We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,500
Open access books available

136,000
International authors and editors

170M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

BOOK CITATION INDEX
CLARIVATE ANALYTICS
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter

Data Assimilation as a Tool to Improve Chemical Transport Models Performance in Developing Countries

Santiago Lopez-Restrepo, Andrés Yarce Botero,

Olga Lucia Quintero, Nicolás Pinel, Jhon Edinson Hinestroza,

Elias David Niño-Ruiz, Jimmy Anderson Flórez,

Angela Maíra Rendón, Monica Lucia Alvarez-Laínez,

Andres Felipe Zapata-Gonzalez,

Jose Fernando Duque Trujillo, Elena Montilla, Andres Pareja,

Jean Paul Delgado, Jose Ignacio Marulanda Bernal,

Bibiana Boada, Juan Ernesto Soto, Sara Lorduy,

Jaime Andres Betancur, Arjo Segers and Arnold Heemink

Abstract

Particulate matter (PM) is one of the most problematic pollutants in urban air. The effects of PM on human health, associated especially with PM of $\leq 2.5 \mu$m in diameter, include asthma, lung cancer and cardiovascular disease. Consequently, major urban centers commonly monitor $PM_{2.5}$ as part of their air quality management strategies. The Chemical Transport models allow for a permanent monitoring and prediction of pollutant behavior for all the regions of interest, different to the sensor network where the concentration is just available in specific points. In this chapter a data assimilation system for the LOTOS-EUROS chemical transport model has been implemented to improve the simulation and forecast of Particulate Matter in a densely populated urban valley of the tropical Andes. The Aburrá Valley in Colombia was used as a case study, given data availability and current environmental issues related to population expansion. Using different experiments and observations sources, we shown how the Data Assimilation can improve the model representation of pollutants.

Keywords: chemical transport model, air quality, data assimilation, LOTOS-EUROS, low-cost networks

1. Introduction

Air pollution is defined as the presence of solid, liquid or gaseous components in the atmosphere that can cause risk and troubles for living beings or goods in general. Air pollution is one of the major environmental problem in modern human history [1]. Environmental pollution can be produced by natural or human actions. Natural sources include forest fires, volcanic emissions, dust, sand, vegetation (as pollen) and wildlife (as methane). The main human sources of air pollution are industry, power generation, transportation, deforestation and cattle raising [2].

The current exponential growth in world population heightens the importance of public health issues related to air quality [3, 4]. In developing countries, decision makers must cope with the environmental demands of expanding and overpopulated urban centers. Short term air quality forecasts and long term mitigation strategies for these centers are usually based on specialized assessments of particulate matter dynamics [5, 6]. The Aburrá Valley houses the city of Medellín and neighboring municipalities. It is the second most populous urban agglomeration in Colombia, and the third densest in the world. The valley traces the course of the Medellín River along 60 km of a deep mountain canyon that ranges in width between 3 and 10 km, and with a height difference of up to 1800 m. Air quality conditions deteriorate severely within the valley twice a year around the time of the arrival of the Intertropical Convergence Zone (March–April, and with lower intensity in October–November), when the atmospheric inversion layer persists throughout the day below the rim of the canyon, thus trapping all of the urban atmospheric contaminants within the lower atmosphere [7]. During these periods, the concentrations of particulate matter below $10\mu m$ (PM_{10}) and $2.510\mu m$ (PM2.5) remain at levels considered hazardous for vulnerable populations and even for the general population (**Figure 1**).

Due to the large stress on human health induced by this air pollution, efforts have been made to monitor, reduce, and prevent episodes in which concentrations of pollutants reach hazard levels. Before measures for reducing air pollution can be implemented it is important to know the actual concentration levels and how these evolve in time over the area of interest. This could be done using a Chemical Transport Model (CTM) to simulate concentrations of trace gasses and particulate matter [8, 9]. In the last 20 years, CTMs have seen a huge growth and development; in consequence a diversity of models exists, differing in their complexity, size of the region of study, and methods used for their development. CTMs can be broken down in four categories according to their dynamic behavior: i) Gaussian, ii)

Figure 1.
Perspective of the air quality in the city of Medellín. (August 26, 2016, www.elmundo.com).

statistic, iii) Lagrangian and iv) Eulerian [8]. Eulerian models are the most widely used and reported for monitoring and predicting the pollution behavior and define the air quality in bigger areas [9]. So, these are frequently used in areas with sizes like countries or continents and have been less used in areas like cities.

