t \,) \tag{3.2}$$

The emission deviation is modeled as an autoregressive model of order one (AR-1) with the following structure of a colored noise process [3]:

$$\delta \mathbf{e}_t = \alpha \cdot \delta \mathbf{e}_{t-1} + \sigma \cdot \sqrt{1-\alpha^2} \cdot \mathbf{w}_{t-1} \tag{3.3}$$

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

$$\mathbf{w}_t \sim N(0, \mathbf{I}) \tag{3.4}$$

Stochastic samples are drawn from a normal distribution with zero mean and standard deviation σ. The temporal correlation coefficient $\alpha \in [0,1]$ 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 $\alpha = 1$ 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 [4]:

$$\alpha = \exp(-|\Delta t|/\tau) \tag{3.5}$$

A stochastic model state is formed by augmenting the state vector (3.1) with the correction factor $\delta \mathbf{e}_t$:

$$\begin{bmatrix} \mathbf{c}_t \\ \delta \mathbf{e}_t \end{bmatrix} = \begin{bmatrix} \mathcal{M}_{LE}\,(\mathbf{c}_{t-1}, \mathbf{e}_{t-1} \cdot (\, 1 + \delta \mathbf{e}_{t-1}\,)) \\ \alpha \cdot \delta \mathbf{e}_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ \sigma \cdot \sqrt{1-\alpha^2} \end{bmatrix} \mathbf{w}_{t-1} \tag{3.6}$$

or simply:

$$\mathbf{x}_t = \mathcal{M}(\mathbf{x}_{t-1}) + \mathbf{G} \cdot \mathbf{w}_{t-1} \tag{3.7}$$

It is feasible to employ a sequential data assimilation method to determine both the state and the correction factor for emission using the augmented vector (10). The non-linear operator $\mathcal{M}$ progresses the augmented state vector $\mathbf{x}_t$ in time, while $\mathbf{G}$ distributes the stochastic forcing $\mathbf{w}_t$ over the elements of the state.

3.2. Methods

3.2.1. TROPOMI Satellite Data

TROPOMI is a sun-synchronous orbit instrument that overpasses at altitudes around 800 km at 13:30 local time. On October 13, 2017, it was launched onboard the low earth orbit polar Sentinel 5 Precursor (S5P) satellite. The TROPOMI spatial pixel resolution is 5.5 x 3.5 km^2, and the NO_2 retrieval uses a wavelength range of 405–465 nm (spectral band 4). The NO_2 concentration retrieval and averaging kernels are necessary for converting TROPOMI data to compare appropriately with Chemical Transport Models (CTM) [5]. The most recent NO_2 satellite measurements for the study region are from the TROPOMI instrument, a spectrometer sensing ultraviolet (UV), visible (VIS), near (NIR), and short-wavelength infrared (SWIR) wavelengths to monitor Ozone (O_3), Methane (CH_4), Formaldehyde (CH_2), Aerosol, Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), and Sulfur Dioxide (SO_2). The Royal Netherlands Meteorological Institute (KNMI) created the TROPOMI NO_2 retrieval method based on the DOMINO NO_2 retrieval algorithm employed on the Ozone Monitoring Instrument (OMI) precursor instrument [6].

Although the high levels of cloudiness in the area continue to be a significant problem for retrieving pollutants, the high density of information of TROPOMI respects OMI increases the potential use of this data over the region. [7] and extends the capabilities from OMI. It has improved the retrieval values of the Vertical Column Densities VCD because of the no use of the intensity offset correction. TROPOMI observations have been applied in inversion studies, demonstrating the ability to track small-scale pollution and emissions [8].

3.2.2. TROPOMI retrieval algorithm

The steps to employ the TROPOMI data available take into account the previous procedures to have the L2 level geophysical products from the raw data in the L0 level to specific radiance and irradiance products in the L1 level. In the first stage, the overall density of the NO_2 slant column is determined using the Differential Optical Absorption Spectroscopy (DOAS) method. Compared to the light that comes from the path of the sunlight reflected from the earth to the direct sunlight reaching the satellite, the extinction of scattering and absorbing photons by tracers affects the reflected information in the spectrum [7]. For the computation of the NO_2 concentration, a radiance model is fitted to these various received spectra, assigning the slant column density of NO_2 as the efficient absorption of NO_2 along this channel.

The second stage divides the slant column into stratospheric and tropospheric parts. The stratospheric and tropospheric components of the slant column density are separated using data assimilation and the TM5-MP CTM. The mechanisms that

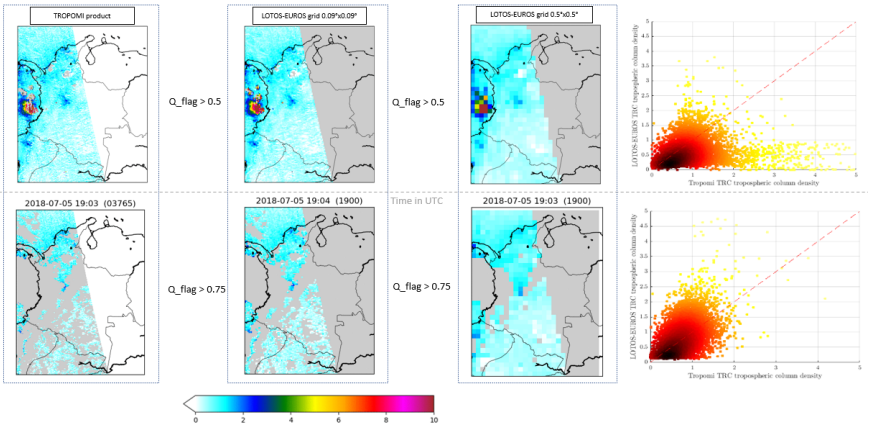


Figure 3.2: Illustration of the TROPOMI data downloaded before the regriding. *Upper panels*, data filtering a quality flag threshold of 0.5. *Lower*, data filtering at a quality flag threshold of 0.75. *Left column panels*, TROPOMI data retrievals for the two quality flags. *Center*, TROPOMI data projected to LOTOS-EUROS grids at resolutions of 0.09° (*center*) or 0.5° (*center-right*). *Right*, scatter plots comparing TROPOMI vs. LOTOS-EUROS pixel values for the resolution of 0.09°).

occur at the stratosphere and tropospheric levels are substantially distinctive. Most NO_2 in the slant column will be confined to the troposphere over highly polluted areas. However, stratospheric NO_2 dominates the slant column's density in remote areas like the oceans. The tropospheric column is calculated by subtracting the stratospheric column from the total column; note that this stage of the retrieval process leverages pre-existing knowledge regarding NO_2 columns from the TM5-MP model.

The final stage of the retrieval procedure is to convert slant column densities to vertical column densities, which is vital since the vertical column can readily be compared to the model's output. Slant columns are converted to vertical columns using air-mass factors (AMFs), denoted by the symbol M. The AMF establishes a link between the density of slant columns and the vertical density and defines the vertical profile of the trace gas. AMFs that are altitude-dependent are determined by retrieval parameters such as the satellite's viewing geometry, the surface albedo, the surface pressure, and cloud cover. These AMFs are used as prior information in the retrieval method since they are produced from a radiative transfer model.

TNO Dutch Organization for Applied Scientific Research (TNO https://www.tno.nl/) developed the module EMIP (EMEP Model Input Processor) for the satellite Sentinel 5P preprocessing task to produce the files to simulate the LOTOS-EUROS model and calculate the value for the appropriate comparison with the model output. The EMIP module evolved into what is known as CAMS Satellite Operator CSO, a toolbox designed to aid in integrating satellite observations into local air quality models. (https://ci.tno.nl/gitlab/cams/cso), which is software for preprocessing satellite information. The toolbox consists of two primary components: a pre-processor capable of downloading and transforming satellite data,

specifically TROPOMI data, and an observation operator that can be incorporated into the model simulation's source code. Using this operator, the module can simulate satellite retrievals and utilize them for data assimilation [9]. CSO handles the TROPOMI tropospheric NO_2 retrieval product (yr) as a profile spanning from the surface to 200 hPa. The first is the download task; once the date was introduced in the configuration script, the download starts from `www.temis.nl` for the TROPOMI Offline (OFFL) product and from `www.data-portal.s5p-pal.com` for the TROPOMI Products Algorithm Laboratory (PAL) product. In the second step, CSO creates netCDF files with selected pixels, such as those within some region or cloud-free pixels. The third step is regarding this task could resample the pixels onto a regular grid, distributing the footprint polygon over the grid cells. Each grid cell is filled with a weighted sum of contributions from pixels that (partly) overlap the cell.

For this comparison, the satellite quality flag defined in the EMIP module was set to 0.75 to avoid artifacts that do not necessarily correspond to NO_2 concentrations.

The S5P/TROPOMI tropospheric NO_2 data are gridded onto the simulation grid using a conventional area-weighted averaging approach. The relative mean is determined after averaging the observations' distance from each grid point. A model simulation of the satellite retrieval could be derived from the average kernel of the satellite product concentration profiles applied to the LOTOS-EUROS concentration to produce simulations of the gridded data. The average kernel integrates the concentrations over height, as the sensitivity of the satellite instrument to tracer densities is height-dependent. The model profile was convolved with the averaging kernels provided in the satellite download data product to simulate the retrieval. [5]. The averaging kernels are applied to the model output at the nearest satellite pass, usually between 11:00 and 12:00 UTC [10].

Satellite information is highly valuable for collecting data. However, it often encounters quality problems, which can be a persistent issue. Additionally, satellite data is scarce, especially in regions with heavy cloud cover [2]. Due to these high cloudiness levels, most of the region's satellite data is filtered out. The remaining pixels are collocated to the grid of the LOTOS-EUROS model ($0.09^\circ \times 0.09^\circ$) simulation to facilitate graphical comparison. Each grid cell is filled with an area-weighted sum of contributions from pixels that (partly) overlap the cell.

3.2.3. Design and Implementation of a Low-Cost Air Quality Network for the Aburrá Valley Surrounding Mountains

There is interest in developing a measurement network study to corroborate previous CTM representation of the valley and collecting data in these allocations for a future operational data assimilation (DA) start. In recent cases, as in [11] for estimating emissions in urban environments and in another case where the IoT-based (Wifi protocol) for suburban environments generates a network for measuring pollutants [12].

The proposed station locations are intended to gather information from areas that connect the urban regions of the Aburrá Valley and the San Nicolás Valley to the east, such as Rionegro. The proposed sites cover the highest regions along the

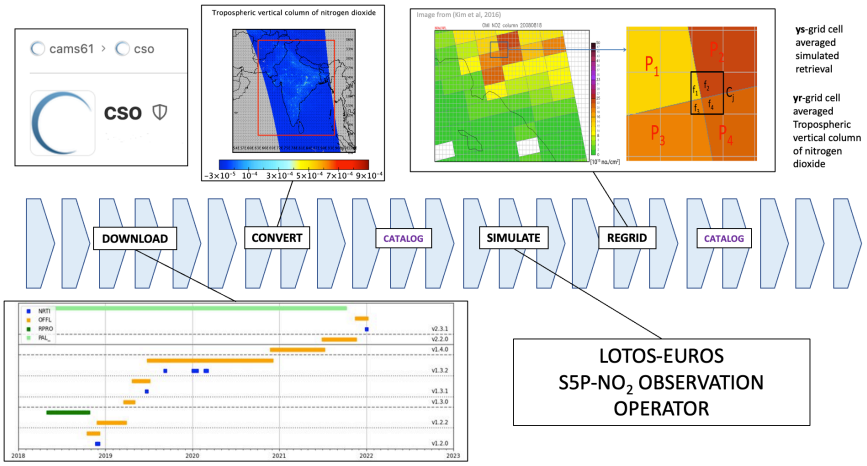


Figure 3.3: CSO pipeline procedure. It starts with downloading different TROPOMI product datasets and then converting or cropping them to a specific region. The CSO plots the steps of the data processed in the catalog. Then, the simulation goes through a model step for column integration, and last, the grid step produces the files to compare these observations directly with the model.

slopes of the Aburrá Valley. This will allow monitoring of the pollutant dynamics of the valley and understanding of the extent of these regional atmospheric linkages while providing sampling with high spatio-temporal resolution. SIATA (Sistema de Alerta Temprana del Valle de Aburr'a) manages the air quality measurement network in the Aburrá Valley (https://siata.gov.co/) and holds authority over varying early warning systems within the region. The Figure 3.4 simulation provides a basis for the proposed network configuration and offers stations that complement the SIATA station network. SIATA stations will serve as reference and calibration stations to ensure the dependability of the resulting data, as it constitutes the operational network for decision-making in the metropolitan area. The simulations employed the LOTOS-EUROS CTM to simulate a nested domain configuration to reach 0.01°× 0.01 °resolution for the particulate matter PM2.5, using two different meteorology input conditions (ECMWF and WRF). The results illuminate how pollutants are carried far away from the Valley by ECMWF meteorology, while for WRF meteorology, the Valley acts as a trap for contaminants.

In terms of low-cost sensors and applications with CTM, the work in [13] shows the comparison between a certified air quality network and a DA experiment with the low-cost sensor network by combining these two sources of information through DA schemes to balance the inaccuracy of these with the possibilities to drive a model with the right trend. In Figure 3.4, the comparisons of the three-dimensional snapshot of the model output over the valley and assimilated outputs of a low-cost sensor network inside the valley are depicted in transversal cuts.

The development of technology for monitoring air quality based on low-cost sensors is an increasing trend for small cities, harbors, and rural areas with increasing pollution-related problems [14–16]. Miniaturization of technology opens up new

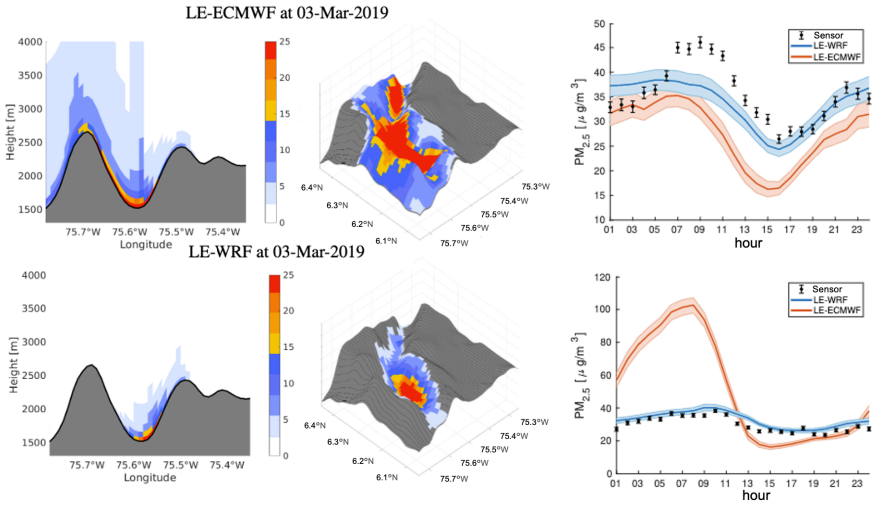


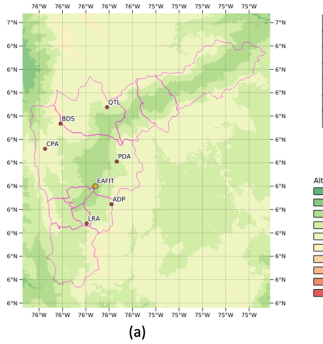
Figure 3.4: Left panel: transversal cut comparisons between the LOTOS-EUROS model simulated with ECMWF and WRF input meteorology. Central panels: a three-dimensional snapshot of the model output over the valley. The colorbar is for the particulate matter concentrations in $\mu g m^{-3}$. Right panels: assimilated outputs of a low-cost sensor network inside the valley with the two meteorology configurations for the LOTOS-EUROS (LE-WRF and LE-ECMWF), respect the observation network.

modular integration opportunities [17]. In this case, robust and rugged aerospace-inspired low-cost monitoring stations were deployed in strategic remote sites to detect the exchange of atmospheric contamination among regions. The proposed network, nested in a cyber-physical system [18], integrates new and existing data into a framework for understanding regional dynamics, evaluating development scenarios, and supporting decision-making and citizen science at the local and regional scales[19]. A review of low-cost sensors for outdoor quality applications is [20] and includes helpful information to understand the difference between existing ways to measure the standard sensor.

Description of the network

At the heights around the Aburrá Valley, measurements are necessary for the CTM comparison. The locations of the network were selected because they are city-representative locations and can be accessed by hill walking routes. Therefore, the nodes are considered energy-autonomous and remotely connected. The species of interest will be particulate matter ($PM_{2.5}$), nitrogen oxides (NO_x), and ozone (O_3), in addition to standard meteorological variables such as relative humidity, temperature, wind magnitude, and speed. However, other compounds, such as ammonia and isoprene, may be interesting to monitor in the future.

The proposed location of the measuring stations is presented in Figure 3.5. These locations are primarily selected for coverage of remote sites not covered by the SIATA network but that may provide data to detect the transport of contaminants in and out of the Aburrá Valley. Because of the urban growth trends, there will



(a)

Station name	Code	Latitude	Longitude
Cerro padre Amaya	CPA	6° 16′ 47.6796" N	-75° 41′ 14.2362" W
Baldias	BDS	6° 20′ 1.3632" N	-75° 39′ 14.7312" W
Quitasol	QTL	6° 22′ 8.5908" N	-75° 33′ 16.2318" W
Pan de Azucar	PDA	6° 15′ 10.5228" N	-75° 32′ 0.6432" W
La Romera	LRA	6° 7′ 12.2988" N	-75° 35′ 52.6704" W
Alto de Palmas	ADP	6° 9′ 40.8414" N	-75° 32′ 42.1152" W

(b)

(c)

Figure 3.5: (a) Locations selected for the air quality network disposition around the mountains surrounding Medellin. Colorbar indicates height over sea level in meters. The red cross is located at the point of view in the right direction from the arrow of the (c) photo. (b) table with the name, code, and coordinates of the stations. (c) Photography of the landscape seen from Medellín to the east mountains of the valley. The Pan De Azúcar (PDA) station location is indicated with an arrow.

be an emphasis on stations that may provide data to understand the atmospheric interconnections between the Aburrá Valley and the San Nicolás Valley.

The panel in Figure 3.6 points to the need for a measurement station in the high levels of the Aburrá Valley from the perspective of a previous DA scenario Figure 3.6. It shows the LOTOS-EUROS model assimilated with the $PM_{2.5}$ high-density network. It is crucial to notice how almost all stations from this assimilation network are below the 2000 m.a.s.l. The results shown in Figure 3.6 showcase the need to collect more information at high altitudes. An example to support this affirmation is seen for the star station, which is possible to see in the right part above the valley (yellow circle), for which the value in both situations for the model was underestimated, not presenting a proper update of the analysis for this area suggesting an increase in the number of observations.

With the measurement tools discussed here (TROPOMI satellite data and low-cost sensor array), we proceed to present them in more detail in the next section.

3.2.4. Airborne data collection

The low-cost sensor developed has been tested in airborne data collection campaigns. The ability to georeference the data means that mobile air quality data can be collected and georeferenced by these units in different conditions. This section presents the two data collection campaigns, one in a helicopter and the other in an aeroplane.

Hellicopter Insitu Pollution Acquisition Experiment

The HIPAE mission took place on 12 March 2019, from 9:30 am to 12 noon local time. Figure 3.7 shows the helicopter's flight path over the city of Medellín. The designed pattern of the flight was at three different altitudes of 600, 1200 and 1800 ft. The aircraft used was the helicopter (Airborne) of the Fuerza Aerea Colombiana (Colombian Air Force), UH-60L FAC 4121, a utility helicopter manufactured in the

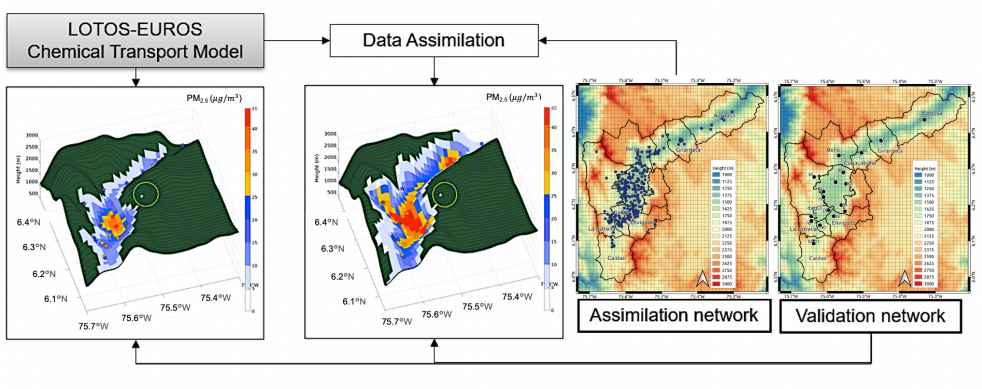


Figure 3.6: Comparison of the LOTOS-EUROS output concentrations before and after the Data assimilation over the 3D topography of the Aburrá Valley. The new measurements in the high part of the valley complement the SIATA networks for assimilation or validate.

USA by Sikorsky Aircraft Industries. The capacity of 22,000 lbs under external load configuration was used to allow an external sling load of 60ft.

Airborne data collection

Between the first of November and March 2021, 15 routes were flown to take measurements on a Colombian Air Force Caravan aircraft Figure Figure 3.8. This unpressurized aircraft flies humanitarian routes daily, and the sensor developed was attached to the cabin as a secondary payload to get data for the purpose of this thesis.