Data assimilation (DA) is a mathematical process that provides integration between measured values (observations) and a dynamic model, to improve the operation of the model. With DA, the output value provided by the model has a smaller error than the output value provided by the model without observations. DA has two key objectives: to improve the operation in predictions of model states; and estimate unknown parameters of the model [10]. DA has been tested in different science fields such as oceanography, climatology, CTMs, and reservoirs characterization [11]. DA allows integrating models and observations out different scales of size and temporal sampling [12]. When two sources of information are combined, DA assumes that both the model and the measurements are subject to errors. These errors are impossible to know with accuracy and need to be specified in statistical and probabilistic terms. DA is not only looking to reduce the model error in space or time with observations; its mission is to digest the observation based on the laws given by the model and to determine the dynamic evolution of the model state that represents better measurements [13].

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

Public air quality monitoring networks often consist of fixed measuring stations equipped with expensive sensors and maintained under rigorous operational and calibration regimes in order to provide high quality data. The high costs associated with establishing and maintaining such stations means that not all cities in developing countries can afford monitoring networks of sufficient spatial coverage [14]. Even in large cities in developed countries, the official air quality monitoring networks do not always provide information at the spatial and temporal resolution required to assess the impact of pollution sources on health [15], as the cost of the equipment makes the necessary density prohibitive. In the metropolitan region of Medellín (Colombia) and its con-urban municipalities for example, there are 21 main PM2.5 monitoring stations, at an average density of 8.25 km2 over the entire area of the 10 municipalities. This has motivated the expansion and improvement of low-cost systems and programs to measure PM [16]. The limited number of studies that have evaluated newer generations of low-cost $PM_{2.5}$ sensors have shown that the most widely used low-cost sensors attain high accuracy when compared to standard monitoring stations (R^2 value ranging from 0.93 to 0.95) [17]. The data provided by these sensors can complement those generated by conventional systems, increasing the data resolution and allowing studies of exposure at the human

level [15, 18]. By data assimilation, the incorporation of air pollution data into CTM increases the ability to grasp local and regional patterns and fill spatial coverage gaps. Additionally, the combination of different sources of information and knowledge (data and model) increases the robustness and reliability of low-cost observations [12, 19].
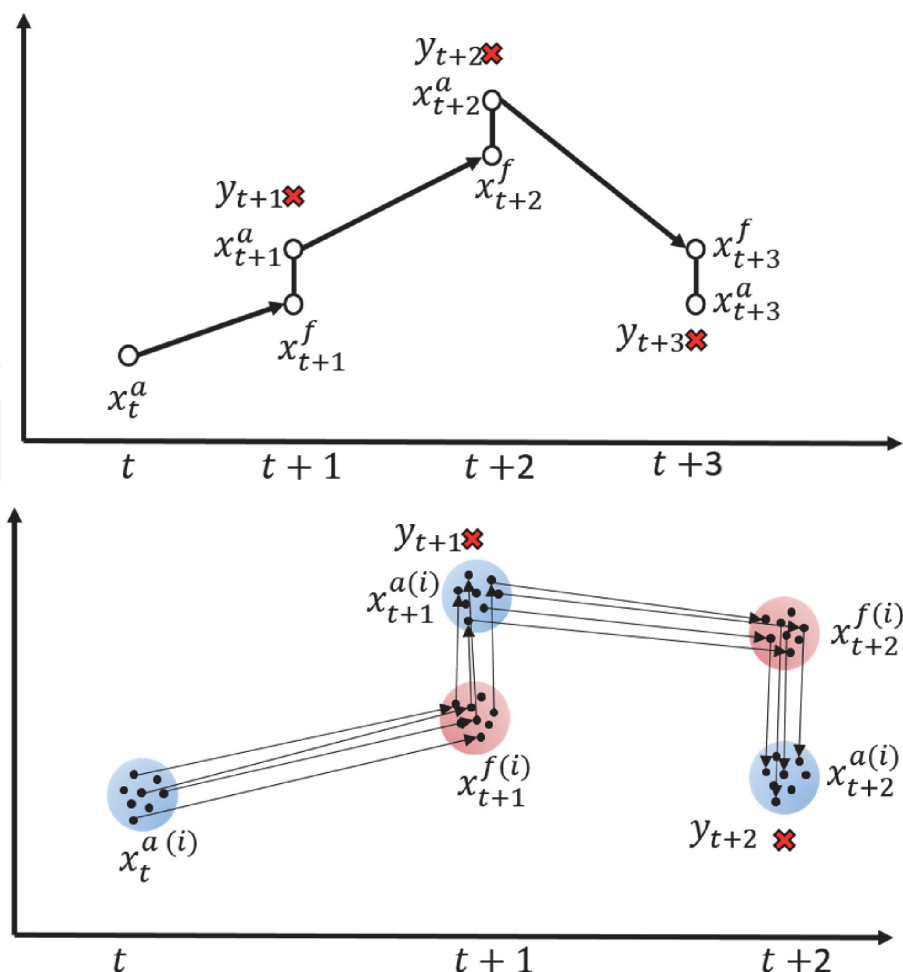
2. The ensemble Kalman filter