In Figure 3.9, we present a crucial aspect of our study involving integrating aircraft data with the LOTOS-EUROS model using the LEnKF (Local Ensemble Kalman Filter) framework. The aircraft dataset, characterized by its 3D plus temporal variability, provides high temporal resolution with data points sampled every minute. A regrinding process was performed to prepare the data with the simulation resolution of interest to enable meaningful comparison and assimilation of the data with the model. The standard LEnKF implementation in the LOTOS-EUROS needs the data in a specific structure, capturing both geopositioning and concentration magnitude information within its operational framework. In particular, the assimilated data, aggregated hourly and over-defined grid areas, play a key role in improving the accuracy and predictive capabilities of the model. This integration highlights the synergy between high-resolution real-world data and modeling techniques, providing valuable insights into atmospheric dynamics and composition.

3.3. Results

3.3.1. TROPOMI observations errors

Figure 3.10 shows an example of the grid satellite data, presenting how the NO_2 footprint demarcates the principal Colombian cities. The right panel shows an ex-

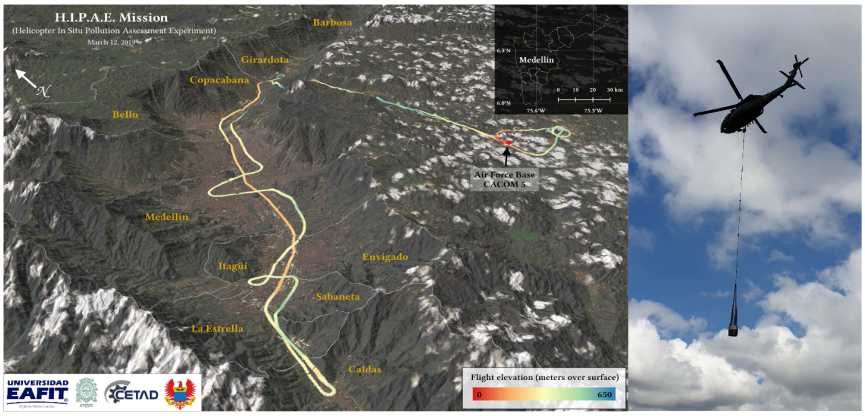


Figure 3.7: Left image is the ap of the mission's flight route, departing from Air Force Base CACOM-5 and entering the Aburrá Valley through the municipality of Copacabana Northeast of Medellín. Indicated in the image are the boundaries of the ten conurban municipalities of the Valley. Land cover derived from pan-sharpened Landsat 8 image (2018-04-09). 3-D rendering based on JAXA's 30 m ALOS Global Digital Surface Model (AW3D30), using QGIS 3.6. Right image Helicopter (Airborne) of the Fuerza Aerea Colombiana (Colombian Air Force), UH-60L FAC 4121,

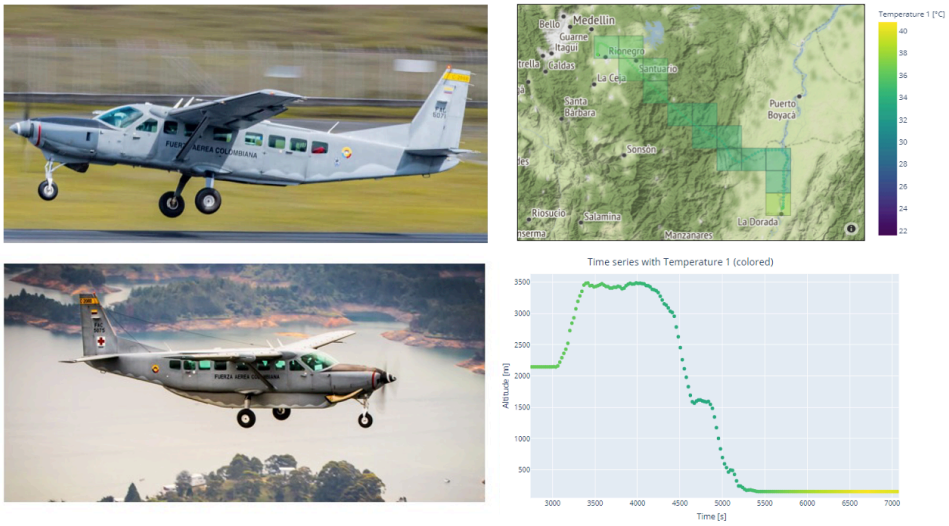


Figure 3.8: The left-hand column of this panel displays two images of the Caravan aircraft utilized to transport the developed sensor. The right-hand column depicts a 2D representation of the spatial temperature pattern in the upper-right panel, with a time series of the same variable indicated in the lower-right corner using the same colormap.

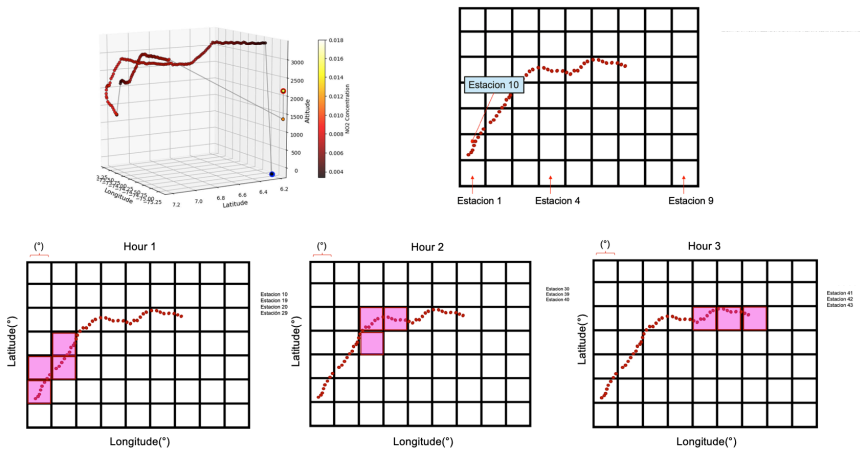


Figure 3.9: The data of the airplane is a 3D plus time variable that is sampled with high temporal resolution (1 value of data per minute), so we had to regrind it to the simulation resolution of interest in order to perform comparison or data assimilation with the model. By default, The LEnKF in the LOTOS-EUROS receives in situ measurements in a structure that permits recall of the geoposicioning and magnitude of the concentration once the filter is working. The data is hourly, and per grid averageded

ample of the quality flags qa values. A quality flag qa for the retrieval is a value provided for each pixel recovered; this number might be between 0 and 1, where 1 is the perfect quality data retrieved scene, and 0 is the worst. This quality qa value is mainly determined by the presence of clouds above the pixel. When the qa value is > 0.75, cloud-covered, partially snow/ice-covered errors and problematic retrievals are removed. A qa-value > 0.5 indicates good-quality retrievals over clouds and scenes covered with snow/ice, which is helpful for model comparison and assimilation studies. For this comparison, the quality flag qa was configured at 0.75 to avoid artifacts that do not necessarily correspond to NO_2 concentrations.

To see the effects due to the quality filter in this region, Figure 3.2 shows the comparison of the TROPOMI output in the first left column and the simulation kernel transformation for two different grid resolutions (0.09 °,0.5 °) and the scatter plot between the tropospheric column density of the satellite and the LOTOS-EUROS simulated column density, showing a large divergence between the satellite simulated data and the model output. After downloading the offline TROPOMI TM5-MP / DOMINO data from (`www.temis.nl`) [21], a pixel quality flag qa selection was applied. However, [7] recommends using pixels with a qa value of 0.52 or higher for data assimilation and model comparison studies.

Data assimilation must get the accuracy of the observations. Van Geffen *et al.* contains an in-depth examination of how uncertainty emerges throughout retrieval. Inaccuracies in the TROPOMI observations result from the three stages of the retrieval method described previously: errors in quantifying slant columns, errors in separating the stratospheric and tropospheric components of slant columns, and errors in tropospheric air mass factors. The overall error is provided in the TROPOMI

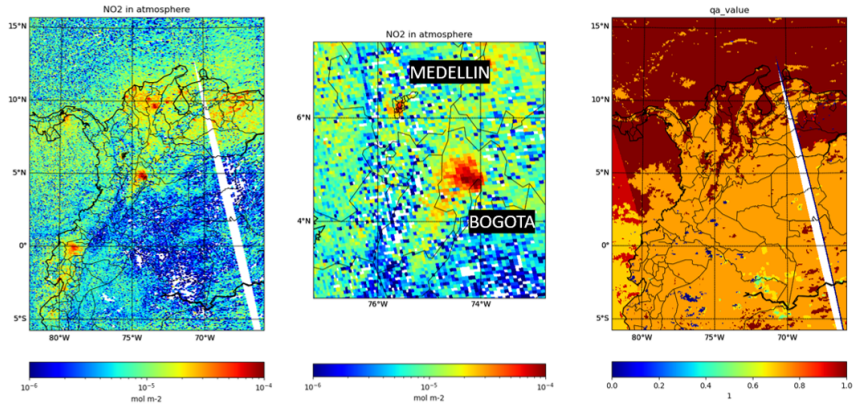


Figure 3.10: TROPOMI NO_2 vertical column density. *Right panel*, TROPOMI Offline level 2 data for the overpass over Colombia on 2019-04-02. *Center panel*, as before, zoomed over Bogotá and Medellí, the two largest (population-wise) Colombian cities. *Right panel*, quality flag values for the corresponding data.

data product.

For data assimilation, it is essential to get estimates of the accuracy of the observations. An in-depth assessment of how uncertainty arises throughout the retrieval is available in [7]. Slant column errors, errors in differentiating stratospheric and tropospheric components of slant columns, and tropospheric air-mass factors errors contribute to the overall error. In the TROPOMI data product, the total error per pixel is given.

Figure 3.11: Available pixels with NO_2 data for (a) April, (b) May and (c) June 2019 with a quality flag of 0.75 and a grid resolution of 0.05°. The areas of unavailable pixels clearly demarcate the Andean corridor during the month due to the high cloudiness of this mountainous area during these months, which is a common situation throughout the year.

In the figure presented, labeled (a), (b), and (c), we show the spatial distribution of available pixels containing NO_2 data for April, May, and June in the year 2019, respectively. These data are presented with a quality flag of 0.75, indicating a

reliable dataset, and have been spatially resolved to a grid resolution of 0.05°. What is striking about these visualizations is the clear delineation of areas where NO_2 data are unavailable, particularly within the Andean corridor. This lack of data is due to the persistent high cloud cover in this mountainous region during these months, a climatic characteristic that persists throughout the year. The figure illustrates the challenges of obtaining air quality information in regions with frequent cloud cover. It highlights the limitations and considerations essential for accurate environmental monitoring and assessment.

3.3.2. TROPOMI versions comparison

The comparison of the OFFL and PAL TROPOMI versions over Colombia shows differences that were compared with the Google Earth information. The Y_r is the satellite value. The comparison shows two points corresponding to an open pit mine in Colombia and a presumably large body of water with high eutrophication activity.

Figure 3.12: (a) shows the value of the OFFL version and (b) shows the PAL version and (c) the comparison of Y_r average tropospheric vertical column of nitrogen dioxide.

The Y_s is the model value integrated with the observations. Here an area that pointed out is located at the south of the country over a river due to the fluvial transport.

Figure 3.13: (a) shows the value of the OFFL version and (b) shows the PAL version, and (c) the comparison of Y_s grid cell averaged simulated retrieval column of nitrogen dioxide.

3.3.3. TROPOMI LEnKF Data Assimilation

Figure 3.14 shows a spatial comparison of NO_2 total column simulations over the Colombian domain, with and without the application of data assimilation techniques. The number of ensembles used in this experiment was 12, generated by applying stochastic noise to NO_x emission inventory values. The spatial localization of the filter was 15 km, and for the temporal correlation, 1 day was set to have persistence in the emissions, which means that the effect of the parameter update maintains a residual effect that is transferred to the following day until the next satellite overpass that generates a new update of the parameters.

In the upper left corner of the figure, we observe the full column simulation for NO_2 without the influence of data assimilation, showing the baseline representation. In the upper right corner, we show the simulation incorporating the TROPOMI data through the LEnTKF data assimilation method, demonstrating the improved accuracy achieved through assimilation [22, 23]. For this experiment, 12 ensemble members were used with a localization radius of 15 km. The emission factor parameters evolve in a stochastic model, which converges to the values suggested in the estimation process. The unassimilated dc parameters are shown in the lower-left corner, maintaining a consistent nominal value of 1 without affecting the baseline emissions. Meanwhile, we visually represent the parameter changes resulting from the data assimilation process in the lower right-hand corner, highlighting the refinements introduced. Finally, the lower middle panel shows the satellite data corresponding to this specific time, providing an additional perspective on the observed NO_2 levels in the region.

Figure 3.14: Figure 1: Spatial Comparison of NO_2 Total Column Simulation with and without Data Assimilation for the Colombia Domain. The total column simulation for NO_2 without data assimilation is shown in the upper left corner. The simulation assimilating TROPOMI data using the LEnKF data assimilation method is displayed in the upper right corner. The lower left corner illustrates the unassimilated dc parameters, which remain at a nominal value of 1 without affecting baseline emissions. The parameter modifications resulting from data assimilation are presented in the lower right corner. Satellite data is depicted in the lower center panel for this specific moment.

The following section explains the custom electronic development of the low-cost sensor device for air quality monitoring in the Aburrá valley, another observation resource to assimilate.

3.3.4. Low-cost sensor hardware architecture

The module used to develop this network is called Simple-4. This device's electronic architecture evolved from the Simple-3 architecture created in 2018 and the previous before (Simple-3, Simple-2, Simple-1, Simple-0) [24, 25]. The Simple Missions represent projects based on the CanSat development (a CanSat is a standard pico-satellite form factor soda can satellite), with a cylindrical array structure, a mass of approximately 250 grams, and a volume of approximately 330 cubic centimeters. While originally designed for deployment aboard rockets or weather balloons, the CanSat-inspired design results in rugged, modular, robust, non-invasive measuring devices that are highly efficient in their communication approaches and energy use.

The subsequent developments to this module facilitated the use of different communications modules and protocols, increased the number of sensors in the payload, and improved the energy system, which consists of an independent board from the previous payload subsystem. The attributes of the modules make them for deployment as remote measuring stations. Figure 3.15 shows the Simple-4 diagram

schematic system.

The different subsystems are explained below:

SimpleVital - On-Board Data Handling (OB&DH). The OB&DH manages, stores, and sends information from the other electronic subsystems to the ground segment through the communication subsystem. This board contains an 8-bit microcontroller that communicates employing standard protocols such as Serial, I2C (Inter-Integrated Circuit), and SPI (Serial Peripheral Interface) to peripheral units such as GPS (Global Positioning System), IMU (Inertial Measurement Unit), barometer, and temperature and communicates with the subsystems of EPS, Payload, and COMM (Communications). Furthermore, this subsystem is responsible for formatting the data for storage in SD memory. This four-layer Printed Circuit Board (PCB) design promotes the integration of more components per area.

SimplePower - Energy Power Subsystem (EPS). EPS regulates the power of satellite subsystems. The charge is stored in a battery bank with 20000 [mAh] in this module. This battery supplies between 3 and 4.2 volts, raised to 5 volts, to be distributed to other subsystems and a stage for overcharge protection. Additionally, this subsystem controls the power supplied by the solar cells (solar array interfaces) through a DET (Direct Energy Transfer) architecture; this enables the extension of the operation time of the module thanks to the additional power already available in the batteries.

SimplePollution - Payload Subsystem. This subsystem is composed of a sensor (MICS-6814) that measures the concentration of gases such as CO (Carbon monoxide), NO_2 (Nitrogen dioxide), C_2H_6OH (Ethanol), H_2 (Hydrogen), NH_3 (Ammonia), CH_4 (Methane), C_3H_8 (Propane), and C_4H_10 (Iso-butane), in addition to these two SPEC ™sensors for measuring NO_2 and O_3 thermodynamic variables such as relative humidity and temperature.

SimpleCOMM -Communications Subsystems. In particular, the Simple-4 design features three different types of radios (LoRa Tx / Rx @ 915 MHz, Dorji Tx / Rx @ 434 MHz, and Radiometrix Tx / Rx @ 434 MHz). Besides, it incorporates Wi-Fi technology. In terms of communication having different radio possibilities allows greater autonomy and modularity in the network and the capabilities to support long, medium and low-range transmissions. Two references were used, a 144 MHz Very High Frequency (VHF) and 434 Ultra High Frequency (UHF). Both frequencies used Narrow Band FM (NBFM) and Audio Frequency Shift Keying (AFSK) modulation at a 1200 baud velocity using the AX.25 (Amateur X.25) protocol, supporting the use of an Automatic Packet Reporting System (APRS) packets for real-time digital communications, thus permitting Tx/Rx coverage. Communication links generated between the Simple-4 CanSat picosatellites and the gateway station ensured a minimum vertical line-of-sight communication range of 300 meters without multipath propagation, rain fade, and attenuation for vegetation phenomena.

3.3.5. Hardware development

The simple-4 device is made up of 5 Printed Circuit Boards PCBs, each one of them housing a subsystem. The first is one 4-layer PCB called SimpleVital, corresponding to the Command On Board & Data Handling (CB&DH) subsystem. The second is

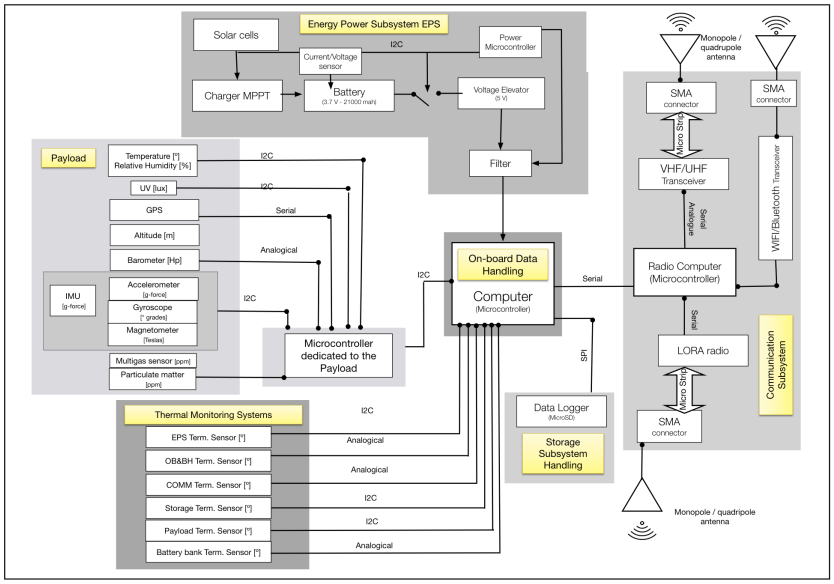


Figure 3.15: Flowchart with the various subsystems that comprise the Simple-4 partial centralized architecture. The Energy Power Supply (EPS) subsystem with the Maximum Power Point Tracker (MPPT) module controls the charge and discharge cycles of the battery and the load consumption. A dedicated microcontroller collects the payload information and, once the data is preprocessed, is delivered to the On Board and Data Handling (OB&DH) to command the communication subsystem. The other secondary system is the thermal monitoring system, a transversal support system conformed by the different thermal sensors in each Printed Circuit Board (PCB) layer.

a three 2-layer PCBs called SimplePayload (Gases / Pollution), the third one the SimplePower (Energy Power Supply (EPS)), the fourth is the SimpleBattery (Battery bank), and the last the SimpleCOMM (Communications Subsystems (COMM)). All those are shown in Figure 3.16

The information transmitted (TX) from the Simple units is received in a gateway (RX) that uploads the information to the cloud service. Different long-range tests were developed in the Aburrá Valley intending to guarantee the success of the network communication challenges as the abrupt terrain as well as possible obstacles, assuring the maximum distance of 18 km from the (TX) location in Bello (Baldias) to the gateway node (RX) in Universidad EAFIT.

3.3.6. Low-cost sensor network evaluation

The nodes established contact with the web server through the configured gateway in the Universidad EAFIT with an incremental distance to understand the system's operation under different types of antennas (Figure 3.17). In this process, two types of antenna, centered and not centered in frequency, were used in two configurations to determine the type of antenna with the best performance. The measurement of the intensity of the signal was compared with the theoretical electromagnetic signal propagation models.

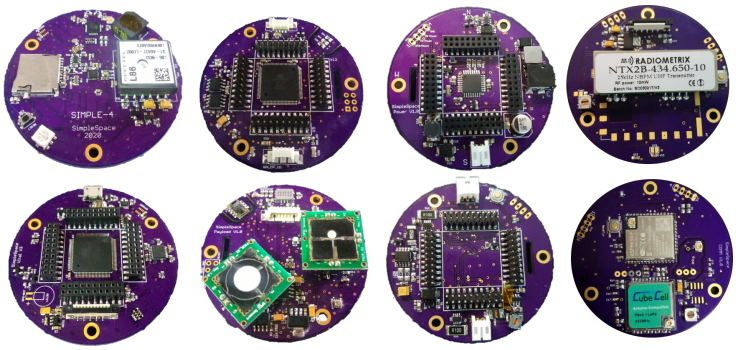


Figure 3.16: The two sides of the PCB, top and bottom, of the a) SimpleVital (Command On Board & Data Handling (CB&DH)),b) SimplePayload (Gases / Pollution),c) SimplePower (Energy Power Supply (EPS)),d) SimpleCOMM (Communications Subsystems (COMM). See the main text for description of the different subsystems.

Test Distance (m)	Location TX(Latitude/Longitude)	Name	Location RX(Latitude/Longitude)	Name
1 6910	6°15′10.9″ N /75°34′52.1″ W	Suramericana, Medellin	6°11′33.0″ N / 75°34′03.1 W	Los Balsos, Medellin
2 11,015	6°16′51.5″ N /75°36′50.1″ W	La Aurora, Medellin	6°11′33.0″ N / 75°34′03.1 W	Los Balsos, Medellin
3 15,998	6°16′51.5″ N /75°36′50.1″ W	La Aurora, Medellin	6°09 ′13.2″ N / 75°32′41.7 W	Alto las Palmas, Envigado

Table 3.1: Locations selected for the long-range transmission test. This gradual increase in successful experiments assures us of the long-range and stable communication link possible with the modules.

In Figure 3.17, in the left image, we examine the relationship between the vertical cut, earth contour, and the Fresnel distances. The x-axis corresponds to the distance in kilometers from the transmission point. In the center, it shows the spatial intensity in the Aburrá Valley with the transmission (TX) and reception (RX) points, and in the right image a model created using different signal intensities and antenna types at different experimental locations to understand the spatial transmission intensity patterns and develop an accurate model for the region.

We conducted different tests to ensure the module's long-term performance and reliability. A thermal camera monitors the electronic Printed Circuit Board (PCB) during complete operation cycles. This enabled us to detect deviations from the nominal operating temperature, as shown in Figure 3.18 (a). By identifying components that were not functioning properly, we were able to take proactive measures to prevent future malfunctions and energy losses.

Additionally, this testing allowed us to identify any potential issues, helping to ensure the longevity and effectiveness of the module. By regularly monitoring the temperature of the PCB and its components, we were able to catch any abnormalities early on and take corrective action to maintain the overall efficiency and reliability of the system. Overall, this testing was critical in ensuring the long-term

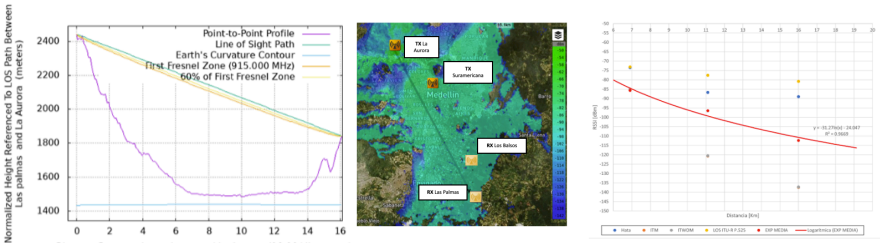


Figure 3.17: **a**) Vertical cut and Fresnel distance. (**b**) Spatial intensity over the Aburra Valley and (**c**) model generated from the different signal intensities and experimental locations.

capabilities of the module and safeguarding against costly and inconvenient malfunctions. Moreover, Figure 3.18 (b) and (c) show some images of the integrity tests carried out in the chamber for the IP5X IP6X evaluation. These tests are essential to ensure the long-term viability of the hardware because they expose the equipment to drastic conditions because the location sites surrounding the valley are not supposed to be visited frequently.

Figure 3.19 shows the time series comparison between the Vaisala AQT400 unit and the module developed for this project for the Ozone, NOx gasses, humidity, and temperature from two sensors. The data from the module overestimated the Vaisala measurements, and once the calibration algorithm is applied, the corrected value is also depicted.

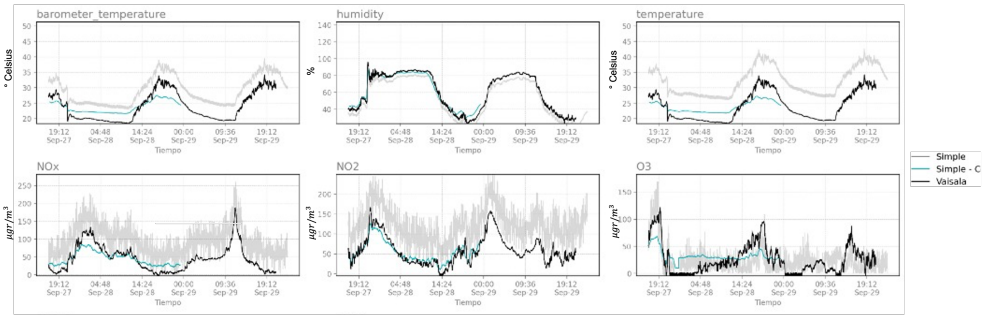


Figure 3.19: Time series comparison for Simple-4 (the grey line is the value measured, and the blue is corrected) for the barometer temperature, humidity, and temperature (upper panels) and NO_2 and O_3 (below panels) from the divide developed with the Vaisala AQT400 (black lines) ®calibrated equipment.

The inter-calibration procedure validated the performance of the sensor under investigation. The sensor was compared with reference sensors to determine observation error and accuracy. Results showed error margins within 10% of nominal, considered acceptable for use in the data assimilation algorithm. The inter-calibration contributed to the robustness of the study's findings by ensuring the reliability of the sensor data.

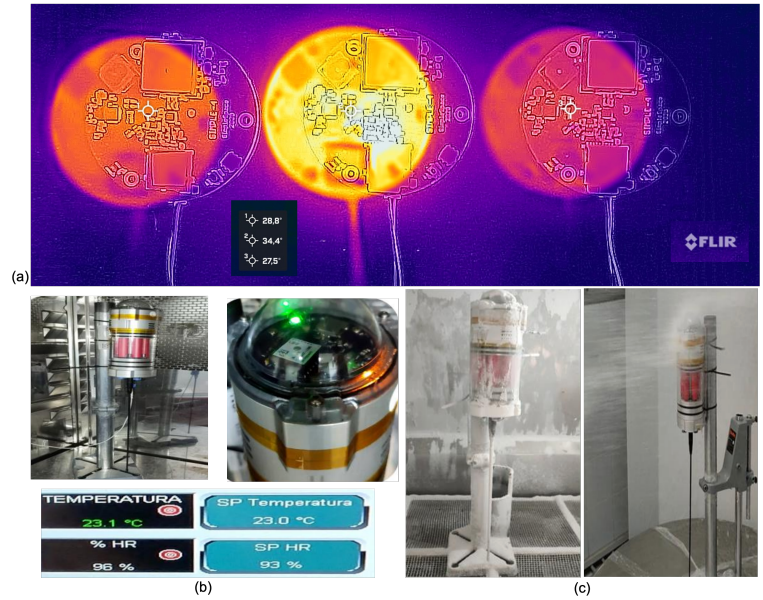


Figure 3.18: (a) Infrared thermal images under operations for different operation routines of the developed module. IP5X IP6X test for the device. (b) Temperature and Humidity stationary chamber test. (c) Dust and wind tests.

3.3.7. Airborne data acquisition, model comparison, and assimillation

The provided figure presents qualitative comparisons of particulate matter 10 (PM10) concentrations between the LOTOS-EUROS model and the low-cost sensor deployed during the HIPAE mission at varying altitudes. The altitudes are categorized into four ranges: (a) 0-1500 meters, (b) 1500-1850 meters, (c) 1850-2200 meters, and (d) 2200-2550 meters. The comparison highlights the agreement or disparities in PM10 levels between the two measurement methods, shedding light on the sensor's performance across different altitude bands. These visual insights contribute to a comprehensive understanding of PM10 distribution in the atmosphere over the Aburrá Valley

Figure 3.21 shows the NO_2 concentration through the flight trajectories of the Caravan aircraft, using assimilation techniques within the LOTOS-EUROS model. The upper panels show four different flight paths, with NO_2 concentrations in parts per million (ppm) displayed on a color scale. The lower left corner introduces a grid labeling system for precise spatial referencing and averaging of measurement points. In the middle panel, a composite of airborne measurements, assumed to be collected on the same day, provides a view of the NO2 distribution and prepares the data for the assimilation experiment. Finally, the lower right corner highlights the column difference resulting from the assimilation of airborne data into the LOTOS-EUROS model, highlighting the importance of data assimilation in refining

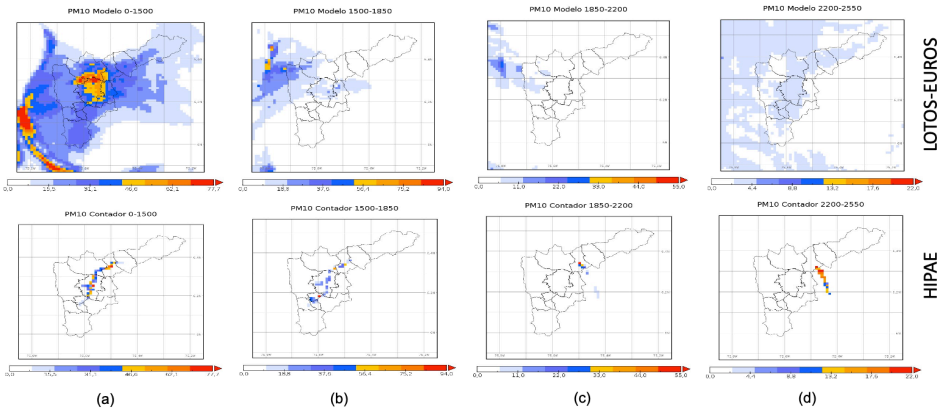


Figure 3.20: Comparisons at different altitudes between the LOTOS-EUROS and the low-cost sensor carried in the HIPAE mission for altitudes between 0-1500 (a), 1500-1850 (b), 1850-2200 (c), and 2200-2550 (d) for concentrations of particulate matter 10.

atmospheric modeling and elucidating the impact on NO2 dispersion dynamics.

3.4. Discussion and conclusion

The previous modeling of atmospheric composition dynamics in the Aburrá Valley showed inconsistencies in concentration and deposition fields based on the meteorological input to the CTM. Accurate meteorology suggests a reduced simulated concentration from the previous ECMWF meteorology used for this valley. Although emissions and meteorology are relevant dynamics in these models, the detailed emissions inventories contain high uncertainty, so the developments of a DA system like this oriented to emission estimations bring a region a capability.

Air quality data at the top of the Valley's surrounding mountains are nonexistent, and the transport hypothesis from the valley still needs to be explored experimentally. In 3.19 the measurement unit's first deployment campaign in the valley's surrounding mountains is shown. A comparison of the particulate sensor (pm10) of the device against two sensors, one inside the valley (Est 295) and the other outside (Est 152), is shown in the right image. This low-cost sensor ground-based measurement could be a feasible solution to increase data availability from non-connected areas. It will be validated if evidence corroborates these contradictions in these difficult-to-reach areas.

Air quality modeling and monitoring are crucial for understanding the sources and distribution of pollution. Incorporating additional spatial and temporal information into model simulations can improve predictions. However, the lack of monitoring in certain areas and the dependence of particulate matter simulation impact on meteorological modeling accuracy pose challenges for fully understanding air pollution. These findings underscore the need for continued research in this field and motivate us to develop a low-cost monitoring system that enables us to get

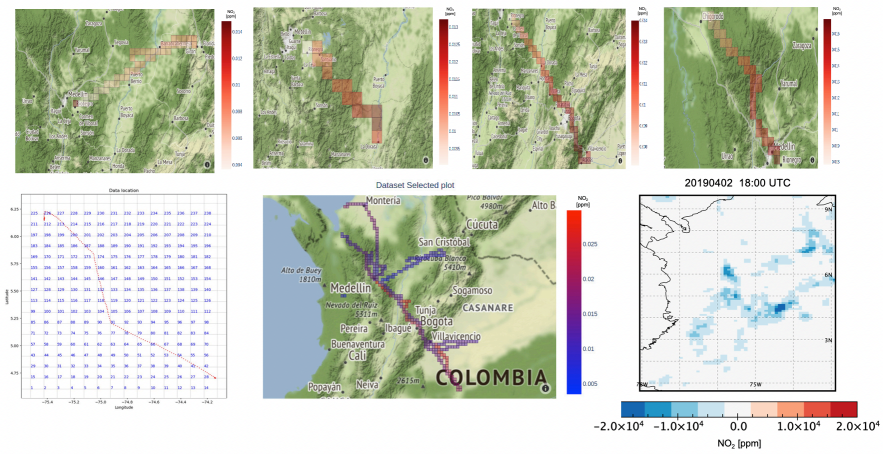


Figure 3.21: The figure shows an example of 4 flight trajectories in the upper panels, with the NO_2 variable shown in ppm in the color scale. In the lower-left corner, the grid label is used to mark the locations and average the values of the measurement points within. In the center bottom panel is a composition of the measurements; for the experiment, we assume that all airborne measurements were taken on the same day. In the lower-right corner, the column difference between the LOTOS-EUROS assimilation of the airborne measurements and the LOTOS-EUROS without assimilation.

this valuable information.

The simulation perspective gives some ideas for this kind of network to verify the predicted removal dynamics corroborating from the observation point of view through DA and validation activities. DA activities have been reviewed for this region in [26], analyzing the state-of-the-art and the subsequent state applicable to the Tropical Andes region. DA has been applied in the Aburrá Valley since 2019 [27] for a high-resolution experiment assimilating particulate matter observations with the LOTOS-EUROS CTM using an ensemble-based technique. From this first DA experiment, the urgent need to expand the data sources on atmospheric pollutant concentrations was identified to improve the performance models at local

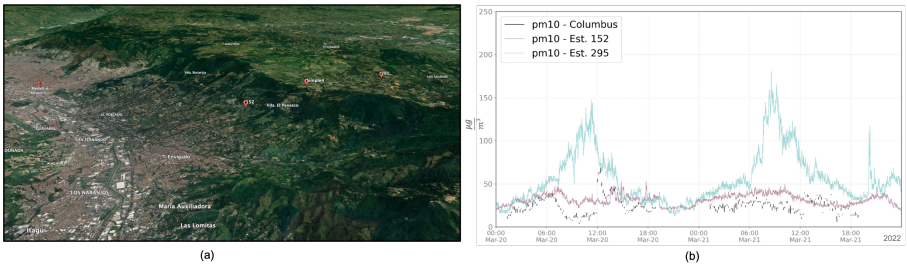


Figure 3.22: (a) Locations of a Simple 4 compared with two sensors of the SIATA network at two altitudes in the southeast part of the Aburrá Valley. (b) PM10 concentrations as a function of time for the three sensors, and this indicates the highest pollution levels inside the Valley.

and regional scales. Low-cost sensors are gaining importance through the years for managing air pollution in the cities [28]. Therefore, more and more electronic technology of this kind is gaining respect and complementing operative air quality networks. Problems such as calibration have recently been reviewed, and methodologies such as [29] can be applied to assess network data through calibration models or specific tools designed to evaluate low-cost gas sensors [**?**]. Also, the correction of hysteresis developed in [30] must be considered further for oxidative-based sensors.

The network is designed to complement the existing air quality network in the area based on various considerations and experiments to test the hardware equipment. This rural network has design challenges such as energy autonomy, case protection against water and dust for electronic equipment, and long-range data transmission. The telemetry signal was analyzed using four radio propagation models and field measurements, allowing a more comprehensive understanding of the network's behavior and signal behavior in the region with this topography. Simulations of energy charge and discharge cycles were used to determine the optimal configuration of solar cells and batteries, and additional studies were conducted on the IP degree of the system. The European Commission's Joint Research Centre (JRC) developed the PVGIS European power calculation tool (https://ec.europa.eu/jrc/en/pvgis). PVGIS uses data from meteorological models and satellite imagery to estimate the solar radiation a PV system receives at a given location, considering the local weather conditions, solar panel orientation, and tilt. This study demonstrated the utility of PVGIS for estimating the potential energy production of PV systems for a low-cost autonomous air quality sensor. Our results suggested that the valley's west side may be optimal for PV systems due to higher solar power availability. However, further research is needed to confirm and expand upon these findings because, in this part of the paper, we had simulated radiances for the past from PVGIS but not yet measured values in the installation points; nevertheless, the simulated values gave a helpful design input.

Airborne data assimilation is an emerging field in Colombia, and this work represents a pioneering effort. While it's a significant first step, further studies are required to validate the conditions for collecting in-situ data from small aircraft. Stemming from this thesis, a project was initiated to enhance the electronic hardware, exploring a range of sensors and optimizing the airflow into the equipment's measurement chamber.

There's a pressing need for further research in comparing data with models and other meteorological and air quality sensors. Establishing this innovative network in high-altitude tropical regions promises to offer crucial empirical data for data assimilation. Beyond this endeavor, there's an ambition to set up similar sensors in Colombia's untouched ecosystems. Taking advantage of the network's extensive transmission range and energy autonomy, we aim to remotely track the spread of urban pollutants. The network's geolocation features and data offerings can also promote data assimilation studies, aiming for heightened spatiotemporal resolution. By extending this network to more regions, we aspire to amplify its potential in monitoring urban pollution influx, thereby deepening our insight into the environmental

repercussions of these contaminants.

References

[1] J. Veefkind, I. Aben, K. McMullan, H. Förster, J. De Vries, G. Otter, J. Claas, H. Eskes, J. De Haan, Q. Kleipool, *et al.*, *Tropomi on the esa sentinel-5 precursor: A gmes mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications,* Remote Sensing of Environment **120**, 70 (2012).

[2] P. A. Tavares, N. E. S. Beltrão, U. S. Guimarães, and A. C. Teodoro, *Integration of sentinel-1 and sentinel-2 for classification and lulc mapping in the urban area of belém, eastern brazilian amazon,* Sensors **19**, 1140 (2019).

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

[4] 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).

[5] H. Eskes and K. Boersma, *Averaging kernels for doas total-column satellite retrievals,* Atmospheric Chemistry and Physics **3**, 1285 (2003).

[6] K. Boersma, H. Eskes, R. Dirksen, R. Van Der A, J. Veefkind, P. Stammes, V. Huijnen, Q. Kleipool, M. Sneep, J. Claas, *et al.*, *An improved tropospheric no 2 column retrieval algorithm for the ozone monitoring instrument,* Atmospheric Measurement Techniques **4**, 1905 (2011).

[7] J. Van Geffen, K. F. Boersma, H. Eskes, M. Sneep, M. Ter Linden, M. Zara, and J. P. Veefkind, *S5p tropomi no 2 slant column retrieval: Method, stability, uncertainties and comparisons with omi,* Atmospheric Measurement Techniques **13**, 1315 (2020).

[8] R. van der A, A. de Laat, J. Ding, and H. Eskes, *Connecting the dots: No x emissions along a west siberian natural gas pipeline,* npj Climate and Atmospheric Science **3**, 16 (2020).

[9] A. Pseftogkas, M.-E. Koukouli, A. Segers, A. Manders, J. v. Geffen, D. Balis, C. Meleti, T. Stavrakou, and H. Eskes, *Comparison of s5p/tropomi inferred no2 surface concentrations with in situ measurements over central europe,* Remote Sensing **14**, 4886 (2022).

[10] I. Skoulidou, M.-E. Koukouli, A. Segers, A. Manders, D. Balis, T. Stavrakou, J. van Geffen, and H. Eskes, *Changes in power plant nox emissions over northwest greece using a data assimilation technique,* (2021), https://doi.org/10.3390/atmos12070900.

[11] L. Johansson, A. Karppinen, M. Kurppa, A. Kousa, J. V. Niemi, and J. Kukkonen, *An operational urban air quality model enfuser, based on dispersion modelling and data assimilation,* Environmental Modelling & Software **156**, 105460 (2022).

[12] S. Metia, H. A. Nguyen, and Q. P. Ha, *Iot-enabled wireless sensor networks for air pollution monitoring with extended fractional-order kalman filtering,* Sensors https://doi.org/10.3390/s21165313.

[13] S. Lopez-Restrepo, A. Yarce, N. Pinel, O. L. 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**, 91 (2021).

[14] 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).

[15] F. Karagulian, M. Barbiere, A. Kotsev, L. Spinelle, M. Gerboles, F. Lagler, N. Redon, S. Crunaire, and A. Borowiak, *Review of the performance of low-cost sensors for air quality monitoring,* Atmosphere **10** (2019), cited By :147.

[16] E. Merico, A. Dinoi, and D. Contini, *Development of an integrated modelling-measurement system for near-real-time estimates of harbour activity impact to atmospheric pollution in coastal cities,* Transportation Research Part D: Transport and Environment **73**, 108 (2019).

[17] W.-Y. Yi, K.-S. Leung, and Y. Leung, *A modular plug-and-play sensor system for urban air pollution monitoring: Design, implementation and evaluation,* Sensors **18**, 7 (2017).

[18] G. Marques, N. Miranda, A. Kumar Bhoi, B. Garcia-Zapirain, S. Hamrioui, and I. de la Torre Díez, *Internet of things and enhanced living environments: measuring and mapping air quality using cyber-physical systems and mobile computing technologies,* Sensors **20**, 720 (2020).

[19] J. Wesseling, H. de Ruiter, C. Blokhuis, D. Drukker, E. Weijers, H. Volten, J. Vonk, L. Gast, M. Voogt, P. Zandveld, *et al.*, *Development and implementation of a platform for public information on air quality, sensor measurements, and citizen science,* Atmosphere **10**, 445 (2019).

[20] M. Penza, *Low-cost sensors for outdoor air quality monitoring,* Advanced Nanomaterials for Inexpensive Gas Microsensors , 235 (2020).

[21] K. B. J.H.G.M. van Geffen, H.J. Eskes and J. Veefkind, *Tropomi atbd of the total and tropospheric no2 data products,* Report S5P-KNMI-L2-0005-RP (2021), https://sentinel.esa.int/documents/247904/2476257/sentinel-5p-tropomi-atbd-no2-data-products.

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

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

[24] J. S. Rodríguez, A. Y. Botero, D. Valle, J. G. Serna, and F. Botero, *Experimental approach for the evaluation of the performance of a satellite module in the cansat form factor for in situ monitoring and remote sensing applications,* International Journal of Aerospace Engineering **2021** (2021), https://doi.org/10.1155/2021/8868797.

[25] A. Yarce, J. S. Rodríguez, J. Galvez, A. Gómez, and M. J. García, *Simple-1: Development stage of the data transmission system for a solid propellant mid-power rocket model, Journal of Physics: Conference Series*, **850**, 012019 (2017).

[26] 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 (2020), 10.1007/s11356-020-08268-4.

[27] 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).

[28] X. Liu, R. Jayaratne, P. Thai, T. Kuhn, I. Zing, B. Christensen, R. Lamont, M. Dunbabin, S. Zhu, J. Gao, *et al.*, *Low-cost sensors as an alternative for long-term air quality monitoring,* Environmental research **185**, 109438 (2020).