The Ensemble-Based DA is a family of methods that uses an ensemble to model the statistics of the first guess (background). In each assimilation step, a forecast from the previous model simulation is used as a first guess, using the available observation this forecast is modified in better agreement with these observations. Due to it is easily implemented, it is relatively low in computational costs (compared with other DA techniques), and has a very general statistical formulation it is one of the most widely used approaches for tackling real-time forecasting problems [20].

The Ensemble Kalman filter (EnKF) is the main Ensemble-based DA method [21]. Based on the Kalman Filter (KF) [22], EnKF is an alternative for nonlinear, high-dimensional systems. EnKF essentially is a Monte Carlo Ensemble-based method, based on the representation of the probability density of the state by an ensemble of N model realizations. Each ensemble member is assumed to be a single sample out of a distribution of the true state [23]. In the first step, a Monte Carlo ensemble of the initial condition is generated to represent the uncertainty in the initial condition. After that, and in the same way that the KF, the EnKF propagates each ensemble using the state-space operator, this step is called forecast step. When observations are available, the EnKF uses them to update each forecast ensemble members and obtain the analysis ensemble, this step is named analysis step. The update is proportional to the differences between the observations and the model outputs, by a gain called Kalman Gain. **Figure 2** shows a graphic representation and a comparison between the KF and EnKF.

3. Forecasting PM10. And PM2.5. in the Aburrá Valley (Medellín, Colombia) via EnKF based data assimilation

Understanding local and regional atmospheric particulate matter transport patterns becomes a top priority for urban valleys in the northern Andes. This work will help establish accurate air quality forecasting systems for the Aburr'a Valley (and other similar areas) and improve decision-making. Chemical Transport Models (CTM) are valuable resources for understanding atmospheric pollutants' dynamics and have thus been widely used in air quality monitoring [8, 9].

Here we use simulations of the LOTOS-EUROS (LE) chemistry transport model (CTM) to investigate the atmospheric contaminant dynamics in the Aburr'a valley, which spans ten municipalities, including Medellín city. The *Sistema de Alerta Temprana del Valle de Aburrá* (SIATA), a ground-based sensor network with stations throughout the valley, can provide particulate material observations. A preliminary exercise is carried out to assimilate these findings into the simulations and assess the system's forecast capacity. Due to the various sources of uncertainty present, this implementation poses a challenge from a scientific standpoint. The topography and scale of the valley and the physical conditions of the area of interest necessitate an extra effort to perform a regional high-resolution model simulation.

Figure 2.
Representation of Kalman filter (upper) and ensemble Kalman filter (lower).

Model inputs (emission inventory and meteorology) are not readily available with the desired resolution and accuracy, which adds to the experiment's uncertainty.

3.1 Material and methods

A data assimilation method for the LOTOS-EUROS chemical transport model has been introduced to boost the PM_{10} and $PM_{2.5}$ forecasts. The system uses an Ensemble Kalman filter with covariance localization, which is based on the specification of emissions uncertainties. The data was gathered from a surface network for the months of March and April 2016, during one of the region's worst air quality crises in recent memory. The SIATA is spread around the five most populous municipalities in the Aburrá Valley, with the bulk of the measuring stations in Medellín. **Figure 3** represents the distribution of observation sites.

Measurements for one station for each species (represented with a star in **Figure 3**) were used for validation, taking two stations with a considerable distance between them to obtain a acceptable spatial representation.

In a first series of experiments, the spatial length scale of the covariance localization and the temporal length scale of the stochastic model for the emission uncertainty were calibrated to optimize the assimilation system. The calibrated system was then used in a series of assimilation experiments. The summarized experimental setup is presented in the **Figure 4**.