[29] D. Suriano and M. Penza, *Assessment of the performance of a low-cost air quality monitor in an indoor environment through different calibration models,* Atmosphere **13**, 567 (2022).

[30] A. Ganji, O. Youssefi, J. Xu, K. Mallinen, M. Lloyd, A. Wang, A. Bakhtari, S. Weichenthal, and M. Hatzopoulou, *Design, calibration, and testing of a mobile sensor system for air pollution and built environment data collection: The urban scanner platform,* Environmental Pollution **317**, 120720 (2023).

4

4DEnVar Data assimilation of TROPOMI for parameter estimation

Perhaps some day in the dim future it will be possible to advance the computations faster than the weather advances and at a cost less than the saving to mankind due to the information gained. But that is a dream.

Lewis Fry Richardson

In this chapter, we present the development of a 4D-Ensemble-Variational (4DEnVar) data assimilation technique to estimate top-down NO_x emissions using the regional chemical transport model LOTOS-EUROS with the NO_2 observations from the TROPOspheric Monitoring Instrument (TROPOMI). The assimilation was performed for a domain in the northwest of South America centered on Colombia, also considering regions in Venezuela and Ecuador. In the 4DEnVar approach, the linearized and adjoint model implementation is avoided by generating an ensemble of model simulations and using this ensemble to approximate the nonlinear model and observation operator. Emission correction parameter locations were defined for positions where the model simulations showed significant discrepancies with the satellite observations. The 4DEnVar method was first implemented in the Lorenz96 model to test it in a more controlled scenario. Afterward, optimal emission parameters for

Parts of this chapter have been published in - Yarce Botero, A., Lopez-Restrepo, S., Pinel Peláez, N., Quintero, O. L., Segers, A., Heemink, A. W. (2021). Estimating NOx LOTOS-EUROS CTM Emission Parameters over the Northwest of South America through 4DEnVar TROPOMI NO2 Assimilation. Atmosphere, 12(12), 1633. https://doi.org/10.3390/atmos12121633

the LOTOS-EUROS model have been estimated using the 4DEnVar data assimilation method, allowing for corrections in areas where ground observations are unavailable. The emission inventories of the region do not correctly reflect current emissions activities, so estimating emissions using different approaches helps reduce uncertainties in this field. The analyzed 4DEnVar concentration was compared with the ground measurements of one local air quality network and the OMI satellite instrument. The assimilation had a low impact on surface concentration, primordially enhancing the spatial and temporal variations in the simulated NO_2 fields.

4.1. Introduction

Colombia is demarcated by the Andes mountain range divided into three branches, with peaks up to 5750 m.a.s.l. This terrain is very rough with variable heights covered with rainforest, a suitable territory for generating clouds. The high-density information of NO_2 instruments such as the TROPOspheric Monitoring Instrument TROPOMI (https://www.tropomi.eu/) increases the potential use of NO_2 air quality information over this region. However, the high levels of cloudiness in the area continue to be a significant problem for retrieving pollutants, as discussed in Chapter 3.

A recent review on using CTM's in the tropical Andean region [1] highlighted the difficulties in modeling atmospheric chemistry in tropical regions with abrupt topography. Previous work in Colombia used the WRF-Chem CTM to evaluate the dynamics behind high PM_{10} and $PM_{2.5}$ episodes under different meteorological conditions for the city of Bogotá [2]. Other works include exploring methane emission sources using satellite data [3] and the dispersion of chemicals such as CO within the Aburrá Valley [4]. Barten *et al.* [5] estimated NO_x sources and sinks in Colombia for 2014 at a 1°x1° resolution using WRF-Chem and OMI satellite observations, identifying lighting as the main contributor to the total nitrogen emission budget. Surface NO_2 concentrations in Colombia were estimated from simulations of the global model GEOS-Chem CTM (resolution 2.5°x2°) and OMI NO_2 column measurements for the year 2007, resulting in the identification of biomass burning as a significant source of atmospheric NO_2 [6].
LOTOS-EUROS is an open-source CTM used for a wide range of applications around the world [7, 8], to support scientific research, regulatory programs, and air quality forecasts[9]. LOTOS–EUROS has been included in various model inter-comparison studies and has been tested for the assimilation of ground-based data and satellite observations [10]. Further details about the LOTOS-EUROS model can be found in Chapter 1. This model has been implemented for DA in studying the dynamics of contaminants in the city of Medellín and the Aburrá Valley [11].

Data Assimilation (DA) is the mathematical technique that integrates observations into numerical model simulations, alleviating the model's weaknesses by improving parameter and state representation and estimation [12]. In this chapter, we present the development of a 4D-Ensemble-Variational (4DEnVar) data assimilation technique [13] assimilating satellite data to estimate, through a top-down approach, emission factors for the LOTOS-EUROS CTM. A more detailed explanation about DA in general can be found in Chapter 1.

This chapter presents the development of a 4D-Ensemble-Variational (4DEnVar) data assimilation technique to estimate top-down NO_x emissions using the regional chemical transport model LOTOS-EUROS with the NO_2 observations from the TROPOspheric Monitoring Instrument (TROPOMI). The 4DEnVar emerged in the Numerical Weather Prediction (NWP) institutions to circumvent the large-scale dynamical model's adjoint development. 4DEnVar combines ensemble information with the variational method, avoiding the explicit development of the observation and forward model adjoint operators or its linearization [14].

The incremental 4Dvar adjoint-free data assimilation procedure, known as 4DEnVar, was implemented for the first time using the TROPOMI satellite in conjunction with the LOTOS-EUROS CTM for the Tropical Andes domain. The objective of this implementation was to estimate parameters that correspond to the emission rate of a particular compound, in this case, NO_x. The parameter emission update has substantially enhanced the agreement between the simulated and observed NO_2 fields; this ensures that The data assimilation implementation effectively extracts NO_x emissions from concentration measurements. Although satellite data for this region have not been used due to the abruptness of the topography causing high cloud cover, we can show how a data assimilation technique can exploit the relationships between observed and unobserved states of a chemical transport model to improve its representation. Readers should be careful in interpreting the emissions presented here.

This chapter is organized as follows. The following section 4.2 presents the 4DEnVar mathematical formulation for the experiments. The emission uncertainties perturbation model that will drive an ensemble of CTM simulations is explained. At the end of this section, an overview of the data used for validation is presented. Subsequently, the results of the 4DEnVar assimilation experiments are described in Section 4.3, focusing on estimating the optimal emission factors and the impact on the model simulations. The final section 4.4 summarizes the results and discusses the usability of this kind of data assimilation technique for future applications.

4.2. Methods

4.2.1. Lorenz 96 model

Small-scale models with chaotic properties and nonlinear behavior are useful to test data assimilation techniques, such as the Lorenz 96 model. Unlike its predecessor, the Lorenz 63, the 96 version is useful for generating a custom number of states from its recursive mathematical formulation. It is defined for $i = 1, ..., n$ where n is the state number with the following expression:

$$\frac{dx_i}{dt} = (\mathbf{x}_{i+1} - \mathbf{x}_{i-2})\mathbf{x}_{i-1} - \mathbf{x}_i \tag{4.1}$$

The index i is cyclic, which means that $\mathbf{x}_{-1} = \mathbf{x}_{n-1}$,$\mathbf{x}_0 = \mathbf{x}_n$ and $\mathbf{x}_{n+1} = \mathbf{x}_1$. $\mathbf{x}_i$ is the state system. The model is highly non-linear and has a strong state-to-state coupling

4.2.2. Chemical Transport Models

Chemical Transport Models (CTMs) are used to simulate and forecast air quality to understand the contaminant dynamics in the atmosphere. CTMs are multivariate models incorporating hundreds of gaseous species and aerosols and their related reactions [12]. Persisting uncertainties associated with emission parameters and numerical approximations of certain model dynamics prevent the exact simulation of reality from CTMs [15–17]. Accuracy in emission inventories is fundamental for proper simulations, yet inventories are generally outdated and principally generated

through bottom-up approaches [18–20].

Data assimilation methods for CTMs are inspired mainly by meteorological DA experiences [12]. Many successful applications have demonstrated the benefits of DA for CTMs, either to produce re-analysis fields and forecasts or with the focus on improving the accuracy of model inputs (such as initial conditions, boundary conditions, emissions) [21]. A common characteristic of these applications is that in regional air-quality simulations, the influence of initial conditions quickly fades over time, as emissions and lateral boundary conditions primarily determine the model fields [22].

Techniques in DA follow either variational or sequential approaches. The variational approach optimizes a cost function that calculates the mismatches between the model and the observations. In contrast, the sequential approach is progressively updated, reconciling the state using the uncertainties defined for the simulated state and the observations [23–25]. A considerable drawback of variational DA is the requirement for a direct model adjoint representation, prohibitive for large-scale models ($\sim 10^6 - 10^9$ state elements), and very expensive to maintain [12]. Hybrid approaches tackle this problem, aiming to take advantage of the characteristics of each method [24, 26]. Ensemble-based approaches avoid the construction of tangent linear and adjoint representations of the forecast model (e.g., [27]), which is interesting for low-budget operational scenarios using numerical models. Emili *et al.* give details on the 4DEnVar methodology and its application to highly nonlinear reactive species in chemical transport data assimilation and state that the emissions are often the most uncertain but also most influential parameters.

Data assimilation experiments with the LOTOS-EUROS CTM mainly used an Ensemble Kalman Filter (EnKF) approach, with most of the applications over Europe [28–30]. In Colombia, the LOTOS-EUROS CTM has been used since 2017 [31], including studies on the assimilation of surface network observations using an ensemble-based LEnKF technique for particulate matter forecasts [11, 32]. Recent applications use variational approaches with the LOTOS-EUROS model in estimating volcanic ash emissions [33**?**] where the vertical profile of volcanic ash injections was estimated using variational techniques that did not require the implementation of the adjoint model (Traj-4DVar). Jin *et al.* [34] used the LOTOS-EUROS model over China to assimilate satellite data to improve the simulation of dust transport from the Gobi Desert to Chinese cities, proposing an adjoint-free 4DVar technique for dust emission parameters estimation.

In this chapter, we used the LOTOS-EUROS CTM to simulate NO_2 concentrations over North-South America, modifying the simulations through 4DEnVar satellite data assimilation for emission parameter estimation. The emission parameter is the quantity responsible for modifying the emission inventories over a region. The increasing availability of observations from both satellites and ground-based instruments allowed reducing the uncertainty of atmospheric chemistry models in many applications [27]. The reason to incorporate these measurements is to improve the numerical models and induce parameter estimation capabilities to overcome the complex tasks of emission inventory development. The principal advantage of the 4DEnVar technique is that it does not require an adjoint model and could also

allow for the introduction of generic model errors by adding stochastic perturbations during the propagation of the ensemble [27]. The variational technique helps assimilate the many observations produced from each satellite overpass, iteratively searching for the optimal values of the quantity of interest.

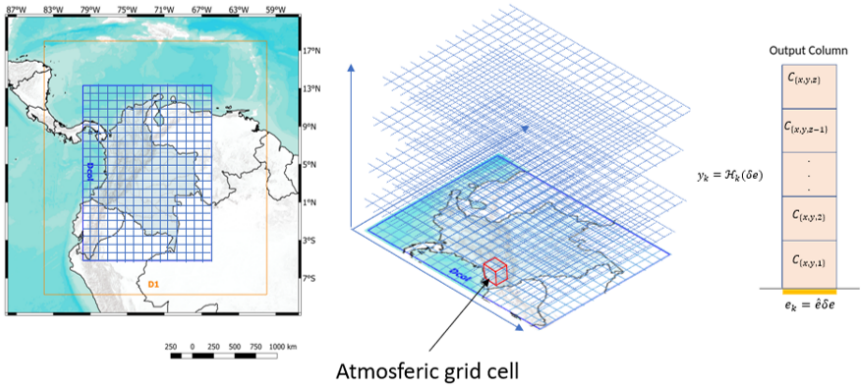


Figure 4.1: The left is a Conceptual illustration of simulation domains and vertical extent and the center and Right, is a conceptual state representation in the gridded model of the column being sampled by the observation operator $\mathcal{H}_k(\delta e)$ relatedto the emission factor δe.

To validate the results, two sources of information were used. First, the version 4.0 Aura Ozone Monitoring Instrument (OMI) Nitrogen Dioxide (NO_2) Standard Product (OMNO2) was used to calculate the tropospheric NO_2 vertical column density (VCD) time-series, which can be found at (https://disc.gsfc.nasa.gov/datasets/OMNO2). Secondly, SIATA (Sistema de Alerta Temprana del Valle de Aburr'a), a network of sensors that offer air-quality measurements for different pollutants in the atmosphere across the Aburr'a Valley region such as O3, SO2, PM10, PM2.5, and PM1. The network is spread across the five most populous municipalities in the Aburr'a Valley, with most of the measurement stations in Medellín inside them into the perimeter demarcated by the mountains of the Valley. SIATA has ground-based sensor measurements of NO2, with stations along the Valley from SIATA's data portal, available at (https://siata.gov.co/descargasiata/index.php/index2/).

4.2.3. 4DEnVar Data assimillation

The 4DEnVar DA technique solves several variational analyses in parallel, introducing flow dependence in the 4DVar DA technique [13, 35]. Multiple trajectories generated through the emission factors' initial parameters' perturbations spawn an ensemble space for the analysis update to estimate the initial parameters gradually through an incremental perspective and a flow-dependent background covariance.

A work that points out details about the viability of a 4DEnVar method dealing with highly nonlinear reactive species in a chemical transport data assimilation could be found in [27]. This work details the viability of a 4DEnVar method dealing with highly nonlinear reactive species in a chemical transport data assimilation.

Let $\mathbf{x}_t \in \mathbb{R}^{n\times 1}$ and $\mathbf{x}^b$ represent the states and parameters of the LOTOS-EUROS model, and the background state at time t respectively. The state $\mathbf{x}_t$ results from propagating the previous state $\mathbf{x}_{t-1}$ one time step using the non-linear model operator $\mathcal{M}_{t-1,t}$:

$$\mathbf{x}_t = \mathcal{M}_{t-1,t}(\mathbf{x}_{t-1}), \quad \mathbf{x}_0 \sim \mathcal{N}(\mathbf{x}_0^b, \mathbf{B}), \tag{4.2}$$

Let $\mathbf{y}_t \in \mathbb{R}^{m\times 1}$ be the set of m number of observations made at time t. $\mathbf{y}_t$ is related to $\mathbf{x}_t$ via the observation operator $\mathbf{H}_t$:

$$\mathbf{y}_t = \mathbf{H}_t\mathbf{x}_t + \nu_t, \qquad \nu_t \sim \mathcal{N}(0, \mathbf{R}_t), \tag{4.3}$$

The mathematical method 4DEnVar was introduced by [13] and its initial formulation comes from the 4DVar method presented in Chapter 1. The cost function described in Eq. 1.4 which is a functional of the n parameters in the augmented quantity $\mathbf{x}_0$ over a time window, is defined from $t = 0$ to $t = T$:

$$\mathcal{J}(\mathbf{x}_o) = \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}_0^b\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{t=0}^{T}\|\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t\|^2_{\mathbf{R}_t^{-1}}, \tag{4.4}$$

where $\mathbf{B}$ is the $n \times n$ background-error covariance matrix and $\mathbf{R}_t$ corresponds to the $m \times m$ observation-error covariance matrix. In an incremental formulation for an updated state from the background state, a state $\mathbf{x}_t$ is estimated as follows:

$$\mathbf{x}_t = \mathbf{x}_t^b + \delta\mathbf{x}_t, \tag{4.5}$$

with $\delta\mathbf{x}_t$ being the increment in the state $\mathbf{x}_t$. Substituting Eq. 4.5 into Eq. 4.4 leads to:

$$\begin{aligned}\mathcal{J}(\delta\mathbf{x}_0) &= \tfrac{1}{2}\|\mathbf{x}_0^b + \delta\mathbf{x}_0 - \mathbf{x}_0^b\|^2_{\mathbf{B}^{-1}} + \tfrac{1}{2}\textstyle\sum_{t=0}^{T}\|\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t^b + \delta\mathbf{x}_t\|^2_{\mathbf{R}_t^{-1}} \\ &= \tfrac{1}{2}\|\delta\mathbf{x}_0\|^2_{\mathbf{B}^{-1}} + \tfrac{1}{2}\textstyle\sum_{t=0}^{T}\|\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t^b - \mathbf{H}_t\delta\mathbf{x}_t\|^2_{\mathbf{R}_t^{-1}},\end{aligned} \tag{4.6}$$

Equation 4.6 represents the incremental strong-constraint 4DVar cost function. Minimizing $\mathcal{J}$ gives particular values of $\delta\mathbf{x}_0$. The cost function in (4.6) could be written in terms of the initial increment $\delta\mathbf{x}_0$ like

$$\mathcal{J}(\delta\mathbf{x}_0) = \frac{1}{2}\|\delta\mathbf{x}_0\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{t=0}^{T}\|\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t^b + \mathbf{H}\mathbf{M}_{(0,t)}\delta\mathbf{x}_0\|^2_{\mathbf{R}_t^{-1}} \tag{4.7}$$

where $\mathbf{M}_{(0,t)}$ is the tangent linear model. The increment is now transformed by what is known as a control variable $\delta\mathcal{X}_{var}$ from $\delta\mathbf{x}_0 = \mathbf{U}\delta\mathcal{X}_{var}$ which is associated with the model space variable $\delta\mathbf{x}_0$ and $\mathbf{U}$ the Control Variable Transforms (CVTs), which is a workaround to represent the covariance matrices needed without knowing them explicitly[14]. Substituting this transformation in Eq. 4.6 generates the following preconditioned cost function:

$$\mathcal{J}(\delta\mathcal{X}_{var}) = \frac{1}{2}\|\mathbf{U}\delta\mathcal{X}_{var}\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{t=0}^{T}\|\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t^b - \mathbf{H}\mathbf{M}_{(0,t)}\mathbf{U}\delta\mathcal{X}_{var}\|^2_{\mathbf{R}_t^{-1}}, \tag{4.8}$$

The next step consist of replacing the CVTs in Eq. 4.8 and use $\mathbf{U} = \mathbf{B}^{1/2}$, having the covariance matrix of the first term like $\mathbf{B} = \mathbf{U}\mathbf{U}^T$:

$$\mathcal{J}(\delta\mathcal{X}_{var}) = \frac{1}{2}\|\mathbf{U}\delta\mathcal{X}_{var}\|^2_{(\mathbf{U}\mathbf{U}^T)^{-1}} + \frac{1}{2}\sum_{t=0}^{T}\|\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t^b - \mathbf{H}\mathbf{M}_{(0,t)}\mathbf{U}\delta\mathcal{X}_{var})\|^2_{\mathbf{R}_t^{-1}} \quad (4.9)$$

If we develop the first term in Eq. 4.9 we have:

$$\begin{aligned}\frac{1}{2}\|\mathbf{U}\delta\mathcal{X}_{var}\|^2_{(\mathbf{U}\mathbf{U}^T)^{-1}} &= \frac{1}{2}(\mathbf{U}\delta\mathcal{X}_{var})^T(\mathbf{U}\mathbf{U}^T)^{-1}(\mathbf{U}\delta\mathcal{X}_{var}) \\ &= \frac{1}{2}(\delta\mathcal{X}_{var})^T\mathbf{I}(\delta\mathcal{X}_{var}) \\ &= \frac{1}{2}\|\delta\mathcal{X}_{var}\|^2_{\mathbf{I}},\end{aligned} \quad (4.10)$$

We now create an ensemble of N realizations of $\delta\mathbf{x}_0$ from a $N(0,\mathbf{B})$ distribution and approximate $\mathbf{B}$ by the covariance of this ensemble with $\mathbf{X}^b \in \mathbb{R}^{n\times N}$ equal to :

$$\mathbf{B} \approx \frac{1}{\sqrt{N-1}}\sum_{i=0}^{N}\delta\mathbf{x}_0^i\mathbf{I}^T\delta\mathbf{x}_0^i = \mathbf{X}^{b^T}\mathbf{X}^b \ . \quad (4.11)$$

Using $\mathbf{U} \approx \mathbf{X}^b$ in the cost function, Eq. 4.9, we have:

$$\begin{aligned}\mathcal{J}(\delta\mathcal{X}_{var}) &= \tfrac{1}{2}\|\mathbf{X}^b\delta\mathcal{X}_{var}\|^2_{(\mathbf{X}^b(\mathbf{X}^b)^T)^{-1}} + \tfrac{1}{2}\textstyle\sum_{t=0}^{T}\|\mathbf{y}_t - \mathbf{H_t}\mathbf{x}_t^b - \mathbf{H}_t\mathbf{M}_{(0,t)}\mathbf{X}^b\delta\mathcal{X}_{var}\|^2_{\mathbf{R_t}^{-1}} \\ &= \tfrac{1}{2}\|\delta\mathcal{X}_{var}\|^2_{\mathbf{I}} + \tfrac{1}{2}\textstyle\sum_{t=0}^{T}\|\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t^b - \mathbf{Y}_t\delta\mathcal{X}_{var}\|^2_{\mathbf{R}_t^{-1}}\end{aligned} \quad (4.12)$$

The core assumption, the need for the tangent linear model, is avoided. To minimize the cost function, the gradient is formulated using an ensemble of model simulations as follows in [14]:

$$\mathbf{Y}_t = \mathbf{H}_t\mathbf{M}_{(0,t)}\mathbf{X}^b \approx \frac{1}{\sqrt{N-1}}\left(\mathbf{H}_t(\mathcal{M}_{0,t}(\delta\mathbf{x}_0^i)) - \overline{\mathbf{y}_t}, ..., \mathbf{H}_t(\mathcal{M}_{0,t}(\delta\mathbf{x}_0^N)) - \overline{\mathbf{y}_t}\right) \ , \quad (4.13)$$

where $\overline{\mathbf{y}_t}$ is the model observation vector at time t based on the ensemble mean. The gradient of eq. 4.12 is calculated as follows:

$$\nabla\mathcal{J}(\delta\mathcal{X}_{var}) = \delta\mathcal{X}_{var} + \sum_{t=0}^{T}\mathbf{Y}_t^T\mathbf{R}^{-1}(\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t^b - \mathbf{Y}_t\delta\mathcal{X}_{var}) = 0 \quad (4.14)$$

Solving Eq. 4.14 for $\delta\mathcal{X}_{var}$ leads to:

$$\delta\mathcal{X}_{var} = -\sum_{t=0}^{T}\mathbf{Y}_t^T\mathbf{R}^{-1}(\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t^b)\cdot\left(\mathbf{I} - \sum_{t=0}^{T}\mathbf{Y}_t^T\mathbf{R}^{-1}\mathbf{Y}_t\right)^{-1} \ . \quad (4.15)$$

This algorithm is for cases where the number of parameters is not too large. By setting the cost function gradient equal to 0, the system of equations for $\delta \boldsymbol{\mathcal{X}}_{var}$ can be obtained. This procedure can also be repeated, starting with the latest estimates of the parameters, by generating a new ensemble of model simulations by perturbing these parameters and by computing a new linear approximation of the observation operator. For large-scale problems, this implementation is not attractive from the computational point of view, and a gradient-based minimization algorithm based on equation 4.12 is to be preferred.

Once $\delta \boldsymbol{\mathcal{X}}_{var}$ is estimated, the following expression calculates the increment of the state:

$$\mathbf{x}_0 = \mathbf{x}_0^b + \delta \mathbf{x}_0 = \mathbf{x}_0^b + \mathbf{X}^b \delta \boldsymbol{\mathcal{X}}_{var} \ . \tag{4.16}$$

The 4DEnVar methodology is used for estimating uncertain parameters. In our application, the emissions are the major sources of uncertainty and are parameterized by introducing multiplicative correction factors.

Let $\delta \mathbf{e} \in \mathbb{R}^{n \times 1}$ with n the number of uncertain parameters, represent the unknown LOTOS-EUROS emission correction factors. The emissions $\mathbf{e}_t$ at time t are calculated according to:

$$\mathbf{e}_t \ = \ \hat{\mathbf{e}}_t \ (\ 1 \ + \ \delta \mathbf{e} \ ) \tag{4.17}$$

with $\hat{\mathbf{e}}_t$ as the nominal emissions at time t from the emission inventory.

The choice of the emission correction parameters to be estimated and the quantification of their uncertainty is critical in designing the assimilation system because this should reflect the most significant uncertainty of the CTM model. Since chemical species can be sensitive to physical and chemical processes, the primary uncertainty source could differ depending on the species of interest.

The choice of the control variables should reflect the largest source of uncertainty of the considered model, which is commonly associated with the model's initial state or chemical emissions. Therefore, its uncertainty is critical for designing an appropriate assimilation algorithm that ensures the correct DA results [27]. Since different chemical species can be sensitive to different physical and chemical processes, the primary source of uncertainty can differ from species to species. [27]. Therefore, the initial conditions are chemical DA's most important control variables. The emissions are often the most influential input parameters choice as control variables [12], and in this chapter implementation, the emission factors were also taken as the parameters to update.

The LOTOS-EUROS simulations exhibit different time evolution of the states of the system in time when compared to real-life observations. This discrepancy is mostly attributed to the uncertainty in the model emission parameters [32, 34]. As previously discussed, the available inventories are not accurate enough in the region of interest. Accurate emissions for specific components such as $PM_{2.5}$ were estimated only recently, for example, for cities like Medellín using an EnKF DA technique in [32]. Nevertheless, for the coarse domain inventories of other pollutants, the emission inventories remain uncertain. Different perspectives have been taken into account for the emission parameter estimation problem using satellite information and data assimilation variational (4DVar) techniques ([36–40], and also from the sequential (EnKF, OI) techniques [41, 42].

4.2.4. LOTOS-EUROS model 4DEnVar setup

The basic idea of the 4DEnVar method is to estimate the sensitivities of the observations concerning changes in the parameters using the full observation model instead of using the linear tangent model and its adjoint that is needed for the gradient calculation. This is done by generating an ensemble of forwarding model simulations that are used to obtain a linear approximation of the operator $\mathcal{H}_t$ [12, 43].

In this chapter, the 4DEnVar methodology is used to estimate uncertain parameters. This chapter shows the viability of this method in dealing with a small number of highly nonlinear reactive species such as NO_2 in a control scenario.

Using all the available observations, we want to estimate the emission correction factors δe. First, the selection of places to estimate those parameters was made, choosing 27 locations that are the main cities and other locations of interest, such as open pit mines and refineries. After the successive implementation of one of the techniques just described, the parameters δe can finally be calculated. The input to this procedure is the CTM model, the satellite observations, and the parametrization of the algorithm (window length, inner loops, convergence criteria), and the output is the set of optimized parameters.

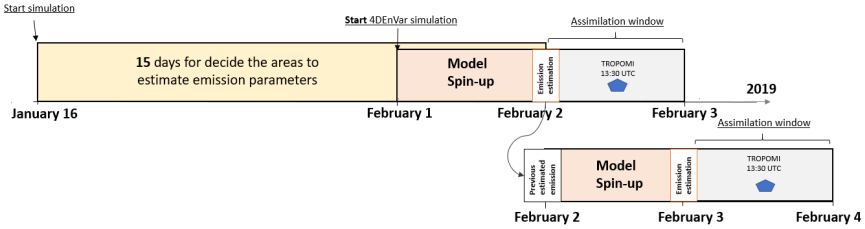


Figure 4.2: Data assimilation (4DEnVar) procedure for emission parameter estimation. Simulations were run for 15 days (January 16-31) before the data assimilation process started. In the window February 1-3, the ensemble was propagated without assimilation, and in this new ensemble subspace, the assimilation was performed for February 2-3 and 3-4.

Figure 4.2 shows the setup of the 4DEnVar assimilation process proposed for the first days of February 2019. A simulation of 15 days over January 16 to February 1 was performed before data assimilation to identify the locations where emission parameters require improvement. Data assimilations were performed over individual days (February 2 and 3). Each assimilation was preceded by a 1-day spin-up period, where the simulated fields stabilized and became independent of the initial conditions. During assimilation, the observations (satellite data available from the TROPOMI instrument at 13:30 local time) were incorporated to estimate the value of the initial emission factor parameters. The optimization procedure consisted of an outer/inner loop iteration to reach the convergence criteria or the maximum number of iterations defined to find the proper parameters. The iteration procedure was terminated when the maximum iteration number was reached or met by the minimal error criteria.

To perform the assimilation window, the 4DEnVar weighted an ensemble of 40 model trajectories generated from perturbing the emission factor parameters based

on how much the ensemble matched the observations and the background state. Based on its minimization procedure, a limit of 15 iterations for the inner loop of the 4DEnVar implementation step was set. Incorporating more than 15 iterations demonstrated no significant reduction of the cost function. The 4DEnVar result (known as analysis in many data assimilation scenarios) is the model value for the emission parameters updated in the previously chosen positions.

4.3. Results

This section is separated into three subsections. First, we focus on implementing the 4DEnVar in the Lorenz96 mode. Then, we pay attention to the observations used for assimilating, and last, we show the DA over the whole of Colombia and major cities.

4.3.1. 4DEnVar in Lorenz96

Figure 4.3 illustrates the temporal evolution of the 40 states; these plots are known as Hovmoller diagrams and are used to see all the states simultaneously for time propagation. The background is the first model realization that is taken as the states to update. The truth value represents a sampled model realization that for generating synthetic observations as the target in the twin experiment to test the ability of the DA system. The analysis shows the filter outcome once the synthetic observations are ingested. Additionally, the figure displays the analysis outcome obtained from the data assimilation (DA) procedure and a comparative visualization of the difference between the analysis and the truth in the right image.

To have another perspective, let's see what happens for three randomly chosen states in Figure 4.4. The dashed black line represents the background state, the red line represents the truth, and the analysis step is in blue. The green dashed line represents the start and end of the assimilation window, which is the defined time for which observations are accepted in the assimilation cycle. Notice here how the initial value at the beginning of the assimilation window is updated to produce a forward simulation step from this new point that predicts more accurately than the background value that the truth suggests.

For the experiment, a number of ensembles from 10 to 100 were defined, increasing to 10 ensemble members in each experiment (Figure 4.5). One ensemble

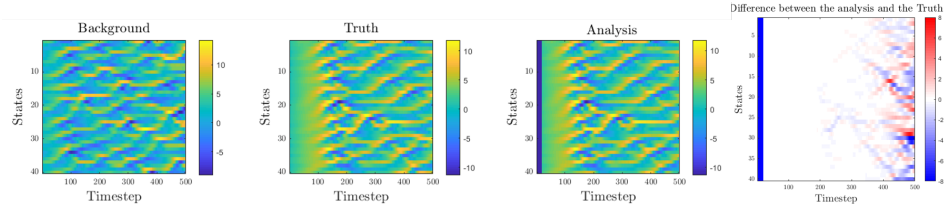


Figure 4.3: The left panel shows the representation in time of the background, then immediately to the right the truth, and then the analysis. On the right is the difference between the analysis output and the background.

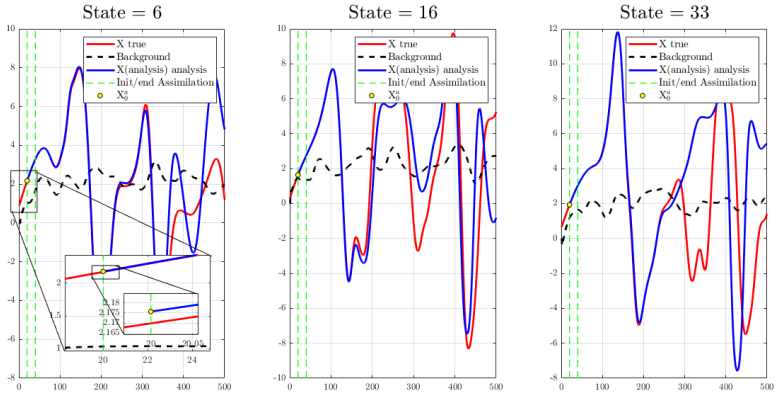


Figure 4.4: Comparison background and analysis in three validation random states (6,16, and 33). The truth state is depicted in red, the background initial state corresponds to the black dashed line, and the analysis output state is in blue. The 4DEnVar method, through the minimization of the cost function, estimates the initial condition for propagating the update version of the model.

is defined as one realization of the model for certain initial conditions.

Once the assimilation step was applied to the model, it was easy to find that by increasing the number of ensembles, the analysis representation improved, approaching the truth value. For the case of the size of the ensemble spread (Figure 4.6), which is also known as inflation, the behavior was the opposite because once the spread size increases, the representation of the analysis deviates more from the truth value.

The case for window size increasing is shown in Figure 4.7. When the window size started to increase, the capability of the method to represent the truth from the analysis worsened.

4.3.2. NO_2 column concentrations: TROPOMI and LOTOS-EUROS

A qualitative comparison of the simulated and observed NO_2 column densities was conducted to detect locations where the TROPOMI observations differed significantly from the LOTOS-EUROS simulations to focus the attention therein for the data assimilation and the emission estimation of the experiment. Figure 4.8 shows a comparison between the 15-day average satellite observations (January 16-February 2019) with the LOTOS-EUROS NO_2 column, as well as the EDGAR NO_x emission inventory on the right image.

The NO_2 concentration in the model and the observations presented high values over the main urban centers, but the amplitude of the simulated NO_2 concentrations often differed. Simulated values over the densely populated area of Venezuela that includes Caracas, Valencia, Barquisimeto, and Maracaibo (a region known as the *oil refinery corridor*) showed much higher values in the model simulation output than the satellite observations, and consequently, showed the highest root-mean-square-error (RMSE; lower-left in Figure 4.8, details performance metrics in the

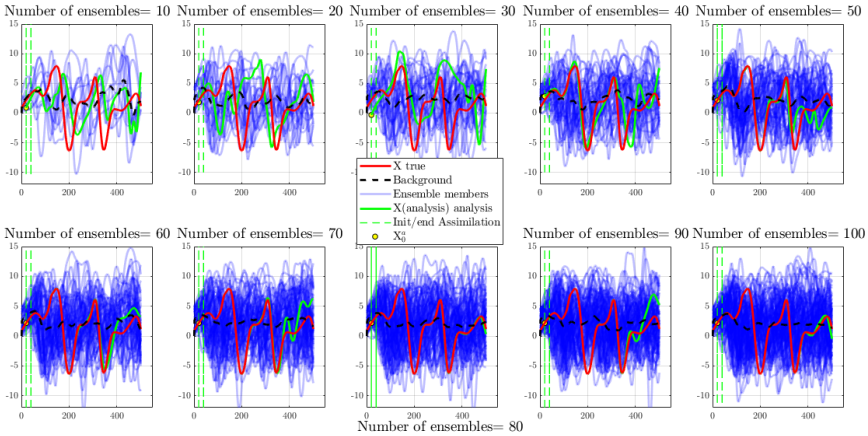


Figure 4.5: Assimilation comparison increasing the ensemble size as indicated. The x-axes correspond to the time dimension. The analysis update from the filter corresponds to the green line, the true is depicted in red, and the blue lines are the ensemble members.

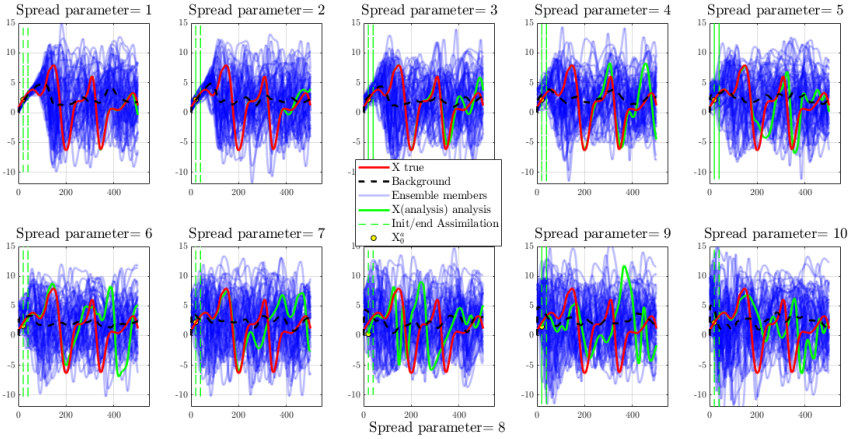


Figure 4.6: Assimilation comparison between the analysis and the true for different spread sizes (1 to 10), the spread corresponds to what is also known as inflation. In the initial condition can be seen as the spread in the seed values for the ensemble members. The analysis update from the filter corresponds to the green line.

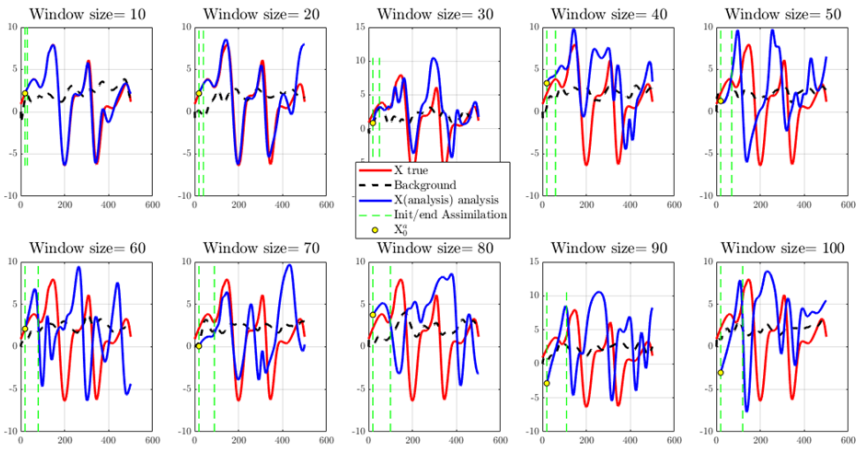


Figure 4.7: Assimilation comparison for different window assimilation sizes. The window size is denoted with the vertical green dashed lines and corresponds to the time for which observations are ingested for the assimilation stage. After a window size of 30, the 4DEnVar analysis no longer starts to fit the truth.

Appendix of this Chapter). A Mean Fractional-bias (MFB; lower-centre in Figure 4.8) calculation showed a broader perspective over the domain, revealing that the model underestimated the NO_2 in the south except for some isolated, overestimated points over the southern Colombian Amazonia. The model also underestimated NO_2 concentrations along the Magdalena River Valley in northern Colombia.

Twenty-seven locations were selected to perturb the model's parameters to generate a subspace of model ensemble trajectories to estimate appropriate values driven by the observations. These are marked as black squares on the central panel of Figure 4.9 denoting the areas selected over the initial nominal emission parameter value. These correspond to the major cities and other areas of interest denoted in the previous Figure 4.8. One is the main oil refinery in Colombia located in the central part of the country in the Magdalena River Valley, and the other is open-pit coal-mine mining in northern Colombia near the Venezuelan border that presented anomalous concentration values in a rural area. The emissions for these 27 locations were updated using the 4DEnVar data assimilation.

Each perturbation location consisted of a 3x3 buffer of 0.09°x 0.09°grid cells. The emission factor perturbation value multiplies the concentrations homogeneously in the area where the parameter was estimated. This buffer of 9 grids is because TROPOMI samples the downwind plume, which might be 1-2 grid cells away from the source.

4.3.3. 4DEnVar data assimilation results over Colombia

Figure 4.10 shows the NO_2 column concentrations from the LOTOS-EUROS before assimilation (free run) and the TROPOMI retrieved product for February 2, 2019. Although the satellite data presented an absence of data (white areas) due to the

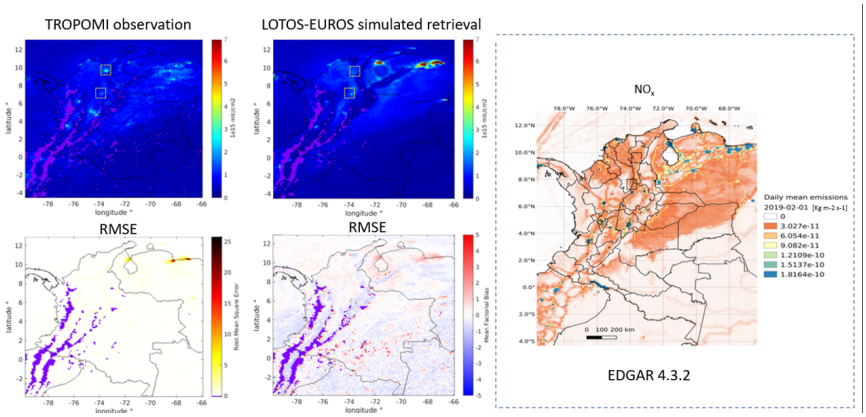


Figure 4.8: In the top left and center, the 15 days mean (January 16 to February 1, 2019) of the TROPOMI tropospheric NO_2 vertical column and the corresponding LOTOS-EUROS simulations. Spatial resolution 0.09 °x 0.09 °. Purple areas correspond to areas where no data was available, mainly due to the cloudiness presence over the area for all that time. The dashed yellow squares correspond to two locations selected to update emissions that are not principal cities, but many concentration spots are seen from the remote sensing instrument to qualitatively detect the right place to estimate unknown or uncertain parameters. In the bottom left and centre, the root-mean-square-error and modified fractional bias between retrieval and simulation. On the right is EDGAR V4.3.2 anthropogenic NO_x emissions, 2012.

quality filter, it was possible to notice the difference between the model simulation and the observation from the remote instrument. The dashed rectangle marks a relatively flat area in Venezuela where good quality observations, due to low cloud cover, are present and where the model overestimates the satellite product denoting areas to update the emission parameters. Other cities like Bogotá, Medellín, Barranquilla, and Quito also present differences between LOTOS-EUROS and satellite information. Open-pit mining pops up in satellite observation and in a small city in the Maracaibo Gulf.

The left panel of Figure 4.11 shows the emission factors obtained with the 4DEnVar technique once it is completed, showing the estimated values of the 27 locations of emission parameters selected based on the preliminary simulation in the domain. For this case, the emission factors went from 0 to 2, suggesting in much of this area reduction or augmenting the magnitude of the parameters due to the overestimation of the model against the satellite observation through the minimization of the cost functional from equation 4.4. In the north part of the domain, mainly in the Venezuelan dashed area, the values suggest reducing emissions below 1, which is the nominal value, it means with this emission factor, the emissions are not modified. In some areas in the center of Colombia, the most populated areas, the new emission suggests an increase with values slightly above 1.

The 4DEnVar analysis output is shown in the center mosaic in Figure 4.11. This graphic shows the NO_2 column that the LOTOS-EUROS model simulated using the updated emission parameters. The image on the right shows the MFB between this analysis step and the background condition; here, it is possible to appreciate the spatial over/underestimations between both scenarios resulting from the updated

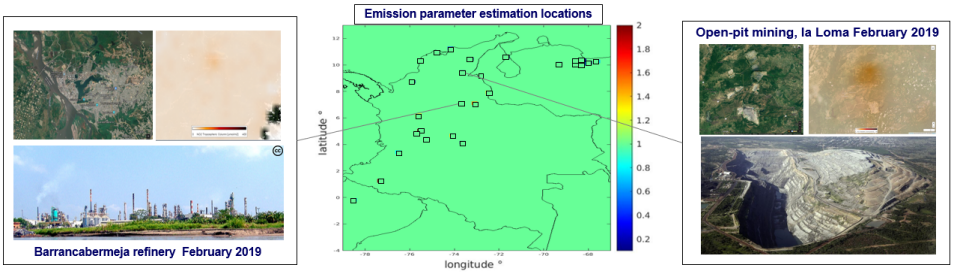


Figure 4.9: Central panel: Locations selected for emission parameter adjustment via 4DEnVar data assimilation. The 27 locations indicated in the *center* panel, corresponding primarily to large urban centers, were chosen for parameter adjustment based on the magnitude of the differences observed in Figure 4.8. The colorbar indicates the emission parameter value, in this case all the emission factor locations are in the default condition which is the nominal with value 1. *Left panel*: location of Colombia's largest oil refinery (near the city of Barrancabermeja). *Right panel*: Drummond open-pit coal mine, a non-urban site whose emissions were poorly represented by the default emissions inventory (Photo Diego Santagruz/Archive journal EL TIEMPO). NO_2 concentration images from https://maps.s5p-pal.com/.