Simulations were conducted with the LE model, adopting a nested domain configuration as depicted in **Figure 5** and detailed in **Table 1**. The data sets used in the model are summarized in **Table 2**.
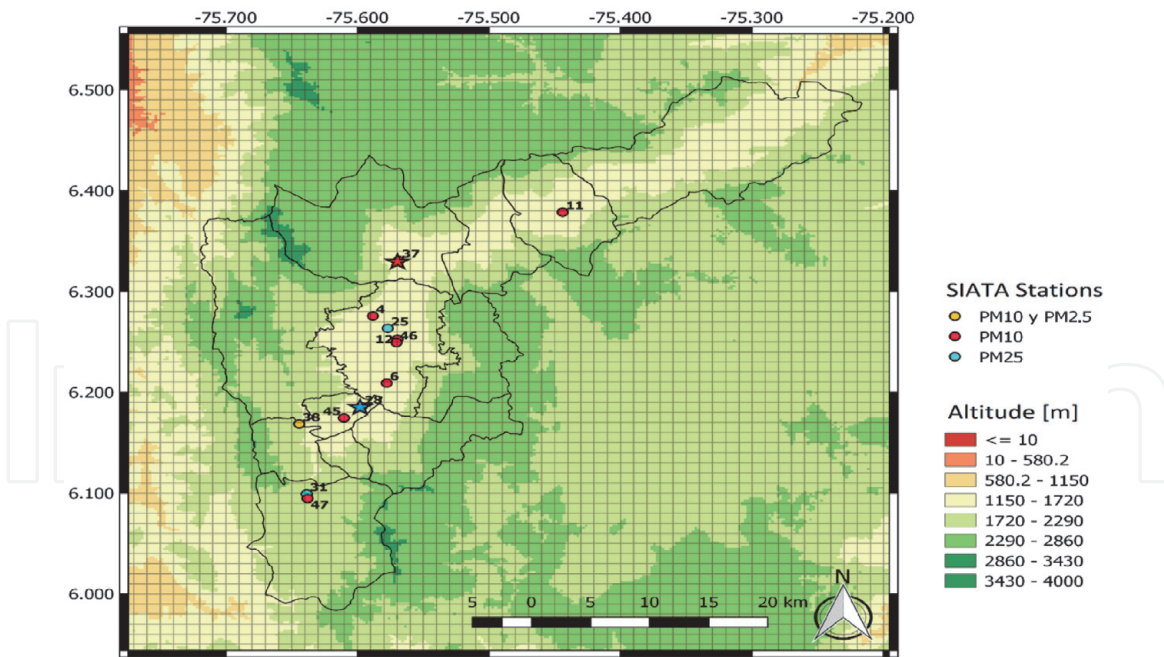
Figure 3.