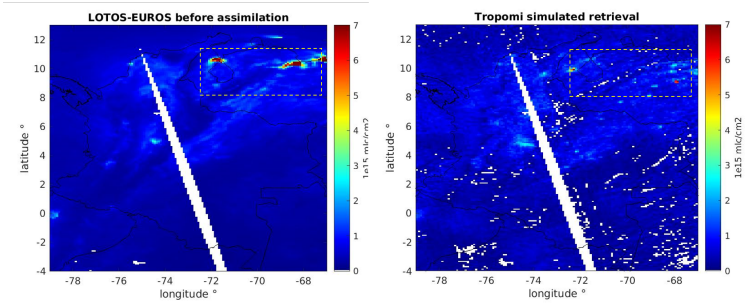


Figure 4.10: Comparison of the tropospheric NO_2 column estimation from LOTOS-EUROS simulations (*left panel*) and TROPOMI retrievals (*right panel*) for February 2, 2019. The dashed rectangle highlights the sizable discrepancies between the two sources observed over the main Venezuelan cities (Caracas, Valencia, Barquisimeto, and Maracaibo). Colorbar indicates the column concentrations.

initial parameter values. The area's values in the dashed square region in the previous Figure suggest a reduction of (0.2-0.6). With the new emission factors, the emissions within the rectangle were strongly reduced, leading to simulated lower magnitude NO_2 columns that were better in agreement with the TROPOMI simulations. Curaçao shows a plume with a high east-west trend in the LOTOS-EUROS that is not appreciated in the TROPOMI data for this day and where we did not put an update emission. It is possible to see how the same plume appears in the analysis simulation, indicating that the model was not modified for this location.

The parameters' value and the comparison of results are presented with the MFB in the right panel, using the emissions estimated to propagate a new forward run (4DEnVar analysis simulated retrieval) and the free run. In the area of Venezuela, it is possible to see how the estimated values of the parameter suggest a reduction of emissions. Moreover, in the city of Bogotá, an underestimation of the model's

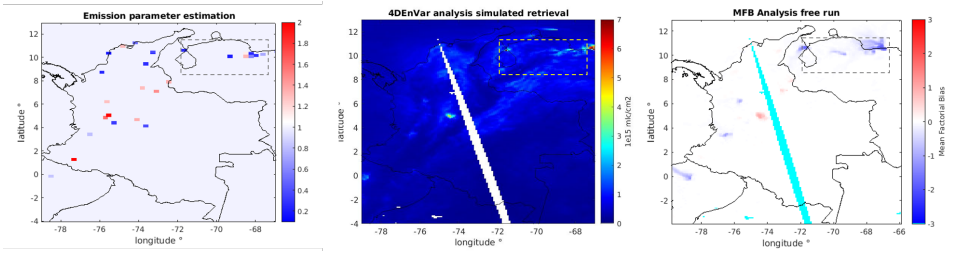


Figure 4.11: *Left panel*: Colorbars represent the NO_2 emission factor adjustment (δe) over the 27 locations illustrated in Figure 4.9 to reconcile the LOTOS-EUROS simulations with the TROPOMI observations, with values < 1 indicating LOTOS-EUROS over-estimations, and values > 1 indicating under-estimations (observations > simulated values). *Central panel*: simulated TROPOMI NO_2 columns from the analysis run using the newly estimated emission factors. Colorbars represent concentrations *Right panel*: Colorbars represent the Mean Fractional Bias in NO_2 column between LOTOS-EUROS simulations with the adjusted emission factors (analysis) and the simulations with the default emission factors (background).

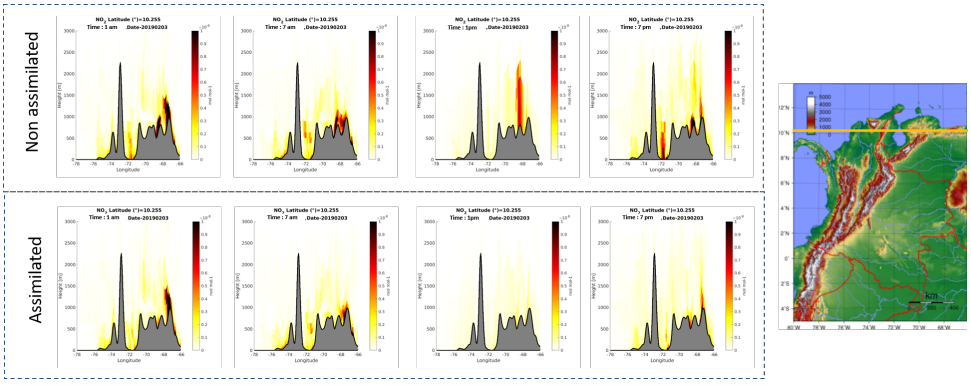


Figure 4.12: Transverse cut profiles for NO_2 concentration over latitude =10.255° for four different time points during the assimilation window day (2019-02-03). The adjustment in emission parameters affected the simulated NO_2 column estimates at various heights and the associated plume. Colorbars indicate concentration levels in mixing ratio. The right panel shows the topography map of Colombia with the transversal cut location

emission was corrected, increasing the emission due to the parameter estimated in that place. The locations where the emission adjustment does not help, presumably, are where the satellite information is incomplete, such as the Andean mountain corridor and regions for which the emission is too low in the current inventory.

Vertical profiles

Figure 4.12 shows a comparison between the assimilated and non-assimilated vertical NO_2 concentrations for the latitude 10.255° for four different local time steps (1 am, 7 am, 1 pm, 7 pm UTC) during the second day of assimilation (2019-02-03). This mosaic compares the changes in the vertical profiles of this gas. This latitude corresponds to the locations of the parameters with the highest emissions in Venezuela, which correspond to the oil refineries. The non-assimilated scenario

displays a higher NO_2 concentration than the assimilated scenario; it is evident how the plume partially disappears for moments like at 1 pm. The emission was updated with the parameters estimated, suggesting this is a decrease in the emissions for the current time for the experiment. These results could be explained by the fact that during the last ten years, there has been a dramatic reduction in oil production in Venezuela due to the U.S sanctions, for which also the refining activity has decreased [44].

4.3.4. Impact over major cities

The 4DEnVar results have been evaluated in more detail for the areas around the cities of Medellín and Bogotá (Figure 4.13). For these cities, respective close-ups are presented, for which it is also possible to appreciate the wind pattern driving the transport of these atmospheric chemicals, which is mainly east-west and sometimes northwest. The TROPOMI retrieval is shown in the two right columns, while the emission factors estimated are shown in the two left columns. For the two days of assimilation, the change in the value of this new emission parameter from the nominal value has an impact on the concentration fields as shown in Figure 4.14. The second assimilation day integrates the emissions being estimated as the new nominal values to start the new assimilation window.

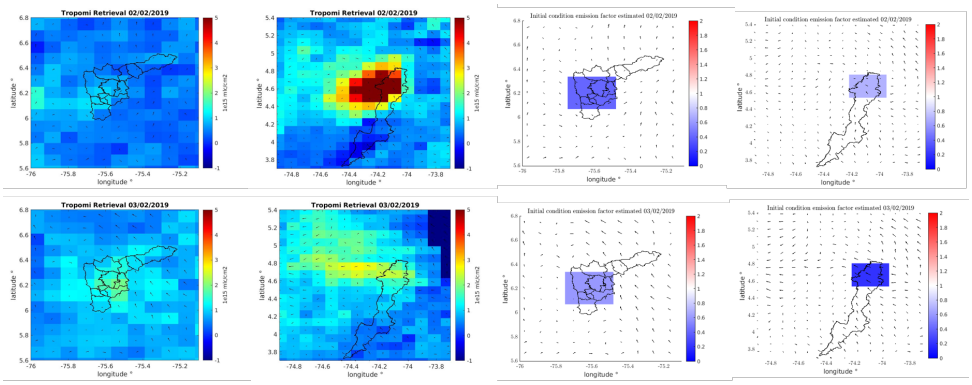


Figure 4.13: Comparison of the TROPOMI retrieval products for 2019-02-02 and 2019-02-03 zooming in the two principal Colombian cities, Medellín and Bogotá. Colorbar on the left panel represents concentrations. The right panels show the corresponding estimated values of the emission parameters that suggest a reduction in the emission for the two cities for the two assimilation days. Colorbar on the right panel represents emission factors.

Figure 4.14 shows the difference between the background NO_2 plumes retrieval and the retrieval with the estimated parameters. The meteorological wind fields drive the concentration plumes. The magnitude reduction due to the emission update is shown for the assimilated scenario, which reduces the concentration spatially along the wind trajectory.

Figure 4.15 shows the vertical cut profiles for Bogotá and Medellín, respectively, for the two assimilation days declared in the schematic of Figure 4.14 and the two scenarios before assimilation and with the assimilation. Bogotá is located over a

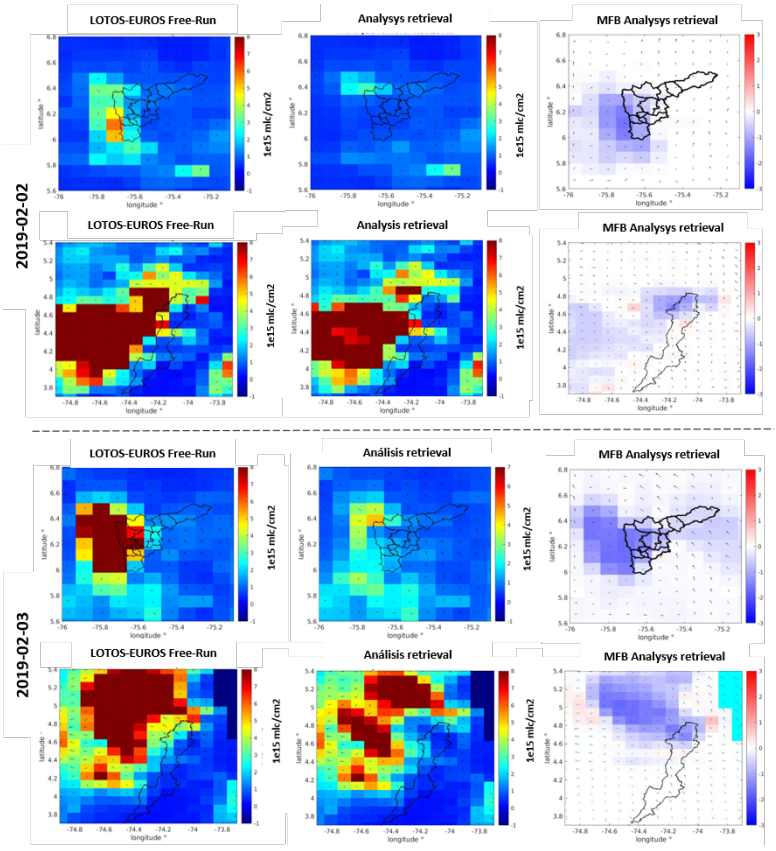


Figure 4.14: LOTOS-EUROS free run and the analysis run generated from the 4DEnVar results with the Mean Fractional Bias comparison for the 2019-02-02 (three first rows) and 2019-02-03 (three last rows). For each situation, the zoom is made for Bogotá and Medellín, and in the right column, the statistics are used to quantify the impact of those changes.

plateau, with the removal process dominant from east to west; the assimilation effect is noticeable regarding the magnitude of the concentration reduction in the plume transported in this direction. Medellín is located in a deep-seated valley (Aburrá Valley). The wind removes the contaminants in the order that overtakes the mountains around the Valley. Here, the assimilation effect is noticeable regarding the magnitude of the concentration reduction gathered inside the Valley.

Comparison with SIATA surface observations and OMI measurements

For the city of Medellín, surface measurements of NO_2 are available along the Valley from the ground-based sensor network measurements from *Sistema de Alerta Temprana del Valle de Aburrá* (SIATA) which provide time series observation with an hourly temporal resolution. Figure 4.16 shows the mean of all the available stations with their standard deviation spread against the non-assimilated and 4DEnVar assimilated model output. The assimilated trajectory starts from the horizontal red

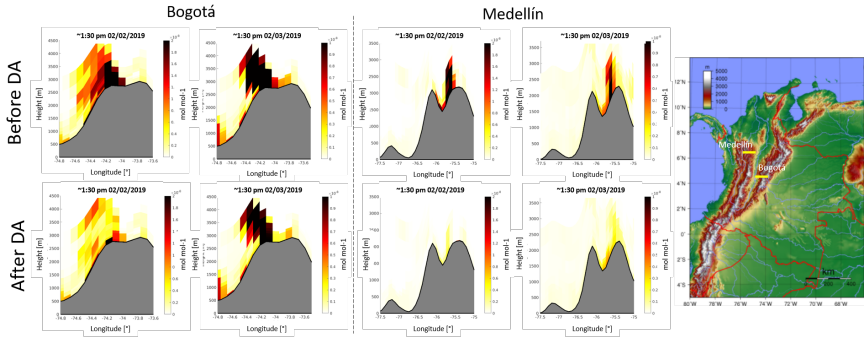


Figure 4.15: Transverse cut for NO_2 concentration over latitude=6.2518° for Medellín (left four panels) and latitude=4.609° for Bogotá (four central panels). The emission parameter update impacts the column of the model concentration output. The right panel shows the topography map of Colombia with the transversal cut location

dashed because the emissions were corrected for the model from this time after DA. The spatial distribution from the NO_2 SIATA network stations is in the right panel, where it is appreciated how the stations are located in the lower part of the Valley.

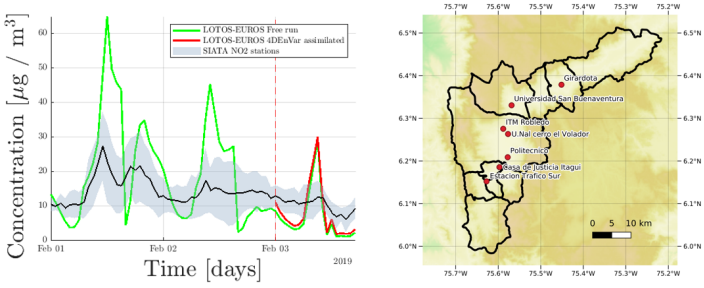


Figure 4.16: In the Left panel, Seven stations from the SIATA network of NO and NO_2, comparing the no assimilated model (green curve) and 4DEnVar assimilated model output (red curve). The black curve corresponds to observations and its representativity error is depicted with the grey uncertainty. Right panel: Location of the NO_2 measuring stations in the Aburrá Valley

Table 2 shows the mean fractional bias (MFB), the root mean square (RMSE), and the correlation factor comparison in the free run and assimilated propagation during the assimilation period against the observation mean. The estimated emission parameter suggests an increase in the emission value ($e > 1$) from the assimilation experiment in this location, which is appreciated in reducing the MFB and RMSE statistic and increasing the correlation factor.

For this area, the comparison was not too relevant upon modifying the initial conditions for the analysis value because the TROPOMI information was not very significant for the experiment period to promote a drastic parameter update. At the same time, in Medellín's vertical cut, the effect was more evident in the concentra-

Observation SIATA as reference	MFB	RMSE	Correlation
free run	-0.4520	7.3050	0.5485
Assimilation	-0.3226	7.0464	0.5864

Table 4.1: Comparison of Mean Fractional Bias (MFB), Root Mean Square (RMSE), and Correlation factor between the free run and assimilated propagation, relative to the observation mean during the assimilation period.

tion of higher levels seen in Figure 4.14. It was possible to see how the analysis searched the mean value from the ground observation. No more NO_2 measurements were available around the domain during this study.

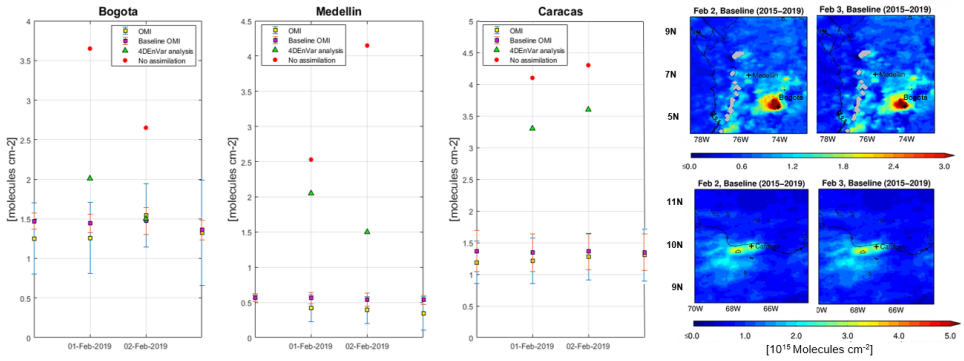


Figure 4.17: Left panel: comparison in time of the non-assimilated (red circles) and assimilated 4DEnVar (green triangles) concentrations with the OMI satellite measures (pink and yellow squares) around Bogotá, Medellin. The right panel shows the spatial comparison over the region in Colombia that covers Bogotá and Medellín in the upper panels and over Venezuela in the below panels.

Another comparison was made with the OMI measurements from the reference data [45]. Only data of excellent quality (VCD Quality Flags of 0) and cloud screened (Effective Cloud Fraction 30 percent) were included in the statistics. In Figure 4.17, two concentration values are shown; the OMI concentration for the days and similar statistics for 2020 were computed for each day during 2015-2019 to produce the ”baseline” value shown. The baseline values were then calculated as a weighted mean of values for each day’s overall years, weighted by the number of points used to compute each median. The magnitude of the concentration value simulated by the model decreases for all cities, approaching the magnitude of the OMI concentrations.

4.4. Conclusions

A 4DEnVar methodology has been developed to combine TROPOMI satellite observations with LOTOS-EUROS regional CTM simulations for the northwest of South America. Cities and municipalities in developing countries without local air quality and meteorological networks can use this technique, which avoids the implementa-

tion of the adjoint of the CTM to estimate coarse emissions inventories. This study defines emission correction factors as the uncertain parameters to be calculated by multiplying the nominal emission inventory. Modifying the correction factors by using the 4DEnVAR technique to assimilate the TROPOMI-NO2 columns improves the spatial and temporal concentration distribution of the NO_2 fields by the estimations of the emission factors. The emission update has substantially improved the agreement between the simulated and observed NO_2 fields. These findings showed that TROPOMI NO_2 concentrations could be used to reconstruct spatial and temporal variable NO_2 components, making enhancing temporal NO2 emission patterns relatively simple in a forward modeling setting.

It should be remembered that the updated inventory that takes the model forecasts closer to the assimilated observations is the subsequent emission inventory obtained from DA. Consequently, it involves a model error and does not reflect the actual emission scenario. One of the most relevant facts is that this technique relies on the generated ensemble's Gaussianity.

Although the satellite data have not yet been exploited for extended periods due to the high cloudiness, we have demonstrated how to use a 4DEnVAr data assimilation technique and take advantage of the relationships between observed and unobserved states of a chemical transport model to improve the model results. Adding satellite information to the model makes estimating a coarse emission inventory possible. It is also a good starting point for establishing higher-resolution emission inventories or improving boundary conditions for high-resolution nested simulations. The particular results from this case study suggest a decrease in the emission values in notorious places like the refinery corridor in the Venezuela region, driven by the drastic decrease in oil production over the last ten years.

In the future, we will refine the surface information, complementing the region's surface information with the satellite information and the meteorology from high-resolution mesoscale models such as WRF, to represent more accurate patterns, as we see in deep-seated narrow valleys such as in Medellín. We will also improve the data assimilation methodology by implementing localization techniques to reduce the well-known problems introduced by using a limited amount of ensembles.

4.4.1. Appendix: Performance metrics

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:

$$\text{MFB} = \frac{2}{M}\sum_{i=1}^{M}\frac{(H(c))_i - y_i}{(H(c))_i + y_i} \tag{4.18}$$

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 [46].

The *root mean square error* (RMSE) represents the sample standard deviation of the differences between predicted values and observed values Eq. 4.19. The RMSE penalizes a high variance as it gives errors with larger absolute values more weight than errors with smaller absolute values [47]:

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left((H(c))_i - y_i\right)^2} \tag{4.19}$$

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 [48]:

$$\text{Corr} = \frac{\sum_{i=1}^{M}\left((H(c)_i - \overline{(H(c)_i}\right)(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{M}\left((H(c)_i - \overline{(H(c)_i}\right)^2}\sqrt{\sum_{i=1}^{M}(y_i - \overline{y})^2}} \tag{4.20}$$

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

References

[1] 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 (2020), 10.1007/s11356-020-08268-4.

[2] A. Casallas, N. Celis, C. Ferro, E. L. Barrera, C. Peña, J. Corredor, and M. B. Segura, *Validation of pm 10 and pm 2.5 early alert in bogotá, colombia, through the modeling software wrf-chem,* Environmental Science and Pollution Research , 1 (2020).

[3] J. Verkaik and L. Ganzeveld, *Evaluation of Colombian Methane Emissions Combining WRF-Chem and TROPOMI* (Wageningen University and Research, Meteorology and air quality, MSc Thesis, 2019).

[4] 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**, 928 (2020).

[5] J. G. M. Barten, L. N. Ganzeveld, A. J. Visser, R. Jiménez, and M. C. Krol, *Evaluation of nitrogen oxides sources and sinks and ozone production in colombia and surrounding areas,* Atmospheric Chemistry and Physics Discussions **2019**, 1 (2019).

[6] J. F. Grajales and A. Baquero-Bernal, *Inference of surface concentrations of nitrogen dioxide (no2) in colombia from tropospheric columns of the ozone measurement instrument (omi),* Atmósfera **27**, 193 (2014).

[7] R. Timmermans, R. Kranenburg, A. Manders, C. Hendriks, A. Segers, E. Dammers, Q. Zhang, L. Wang, Z. Liu, L. Zeng, *et al.*, *Source apportionment of pm2. 5 across china using lotos-euros,* Atmospheric Environment **164**, 370 (2017).

[8] M. Thürkow, I. Kirchner, R. Kranenburg, R. Timmermans, and M. Schaap, *A multi-meteorological comparison for episodes of pm10 concentrations in the berlin agglomeration area in germany with the lotos-euros ctm,* Atmospheric Environment **244**, 117946.

[9] M. Schaap, R. M. Timmermans, M. Roemer, G. Boersen, P. Builtjes, F. Sauter, G. Velders, and J. Beck, *The lotos euros model: description, validation and latest developments,* International Journal of Environment and Pollution **32**, 270 (2008).

[10] A. M. Manders, P. J. Builtjes, L. Curier, H. A. Denier van der Gon, C. Hendriks, S. Jonkers, R. Kranenburg, J. J. Kuenen, A. J. Segers, R. Timmermans, *et al.*, *Curriculum vitae of the lotos–euros (v2. 0) chemistry transport model,* Geoscientific Model Development **10**, 4145 (2017).

[11] 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).

[12] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen, *Data assimilation in the geosciences: An overview of methods, issues, and perspectives,* Wiley Interdisciplinary Reviews: Climate Change **9**, e535 (2018).

[13] C. Liu, Q. Xiao, and B. Wang, *An ensemble-based four-dimensional variational data assimilation scheme. part i: Technical formulation and preliminary test,* Monthly Weather Review **136**, 3363 (2008).

[14] R. Bannister, *A review of operational methods of variational and ensemble-variational data assimilation,* Quarterly Journal of the Royal Meteorological Society **143**, 607 (2017).

[15] K. Aleksankina, S. Reis, M. Vieno, and M. R. Heal, *Advanced methods for uncertainty assessment and global sensitivity analysis of an eulerian atmospheric chemistry transport model,* Atmospheric Chemistry and Physics **19**, 2881 (2019).

[16] K. Aleksankina, M. R. Heal, A. J. Dore, M. V. Oijen, and S. Reis, *Global sensitivity and uncertainty analysis of an atmospheric chemistry transport model: the frame model (version 9.15. 0) as a case study,* Geoscientific Model Development **11**, 1653 (2018).

[17] V. Mallet and B. Sportisse, *Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: An ensemble approach applied to ozone modeling,* Journal of Geophysical Research: Atmospheres **111** (2006), https://doi.org/10.1029/2005JD006149.

[18] M. Guevara, S. Lopez-Aparicio, C. Cuvelier, L. Tarrason, A. Clappier, and P. Thunis, *A benchmarking tool to screen and compare bottom-up and top-down atmospheric emission inventories,* Air Quality, Atmosphere & Health **10**, 627 (2017).

[19] 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**, 49 (2018).

[20] F. Liu, R. J. van der A, H. Eskes, J. Ding, and B. Mijling, *Evaluation of modeling no 2 concentrations driven by satellite-derived and bottom-up emission inventories using in situ measurements over china,* Atmospheric Chemistry and Physics **18**, 4171 (2018).

[21] M. Bocquet, H. Elbern, H. Eskes, M. Hirtl, R. Žabkar, G. Carmichael, J. Flemming, A. Inness, M. Pagowski, J. Pérez Camaño, *et al.*, *Data assimilation in*

atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models, Atmospheric chemistry and physics **15**, 5325 (2015).

[22] E. M. Constantinescu, A. Sandu, T. Chai, and G. R. Carmichael, *Ensemble-based chemical data assimilation. i: General approach,* Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography **133**, 1229 (2007).

[23] 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).

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

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

[26] A. C. Lorenc and M. Jardak, *A comparison of hybrid variational data assimilation methods for global nwp,* Quarterly Journal of the Royal Meteorological Society **144**, 2748 (2018).

[27] E. Emili, S. Gürol, and D. Cariolle, *Accounting for model error in air quality forecasts: an application of 4denvar to the assimilation of atmospheric composition using qg-chem 1.0,* Geoscientific Model Development **9**, 3933 (2016).

[28] R. 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).

[29] H. Eskes, L. Curier, and A. Segers, *Assimilation of surface and satellite observations with the lotos-euros air quality model and the ensemble kalman filter technique,* EGUGA , 12805 (2012).

[30] A. Barbu, A. Segers, M. Schaap, A. Heemink, and P. Builtjes, *A multi-component data assimilation experiment directed to sulphur dioxide and sulphate over europe,* Atmospheric Environment **43**, 1622 (2009).

[31] P. N., S. J., R. A. Posada J., Q. L., and A. Yarce, *Potential urban pollution impacts on protected areas in colombia through atmospheric teleconnections,* CMAS Community Modeling and Analysis system , 3 (2017).

[32] L. S. A, Y. N, A. Pinel, S. O, Q. A, and Heemink, *Forecasting PM_{10} and $PM_{2.5}$ in the aburrá valley (medellín, colombia) via enkf based data assimilation,* Mathematics in science and engineering (2020), https://doi.org/10.1016/j.atmosenv.2020.117507.

[33] G. Fu, A. Heemink, S. Lu, A. Segers, K. Weber, and H. X. Lin, *Model-based aviation advice on distal volcanic ash clouds by assimilating aircraft in situ measurements,* Atmospheric Chemistry and Physics **16**, 9189 (2016).

[34] 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).

[35] C. Liu, Q. Xiao, and B. Wang, *An ensemble-based four-dimensional variational data assimilation scheme. part ii: Observing system simulation experiments with advanced research wrf (arw),* Monthly Weather Review **137**, 1687 (2009).

[36] J.-i. Kurokawa, K. Yumimoto, I. Uno, and T. Ohara, *Adjoint inverse modeling of nox emissions over eastern china using satellite observations of no2 vertical column densities,* Atmospheric Environment **43**, 1878 (2009).

[37] T. Chai, G. R. Carmichael, Y. Tang, A. Sandu, A. Heckel, A. Richter, and J. P. Burrows, *Regional nox emission inversion through a four-dimensional variational approach using sciamachy tropospheric no2 column observations,* Atmospheric Environment **43**, 5046 (2009).

[38] T. Zijlker, *Estimating nox emissions using s5-p tropomi: An adjoint-free 4dvar approach,* TuDelft university (2020).

[39] Y. Wang, J. Wang, X. Xu, D. K. Henze, Z. Qu, and K. Yang, *Inverse modeling of so 2 and no x emissions over china using multisensor satellite data–part 1: Formulation and sensitivity analysis,* Atmospheric Chemistry and Physics **20**, 6631 (2020).

[40] M. J. Cooper, R. V. Martin, D. K. Henze, and D. Jones, *Effects of a priori profile shape assumptions on comparisons between satellite no 2 columns and model simulations,* Atmospheric Chemistry and Physics **20**, 7231 (2020).

[41] K. Miyazaki, H. Eskes, and K. Sudo, *Global no x emission estimates derived from an assimilation of omi tropospheric no 2 columns,* Atmospheric Chemistry and Physics **12**, 2263 (2012).

[42] C. Ma, T. Wang, A. P. Mizzi, J. L. Anderson, B. Zhuang, M. Xie, and R. Wu, *Multiconstituent data assimilation with wrf-chem/dart: Potential for adjusting anthropogenic emissions and improving air quality forecasts over eastern china,* Journal of Geophysical Research: Atmospheres **124**, 7393 (2019).

[43] W. Skamarock, J. Klemp, J. Dudhia, D. Gill, Z. Liu, J. Berner, W. Wang, J. Powers, M. Duda, D. Barker, *et al.*, *A description of the advanced research wrf model version 4 (no. ncar/tn-556+ str),* National Center for Atmospheric Research: Boulder, CO, USA (2019), doi:10.5065/1dfh-6p97.

[44] C.-W. Su, K. Khan, R. Tao, and M. Umar, *A review of resource curse burden on inflation in venezuela,* Energy **204**, 117925 (2020).

[45] L. N. Lamsal, N. A. Krotkov, A. Vasilkov, S. Marchenko, W. Qin, E.-S. Yang, Z. Fasnacht, J. Joiner, S. Choi, D. Haffner, *et al.*, *Ozone monitoring instrument (omi) aura nitrogen dioxide standard product version 4.0 with improved surface and cloud treatments,* Atmospheric Measurement Techniques **14**, 455 (2021).

[46] 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.

[47] 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).

[48] 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 .

5

Conclusions and recommendations

5.1. Conclusions

The road to develop simulation operational scenarios to represent accurately the air quality through the incorporation of real-time satellite and ground observation in this region is beginning. This thesis, which was formulated based on the objectives proposed in chapter 1, presents the conclusions in the following manner:

Development of a CTM for the Colombia region using the LOTOS-EUROS modeling system.

The LOTOS-EUROS model was implemented in the northwestern region of South America with a focus on the Colombian territory. To enhance the model's input information and deliver the best conceivable representation of atmospheric chemical concentrations in the target area, updates were carried out on land use, orography, and meteorology data. In Chapter 2 these updates aimed to improve the representation of complex interactions and dynamics that contribute to the distribution of pollutants in the atmosphere by reducing model uncertainties in LOTOS-EUROS. Simulations were conducted considering experiments on point sources to quantify the impact of pollutant emissions from major Colombian cities at different times throughout the year. The analysis sought to identify natural areas that may be vulnerable to atmospheric deposition pollutants by measuring levels of pollutants per unit area in the designated protected areas of interest.

To evaluate the potential benefits of satellite measurements in improving emission data within the LOTOS-EUROS model, specifically for the Colombian region.

This dissertation explores and utilizes satellite measurements for the northern region of South America, focusing on the data of NO_2 from the TROPOMI instrument. In Chapter 3 a methodology was developed to prepare the satellite information for assimilation into the LOTOS-EUROS model. The preprocessing procedure

implemented involved downloading the data, cropping to a specified region, and integrating it into a column format to ensure compatibility with the model's assimilation algorithms. By utilizing this methodology, the satellite data was proficiently integrated into the CTM, thus enhancing the accessible information and refining the model's output accuracy. The data was prepared accordingly for the ensuing Data Assimilation procedures.

Evaluate the influence of a low-cost sensor network on the improvement of emission data within the LOTOS-EUROS model, focusing on the Colombia region.

We have created an affordable hardware solution that functions as a ground node and a gateway in air quality networks. In Chapter 3 these low-cost sensors were designed and integrated into the model to enhance satellite-derived concentrations using data assimilation techniques. Though cheaper and easier to install than traditional ground-based sensors, these sensors provide essential real-time measurements of pollutant concentrations. We examine the intricacies associated with building and producing a device appropriate for in-situ air quality assessments and long-range data transmission from Colombia. Despite Colombia not being a primary producer of electronic technology, it plays an integral role in its integration. Thus, future endeavors should focus on striking a balance between operational production and the costs associated with technological integration. By integrating the data from these sensors, subsequent operational Chemical Transport Models could profit from these evaluations to increase their overall effectiveness and deliver more dependable estimations of pollutant concentrations. The development of this innovative device is motivated by the need for novel data sources to improve the Data Assimilation of LOTOS-EUROS in Colombia.

To use adjoint-free data assimilation techniques to estimate uncertain parameters within the LOTOS-EUROS model.

Implementing Chemical Transport Models (CTMs) is indispensable for quantifying the concentrations of diverse atmospheric pollutants in regions with insufficient measured information. However, these models inherently carry uncertainties, needing the integration of additional data sources to improve their performance. One effective approach to address the limitations of CTMs is the assimilation of satellite-derived concentrations and low-cost sensors.In Chapter 3 the Local Ensemble Kalman Filter (LEnKF) and in Chapter 4, the 4D Ensemble Variational (4DEnVar) , were successfully implemented over the study region. These techniques allowed the correction of emission parameters, which play a crucial role in accurately simulating pollutant concentrations.

In summary, the successful implementation of CTMs combined with the assimilation of satellite-derived concentrations and low-cost sensor data has significantly advanced our ability to quantify atmospheric pollutant concentrations and estimate emissions and wind direction parameters. The updates and improvements made to the LOTOS-EUROS model, the development of a preprocessing methodology for satellite information, and the integration of low-cost sensors have collectively contributed to a more comprehensive and accurate representation of atmospheric chemistry over the study region. Additionally, in the recommendations section us-

ing the stream function data assimilation expanded our capabilities in optimizing emission parameters and capturing directional information in chemical transport modeling. These advancements hold significant potential for addressing environmental challenges and informing policy decisions related to air quality management in Colombian territory.

5.2. Recommendations

Based on the research and findings, the subsequent recommendations are advanced for maintaining a current and effective Chemical Transport Model:

1. Transition to Higher Resolution Meteorology Input: Shift from the ECMWF meteorological input to non-hydrostatic formulations, such as WRF or HARMONIE. This would improve vertical transport and dynamics representation, providing greater detail for regional scale modeling.

2. Expand the Use of Virtual Sensors: Encourage the integration of both physical and virtual sensors in the modeling process. Virtual sensors, through their intelligent capabilities, can provide estimations of other vital parameters, enhancing the accuracy and robustness of the model.

3. Optimized Utilization of Satellite Data: Recognizing the scarcity of satellite data for certain regions, it is imperative to maximize its utility. This can be achieved by integrating available data through data assimilation into models better suited for regional representation. Continuously explore and incorporate new satellite data sources. Different satellites offer varied standardization mechanisms for concentration retrievals, which could potentially enrich the model.

4. Emphasis on Wind Field Corrections: Prioritize the correction of wind fields. It is evident from different meteorological sources that there exist systematic discrepancies in wind direction data. Addressing these will significantly improve the accuracy of the inversion problem when estimating emissions.

5. Leverage Data Assimilation Strategies with Low-Cost Sensors: Actively incorporate data assimilation techniques to enhance the performance of low-cost sensor technologies. This mutualistic relationship allows for a more refined and cost-effective model, providing better accuracy and reliability.

Incorporating these recommendations will not only streamline the process of maintaining an up-to-date Chemical Transport Model but also ensure that the model remains responsive to the dynamic needs and realities of the region it represents.

5.3. On correcting wind fields

Other parameters were noted as important because they introduce uncertainty to the system. The wind direction is significant to take into account because it is possible to observe variations in direction across different weather models concerning plumes of pollutants detected from satellite observations. In the material in the appendix, we propose a data assimilation (DA) approach for estimating emission and wind direction parameters in an advection-diffusion model. To evaluate the effectiveness of the proposed approach, a low-dimensional advection-diffusion model was used with varying numbers of observations. The model's emission and wind

parameters were assumed to be both uncertain. The parameters are estimated using the Ensemble Kalman filter with an augmented state vector for the parameters. These sequential data assimilation (DA) techniques utilize the ensemble of multiple model realizations to decrease state and parameter representation uncertainty. An associated stream function with a divergence-free condition governs the wind fields. Estimation of this stream function through the assimilation process allows for corrections to the wind fields without violating physical conservation laws.

Appendix: Wind and emission parameter estimation data assimilation with an associate stream function

THE ROAD TO WISDOM
The road to wisdom? Well, it is plain and simple to express:
Err
and err
and err again
but less
and less
and less.

Piet Hein

The present study proposes a data assimilation (DA) approach for estimating emission and wind direction parameters in an advection-diffusion model. This implementation aims to improve the prediction of a chemical transport model over long distances by updating the emission operator in the model using DA techniques. As a first step, we want to test the method in a small-scale scenario. A low-dimensional advection-diffusion model was utilized to evaluate the effectiveness of the proposed approach under various sampling observation numbers. The model's emission and wind parameters are perturbed as a source of uncertainty. The parameters are sequentially estimated with the Ensemble Kalman filter with an augmented state vector. These sequential DA techniques exploit the ensemble of multiple model realizations to reduce uncertainty in the state and parameter representation. An associated stream function with a divergence-free condition controls the wind fields, and the estimation of this stream function through the assimilation process allows corrections of the wind fields without violating physical

Parts of this chapter have been published in Eurosim proceedings conference 2023 and submitted in the conference Springer journal 2023

laws. The technique's performance was compared against validation observations such as the Root-Mean Square (RMS). This study demonstrates the potential of the proposed DA approach for improving the prediction of transport in the advection-diffusion model through parameter estimation.

A.1. Introduction

One of the primary sources of uncertainty in CTMs comes from the emission inventories, which could differ significantly from the technique's location and time against reality. We inspired or contributed to a real-life scenario where we have a CTM model in a region with coarse meteorology input information, so the winds are the secondary source of uncertainty between the model and real directions.

Emission parameter estimation has been the center of attention for DA around the globe for CTM implementations, mainly because the emission inventories tend to have high uncertainty values, with some different degrees of uncertainty between regions and source sectors [1, 2]. In South America, updating emission inventories can be challenging due to historical institutional limitations, making it difficult to have up-to-date information available. This significant variability in CTMs from the emission parameters [3] has been given the principal attention in DA techniques to incorporate observations into numerical models for determining the most appropriate state and parameter sets that produce a new state that follows this observation [4]. DA method is commonly used to incorporate measurements into the system, estimate its states and parameters, and enhance subsequent simulation forecasts, making it a valuable tool for improving model performance.

The state and model parameters, emissions and winds, can be estimated simultaneously using the Ensemble Kalman Filter (EnKF) to estimate the error covariances among several variables with a small-scale model of a CTM, an emission advection-diffusion model. EnKF is the DA technique that generates an update at the first moment of the probability density function found in the analysis once new observations are available[5, 6]. An ensemble of model propagation is created by perturbing emission factors and wind direction parameters. The method for estimating the parameters using the data provided by the observations uses covariances that include the relationship between states and initial conditions and compliance with the observed stated. The "augmented state space" vector technique, proposed by Jazwinsky in 1970 [7], estimates parameters. Here, both the state variables and the model parameters are part of the state vector and can be estimated through the sequential steps of the filter. This approach has been used in different parameter estimation studies, specifically with the LOTOS-EUROS mode" as CTM [8, 9].

This chapter is organized as follows: In Section A.2, we introduce the specific methodology for the EnKF, including the stochastic model for parameter perturbation and propagation, as well as the technical details of the code and stream function formulation. In Section A.3, we focus on the results of this technique with the advection-diffusion model. Section A.4 presents our discussion and concluding remarks. This study offers valuable information on the effectiveness of stream function DA in improving CTM and can contribute to developing more accurate air quality monitoring systems.

A.2. Parametrization

This section presents the representation of the stochastic parameters. Finally, the stream function formulation is introduced.

A.2.1. Stochastic uncertainty representation for wind and emission parameters

A stochastic representation of the parameter uncertainty is required to implement the DA method. The emissions employed by the model operator are consequently modeled as a stochastic process with a factor of random variation.

$$\hat{\mathbf{e}}_t \;=\; \mathbf{e}_t \;(\; 1 \;+\; \delta\mathbf{e}_t) \;. \tag{1}$$

Here, $\mathbf{e}_t$ is the nominal emission from the emission inventory. The emission deviation is modeled as a colored noise process [7] as follows:

$$\delta\mathbf{e}_t \;=\; \alpha\;\delta\mathbf{e}_{t-1} + \sigma\sqrt{1-\alpha^2}\;\mathbf{w}^{\mathbf{e}}_{t-1}\;, \tag{2}$$

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

$$\mathbf{w}^{\mathbf{e}}_t \;\sim\; \mathcal{N}(0,\mathbf{I})\;. \tag{3}$$

A.2.2. Stream function formulation

Consider the scalar stream function:

$$\mathbf{\Psi}(x,y)\;, \tag{4}$$

where

$$\mathbf{u} = \frac{\partial\mathbf{\Psi}}{\partial y}\;, \tag{5}$$

$$\mathbf{v} = -\frac{\partial\mathbf{\Psi}}{\partial x}\;. \tag{6}$$

This scalar stream function constructs the vector field, satisfying the divergence-free property $\nabla\cdot f = 0$:

$$f(x,y) = \begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix} \tag{7}$$

The stream function $\mathbf{\Psi}(x,y)$ is a scalar function describing fluid flow in a two-dimensional space. The stream function divergence-free property $\nabla\cdot f = 0$ means the total flow across any closed curve in the fluid domain is zero. This feature is beneficial because it allows physical constraints, such as continuity or conservation laws. The proposed approach involves perturbing the stream function instead of the two wind velocities. The stream function's divergence-free property is characteristic of an auxiliary space in DA used for parameter estimation, which can have some

properties that help describe the magnitudes related to the model space. Section A.3 provides an illustrative example of this method, where we select a suitable stream function to demonstrate its effectiveness. The stream function by the model operator are consequently modeled as a stochastic process with a factor of random variation.

$$\mathbf{\Psi}_t \;=\; \beta\, \mathbf{\Psi}_{t-1} + \eta\sqrt{1-\beta^2}\, \mathbf{w}^{\Psi}_{t-1}\;, \tag{8}$$

where $\mathbf{w}^{\Psi}_t$ is a white noise process with zero mean and $\mathbf{Q}$ the correlation matrix:

$$\mathbf{w}^{\Psi}_t \;\sim\; \mathcal{N}(0, \mathbf{Q})\;. \tag{9}$$

The spatial correlation has to be designed according to the similarity of the natural phenomena where wind fields have long-range correlations. Moreover, the stream function divergence-free property reduces the dimensionality of the DA problem. The stream function can transform the state vector into a new space with a smaller dimension, reducing computational costs. The stream function can be used as an auxiliary space for DA by transforming the state vector.