SIATA sensor network for PM_{10} and $PM_{2.5}$. The stars represent observation points for validation and the circles represent observations points for assimilation. Taken from [24].
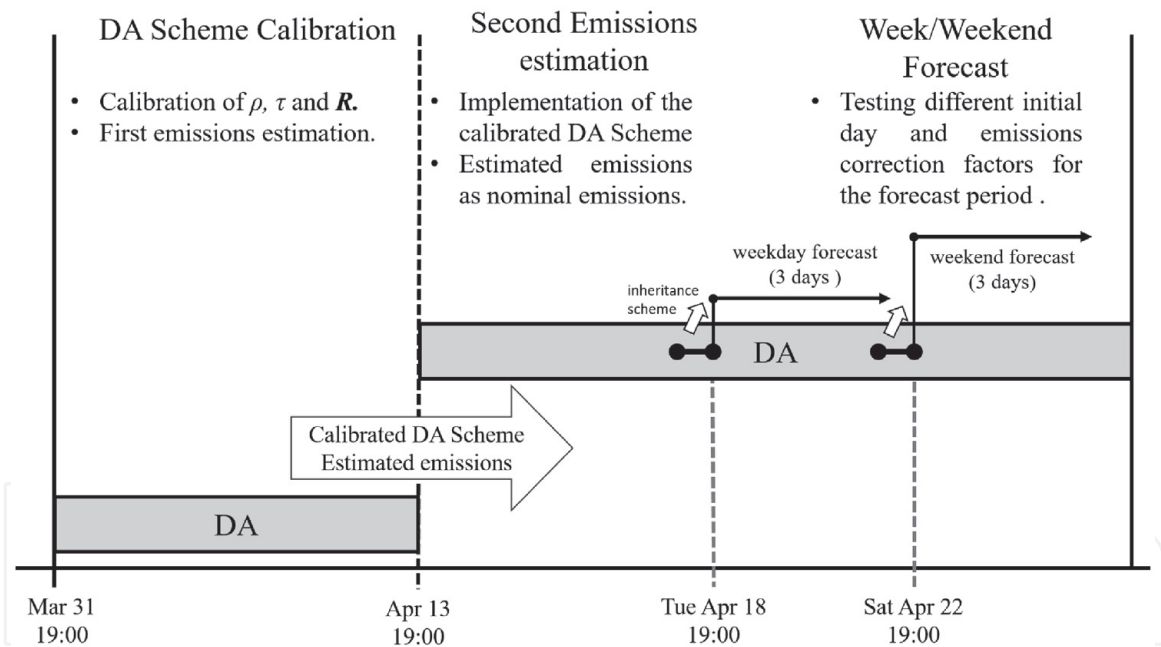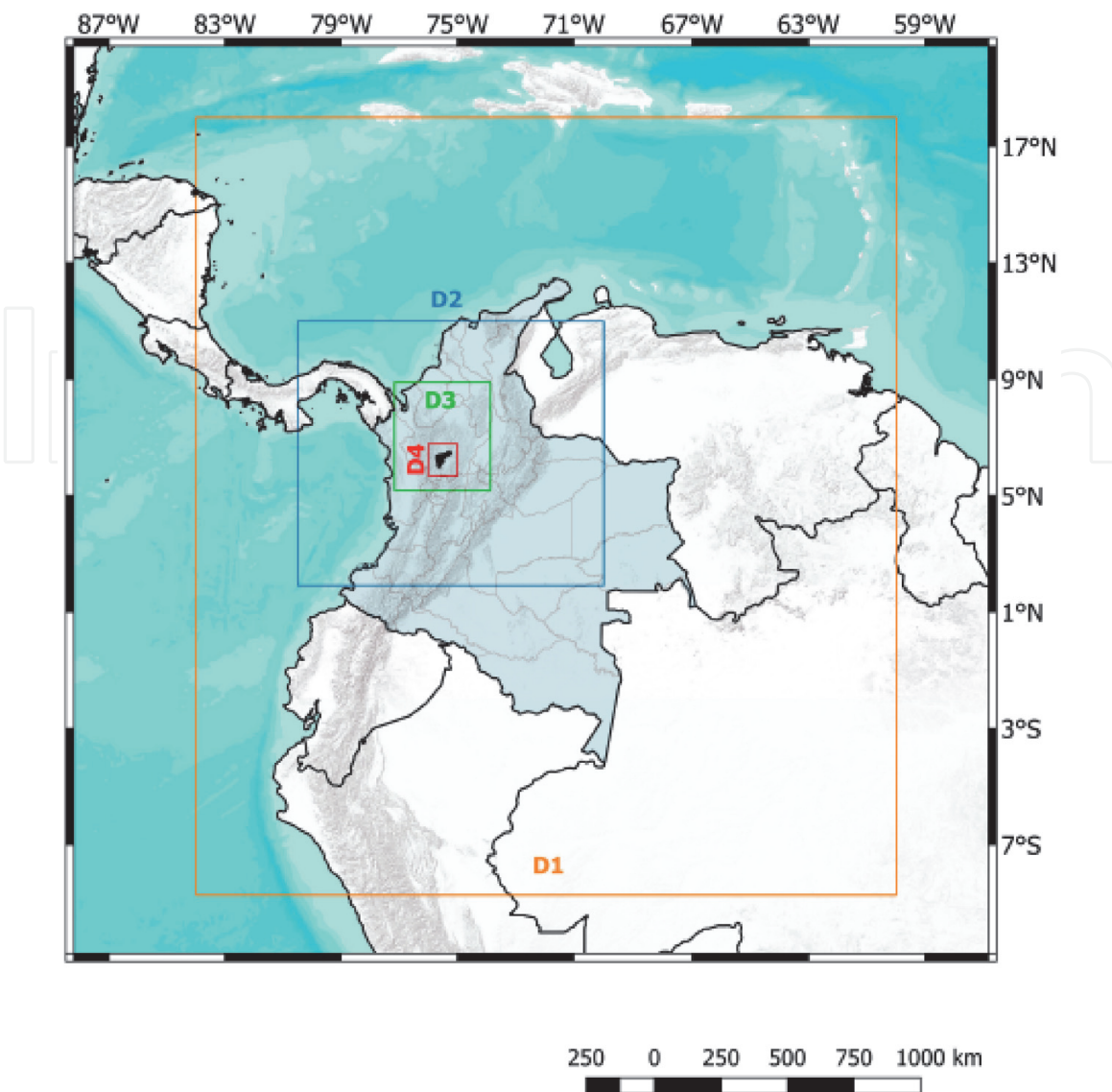