From the parametrization, we create the following augmented vector and calculate the approximate covariance matrix of this augmented system:

$$\begin{bmatrix} \mathbf{x}_t \\ \delta\mathbf{e}_t \\ \Psi_t \end{bmatrix} = \begin{bmatrix} \mathcal{M}_{t-1,t}\left(\mathbf{x}_{t-1}(\delta\mathbf{e}_{t-1}, \mathbf{\Psi}_{t-1})\right) \\ \alpha\ \delta\mathbf{e}_{t-1} \\ \beta\ \mathbf{\Psi}_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ \sigma\sqrt{1-\alpha^2} \\ \eta\sqrt{1-\beta^2} \end{bmatrix} \cdot \mathbf{W}_{t-1} \tag{10}$$

Using the augmented vector (10), it is possible to estimate the state of the emission correction factor and the stream function using a sequential DA approach. The nonlinear operator $\mathcal{M}$ propagates the augmented state vector x in time, while the right part of the expression corresponds to the stochastic forcing $\mathbf{W}_{t-1}$ over the elements of the state, where $\mathbf{W}_{t-1}$ correspond to

$$\begin{bmatrix} 0 \\ \mathbf{w}^{\mathbf{e}}_{t-1} \\ \mathbf{w}^{\mathbf{\Psi}}_{t-1} \end{bmatrix} \tag{11}$$

A.3. Results

The advection-diffusion model is presented in this section, followed by the different parameter estimation results for the homogeneous wind field estimating only one parameter and for the stream function case, which leads to wind direction corrections and magnitudes estimating the number of parameters as the stream function size.

A.3.1. Advection-diffusion model

Testing an implementation with a simplified model is widely used in the DA field to check the behavior of new techniques designed. This section considers a 2D advection-diffusion model to which synthetic data is fed through data assimilation

(DA). The advection-diffusion model for a gas concentration C is described by the following equation in the directions x and y with constant diffusion coefficients D_x and D_y, and velocities in the two directions U and V:

$$\frac{\partial C}{\partial t} = -U\frac{\partial C}{\partial x} - V\frac{\partial C}{\partial y} + D_x\frac{\partial^2 C}{\partial x^2} + D_y\frac{\partial^2 C}{\partial y^2} + E_k \ . \tag{12}$$

Synthetic data in this scenario correspond to the data sampled from the true model states of the advection-diffusion model propagation with different initial conditions. The true model simulation is one simulation of the model that is taken as reality. This model simulation is perturbed, and then this synthetic data sampled here is the observation dataset to assimilate. The change in parameters presents a challenge for DA and predictability. The size of this simple test model is 3600 states (60 x 60 grid) for the concentrations. The emission parameter depends on the number of sources to be updated, and the wind direction parameter is two times the number of states because the horizontal wind directions are u and v. In this case, one advantage of the stream function is reducing the number of wind direction parameters in half because once the stream function parameters are estimated, the wind direction is calculated in the model space through the gradient.

For the discretization form, the central step second order finite difference [10] was used to generate the propagating following expression from an initial state. The boundary condition for solving the experiment was the Dirichlet homogeneous zero or null value fixed in the contour [11].

A.3.2. Results advection-diffusion model

Figure (1) provides a schematic normally-distributed perturbation for the wind direction along with a comparison of various standard deviation values of the wind direction perturbation for the advection-diffusion model as a grey shade region that describes a set of variations in the direction of the wind from the ideal or true value. In comparison, the different standard deviation values of the wind direction perturbation in the advection-diffusion model Eq. 12 determine the region of uncertainty expanded where the variation in the modeled wind direction occurs.

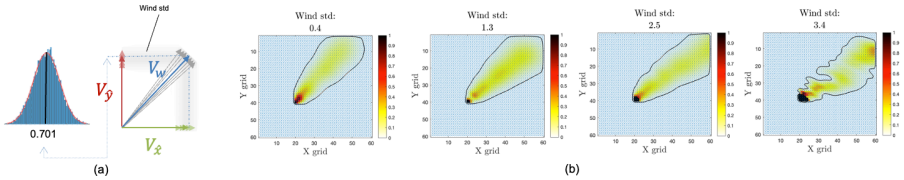


Figure 1: (a) Schematic perturbation for the wind direction and (b) Comparison for different standard deviation values of the wind direction perturbation for the advection-diffusion model. Colorbar indicates concentration.

The Figure provides insights into how perturbations in the wind direction can impact the accuracy of the advection-diffusion model. It highlights the importance

of accurately estimating wind direction in improving model accuracy. Furthermore, comparing the different standard deviation values of the wind direction perturbation is a useful guide for determining the maximum level of uncertainty or variability that should be considered when modeling wind direction for various applications. The wind fields V_w were homogeneously perturbated in direction, assuming that the magnitudes remain constant and equal to $|V_w| = 1$. The noise distribution for generating the distribution is $\mathcal{N}_y \sim \mathcal{N}(\frac{1}{\sqrt{2}}, R)$

We use the following linear function to test the case with the scalar stream function as an auxiliary space to perform the DA for the parameter estimations.

$$\boldsymbol{\Psi}_b(x, y) = u_b y - v_b x \ , \tag{13}$$

where

$$u_b = \frac{\partial \boldsymbol{\Psi}_b}{\partial y} \ , \tag{14}$$

$$v_b = -\frac{\partial \boldsymbol{\Psi}_b}{\partial x} \ . \tag{15}$$

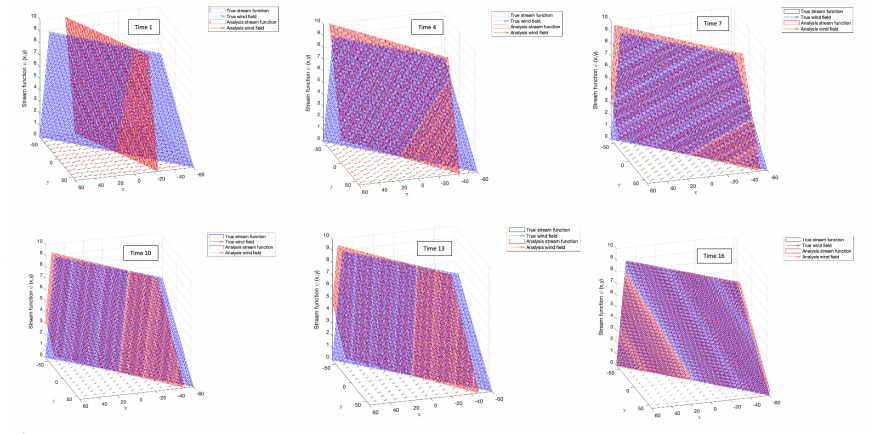


Figure 2: Comparison of the stream function correction after time through the sequential DA, which estimates emission and wind direction parameters, the true and analysis wind direction correspond to the gradient direction of the stream function, which ends align through time.

Figure 3 compares the stream function correction after sequential DA, which involves estimating emission and wind direction parameters. This Figure illustrates the true and analysis wind direction corresponding to the gradient direction, depicted with the red and blue arrows, respectively. Over time, the ends of the arrows align, indicating that the DA process successfully corrected the stream function. This result highlights the potential of DA techniques to improve model accuracy and the importance of accurately estimating emission and wind direction parameters.

The notion of the true state as a homogeneous field is unrealistic, so we created a true state by designing the covariance structure. The following diagram depicts the process of employing a diagonally symmetric matrix, with magnitudes defined by the next expression.

$$\mathbf{C}_{i,j} = \sigma \exp(\beta \cdot (r_{i,j})^2) \tag{16}$$

Where i and j are the row and column matrix members of the matrix and the $r_{i,j}$ in the distance between states.

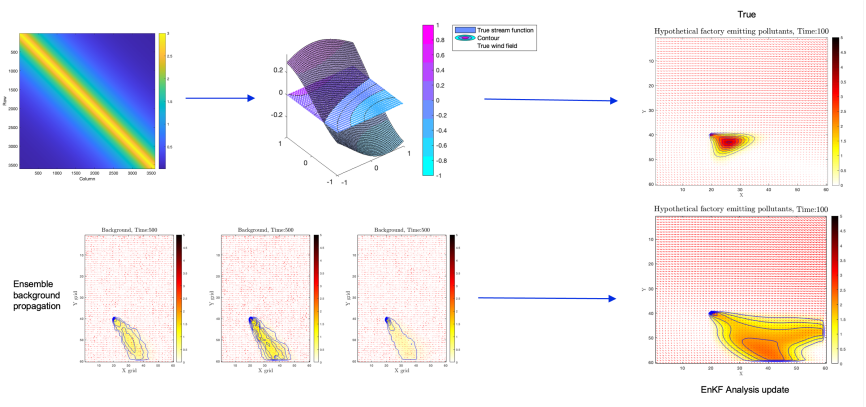


Figure 3: Schematic with the process to generate more realistic structures from a field with correlation structure $\mathbf{C}$ that are created from the covariance. The gradient of this true field constitutes the wind-driven field in the two horizontal directions. This is a preliminary result

The filter is applied, correcting both parameters emissions and wind directions in the stream function space. In this case, different than the homogeneous field estimation, the number of parameters of the wind direction is not one but is the number of states that increase the complexity.

A.4. Discussion

DA techniques have been increasingly incorporated into CTM to improve their accuracy. Most of these models use DA to improve emission estimates by ingesting new data. The next step in our research is to test the stream function DA technique with advection-diffusion in a larger model, such as the LOTOS-EUROS, to evaluate its effectiveness. The results of this work can provide valuable insights into the application of DA techniques in different domains and contribute to the ongoing efforts to improve the accuracy of CTMs. Atmospheric trace gasses measurements retrieved from Satellite instruments can perceive different plumes of pollutants emitted from sources, and recently, different works propose techniques to detect them [12–14], thus estimating the wind direction through the incorporation to the model of concentration plumes can act as a virtual sensor of wind direction.

One of the significant advantages of the proposed approach is the estimation of the associated stream function that controls the wind fields. The assimilation of

this stream function allows for the correction of the wind fields while preserving the physical laws. Considering the uncertainties associated with wind direction, the proposed approach provides a more accurate representation of chemical transport in advection-diffusion models. This study demonstrates the potential of the proposed DA approach for improving the accuracy of CTM in advection-diffusion models. The next crucial step towards advancing our research is introducing spatial correlation to the stream function, thereby generating a diverse range of spatially varying velocity patterns to introduce divergence-free structures on the wind fields to be estimated.

References

[1] S.-Y. Park, U. K. Dash, J. Yu, K. Yumimoto, I. Uno, and C. H. Song, *Implementation of an ensemble kalman filter in the community multiscale air quality model (cmaq model v5. 1) for data assimilation of ground-level pm 2.5,* Geoscientific Model Development **15**, 2773 (2022).

[2] E. Solazzo, M. Crippa, D. Guizzardi, M. Muntean, M. Choulga, and G. Janssens-Maenhout, *Uncertainties in the emissions database for global atmospheric research (edgar) emission inventory of greenhouse gases,* Atmospheric Chemistry and Physics **21**, 5655 (2021).

[3] M. Bocquet, H. Elbern, H. Eskes, M. Hirtl, R. Žabkar, G. Carmichael, J. Flemming, A. Inness, M. Pagowski, J. Pérez Camaño, *et al.*, *Data assimilation in atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models,* Atmospheric chemistry and physics **15**, 5325 (2015).

[4] J. Mo, S. Gong, J. He, L. Zhang, H. Ke, and X. An, *Quantification of so2 emission variations and the corresponding prediction improvements made by assimilating ground-based observations,* Atmosphere **13**, 470 (2022).

[5] G. Evensen and P. J. Van Leeuwen, *An ensemble kalman smoother for nonlinear dynamics,* Monthly Weather Review **128**, 1852 (2000).

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

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

[8] I. Skoulidou, M.-E. Koukouli, A. Segers, A. Manders, D. Balis, T. Stavrakou, J. van Geffen, and H. Eskes, *Changes in power plant nox emissions over northwest greece using a data assimilation technique,* (2021), https://doi.org/10.3390/atmos12070900.

[9] S. Lopez-Restrepo, E. D. Nino-Ruiz, L. G. Guzman-Reyes, A. Yarce, O. Quintero, N. Pinel, A. Segers, and A. Heemink, *An efficient ensemble kalman filter implementation via shrinkage covariance matrix estimation: exploiting prior knowledge,* Computational Geosciences **25**, 985 (2021).

[10] P.-S. Laplace and R. Courant, *The finite difference method,* .

[11] A. H.-D. Cheng and D. T. Cheng, *Heritage and early history of the boundary element method,* Engineering analysis with boundary elements **29**, 268 (2005).

[12] D. P. Finch, P. I. Palmer, and T. Zhang, *Automated detection of atmospheric no 2 plumes from satellite data: a tool to help infer anthropogenic combustion emissions,* Atmospheric Measurement Techniques **15**, 721 (2022).

[13] A. K. Georgoulias, K. F. Boersma, J. Van Vliet, X. Zhang, P. Zanis, J. de Laat, *et al.*, *Detection of no2 pollution plumes from individual ships with the tropomi/s5p satellite sensor,* Environmental Research Letters **15**, 124037 (2020).

[14] G. Kuhlmann, G. Broquet, J. Marshall, V. Clément, A. Löscher, Y. Meijer, and D. Brunner, *Detectability of co 2 emission plumes of cities and power plants with the copernicus anthropogenic co 2 monitoring (co2m) mission,* Atmospheric Measurement Techniques **12**, 6695 (2019).

Acknowledgements

I would like to extend my deepest gratitude to a group of remarkable individuals, whose support and guidance have been invaluable throughout my journey. Their contributions have been a cornerstone of my success, and for this, I am eternally grateful.

- **Arnold W. Heemink:** For his unwavering patience and the wisdom he imparted. His willingness to listen and share insightful advice has been a guiding light in my academic pursuit.
- **Olga Lucia Quintero:** Her mentorship, trusted advice, and the bond of friendship we've built over the years have been nothing short of inspirational.
- **Nicolas Pinel Pelaez:** For his companionship in sculpting the research problem of this thesis and for the enduring friendship that ensued. His motivation was a crucial element in my journey.
- **Arjo Segers:** I am grateful for his charisma and the patience he showed me. His ability to clarify doubts and offer perspectives on problem-solving has been invaluable.
- **My Mother, Sylvia Botero Hoyos:** For the gift of life, a nurturing education, and her unwavering support during the most challenging times. Her strength and love have been my bedrock.
- **My Father, Luis Fernando Yarce Ospina:** For his life lessons, educational guidance, and constant support through thick and thin. His wisdom has been a constant source of inspiration.
- **Maria Juliana Yepes Burgos:** The love of my life and my best friend. Her courage in joining me on the uncertain path of emigration has enriched my life immeasurably.
- **Santiago Lopez Restrepo:** His support was critical in helping me navigate complex mathematical concepts. I am equally thankful for his humor, which brought light to darker times.
- **Jhon Edinson Hinestroza:** For his camaraderie during challenging phases of the PhD program, enriching mathematical discussions, and unwavering friendship.
- **Alvaro Gonzales Garcia:** I am grateful for the time he spent discussing ideas with me, his companionship, and his valuable input in reviewing and editing this thesis.

- **Enrique Guarnieri** and **Amey Vasulkar:** For their friendship and support that greatly enhanced my student life at the TU.
- **Enrico Dammers:** His patience and assistance in clarifying my doubts across various domains – from modeling and observations to data assimilation – have been a great help.
- **Michiel Van Weele:** For his understanding and support during my transition to a new job, and his readiness to aid in my integration into the KNMI environment.

Each of these individuals has played a significant role in my academic and personal growth, and I am profoundly grateful for their contributions to my journey. Thank you all for being a part of this significant chapter in my life.

Andrés Yarce Botero, Delft, 2024

Curriculum Vitæ

Andrés YARCE BOTERO

Andrés Yarce Botero was born the 24-12-1987 in Medellin, Colombia. Due to his father's work his family lived in Bogotá and from 1994–2005 he went to primary school in Colegio Italiano Leonardo Da Vinci, Bogotá Colombia (1994–2000) and in Colegio Padre Manyanet, Medellin Colombia (2001–2005) once his family moved back to Medellin.

From 2006–2014 he studied in Physics in Universidad de Antioquia, Medellin Colombia (2006–2009) not finishing due and later studied Physical Engineering in Universidad EAFIT, Medellin Colombia (2010–2014).

After finishing the bachelor starts a MsC in Applied Physics Universidad EAFIT, Medellin Colombia (2015–2016).

From 2017-2022 he will pursue a PhD in Applied Mathematics/Mathematical Engineering TuDelft in Delft, the Netherlands and in Universidad EAFIT, Medellin, Colombia with the thesis entitled Data assimilation in a LOTOS-EUROS chemical transport model for Colombia using satellite measurements.

During his Ph.D. he worked as a postdoc at TuDelft-KNMI-VU PIPP project "High Resolution to Exploit Nitrogen Chemical Compound Observations from Space in the Netherlands" (2021—2024).

List of Publications and contributions

13. **Yarce Botero, A.**, Lopez-Restrepo, S., Pinel Peláez, N., Quintero, O. L., Segers, A., Heemink, A. W., *Medellin Air Quality Initiative (MAUI).*, Environmental Sustainability: Preparing for Tomorrow **18**, 99 (2021).

12. **Yarce Botero, A. et al.**, *Design and Implementation of a Low-Cost Air Quality Network for the Aburra Valley Surrounding Mountains.*, Pollutants 3.1 **17**, 549 (2023).

11. **Yarce Botero, A.**, Lopez-Restrepo, S., Pinel Peláez, N., Quintero, O. L., Segers, A., Heemink, A. W., *Estimating NOx LOTOS-EUROS CTM Emission Parameters over the Northwest of South America through 4DEnVar TROPOMI NO2 Assimilation*, Atmosphere, 12(12), 1633. (2021).

10. **Yarce Botero, A.**, Quintero, O. L., Segers, A., Heemink, A. W., *Estimating Wind and Emission Parameters in an atmospheric transport model*, Eurosim (2023)(Submitted).

9. **Yarce Botero, A.**, Santiago Lopez-Restrepo, Nicolas Pinel, O.L Quintero, Arjo Segers, and A.W. Heemink. (2017). *Characterization and analysis of satellite and ground data available for the Aburrá Valley (Medellin Metropolitan Area) as inputs for air quality models*, 3th CMAS South America Conference, Vitoria-Brazil .

Other contributions

8. Lopez-Restrepo, S., **Yarce Botero, A.**, Pinel Peláez, N., Quintero, O. L., Segers, A., Heemink, A. W., *Data Assimilation as a Tool to Improve Chemical Transport Models Performance in Developing Countries*, Environmental Sustainability: Preparing for Tomorrow **18**, 99 (2021).

7. Lopez-Restrepo, S., **Yarce Botero, A.**, Pinel, N., Quintero, O. L., Segers, A., Heemink, A. W., *A Knowledge-Aided Robust Ensemble Kalman Filter Algorithm for Non-Linear and Non-Gaussian Large Systems*,Frontiers Appl. Math. Stat.8 (2022)

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

5. Lopez-Restrepo, S., **Yarce Botero, A.**, Pinel, N., Quintero, O. L., Segers, A., Heemink, A. W., *Urban Air Quality Modeling Using Low-Cost Sensor Network and Data Assimilation in the Aburrá Valley, Colombia*,Atmosphere 2021, 12(1), 91

4. Lopez-Restrepo, S., Niño Ruis, E. D., Guzman Reyes, L. **Yarce Botero, A.**, Pinel, N., Quintero, O. L., Segers, A., Heemink, A. W., *An efficient ensemble Kalman Filter implementation via shrinkage covariance matrix estimation: exploiting prior knowledge*,Computational Geosciences 25,985 (2021)

3. Hinestroza-Ramirez, J. E., Lopez-Restrepo, S., **Yarce Botero, A.**, Segers, A., Rendon-Perez, A. M., Isaza-Cadavid, S., ... Quintero, O. L. , *Improving Air Pollution Modelling in Complex Terrain with a Coupled WRF–LOTOS–EUROS Approach: A Case Study in Aburrá Valley, Colombia*,Atmosphere, 14(4), 738. (2023)

2. Hinestroza-Ramirez, J., Rengifo-Castro, J., Quintero, O., **Yarce Botero, A.**, Rendon-Perez, A,*Non-Parametric and Robust Sensitivity Analysis of the Weather Research and Forecast (WRF) Model in the Tropical Andes Region.*,Atmosphere, 14(4), 686. (2023)

1. Hinestroza-Ramirez, J., Soto Barbosa,J.E, **Yarce Botero, A.**, Suarez Higuita, D.A., Lopez Restrepo,S.,..., Quintero Montoya,O.,textbf*Evaluation of the 3DVAR Operational Implementation of the Colombian Air Force for Aircraft Operations: A Case Study*,Climate, 11(7), 153, (2023)

Software implementation

The DALo96 platform, based on the Lorenz model, serves as a key benchmark for testing data assimilation algorithms. This tool simplifies the evaluation of data assimilation techniques across different algorithms, particularly when the forcing parameter is perturbed.

Offering a wide range of data assimilation features, DALo96 is a suitable platform for students implementing software applications. The aim of DALo96 is to illuminate the core principles of various Data Assimilation (DA) methods using the Lorenz 96 model, by adjusting model parameters or assimilation technique parameters. Some of the notable DA techniques available include EnKF, EnKF Schur product covariance localization, EnKF modified Cholesky, EnKS, and EnKS modified Cholesky.

In several configurations, the forcing parameter is subject to additive noise, producing an ensemble space that provides flexibility for multiple model propagations. This parameter can be easily adjusted with a slider and perturbed using a knob. The platform also permits the selection of different data assimilation methods and the calculation of error metrics between the analysis step and the true state.

Figure 4: Data Assimilation LOrenz 96 platform DALO96 https://github.com/ayarceb/DALo96