Figure 4.
Graphic representation of the experimental setup. Taken from [24].

3.2 Results

Estimated PM_{10} emissions and EDGAR nominal emissions are shown in **Figure 6**. The emissions hot-spots occur in rural zones with limited human activity in the EDGAR database. The estimated emissions attempt to remedy this behavior by projecting the most of the pollution into the metropolitan region of the valley (**Figure 6**).

The assimilated PM_{10} concentration match closely those measurements at the Universidad San Buenaventura (center of the valley) from April at 19:00 UTC-5 through April 25 at 11:00 UTC-5 (see **Figure 7**). The peak around 18:00 (and usually all day up to that hour) may be unreliable, which may be because of EDGAR's

Figure 5.
Four nested domains for metropolitan area of Aburrá Valley assesment. Taken from [24].

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

Table 1.
Nested domain specifications.

temporal emissions factors. Additionally, concentrations can be increased by the meteorological fields. Note that the daily cycle for the assimilated model remains closer to the observations than the model without assimilation.

Figure 8 shows a similar comparison for the $PM_{2.5}$ station. The model in a free run tends to over estimate the $PM_{2.5}$ concentrations (see peaks in 15 April at 23:00 UTC-5, 24 April at 22:00 and 25 April at 23:00 UTC-5). The results of the assimilation process offer a better average estimation. The daily cycle of $PM_{2.5}$ within the Aburrá valley is related to the industrial and mobile sources emissions profile and the meteorological conditions inside the valley.

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

**Table 2.**
Data set used in the D4 domain.
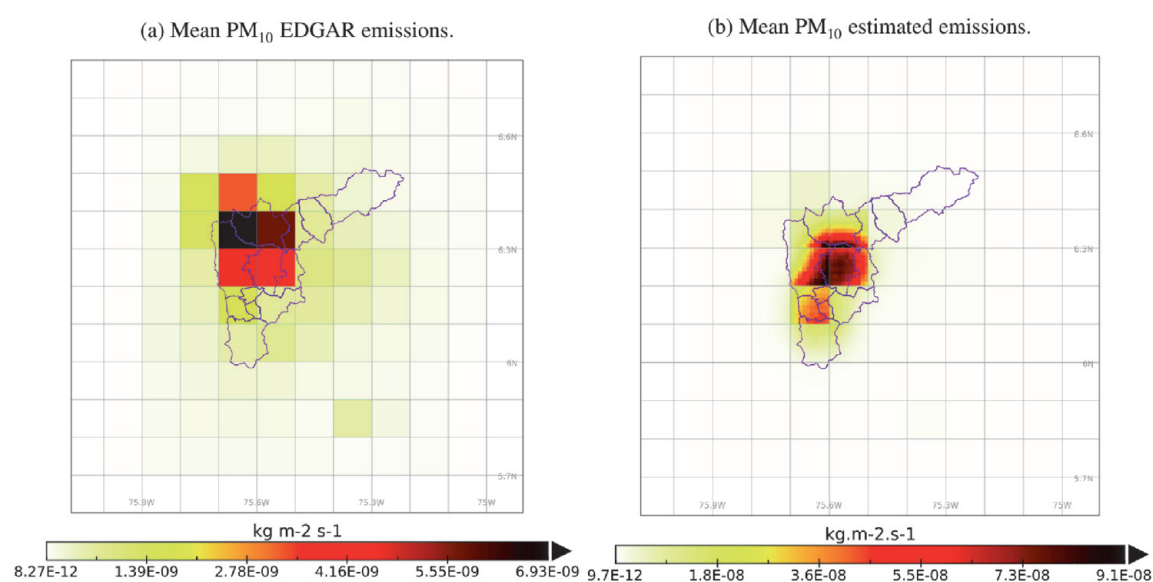
Figure 6.
Comparison between EDGAR PM_{10} and estimated PM_{10} emissions. Taken from [24].

3.3 Conclusions

Poor air quality is a current environmental problem in several Colombian cities. To be prepared for air quality degradation requires accurate and reliable data for decision-making in South America. This study shows that the LOTOS-EUROS model can function in areas with more complex topography, such as the Abura Valley, and encourages the development of fine-tuned weather forecasting systems to support the target. The use of regional, ground-based pollutant data from the SIATA sensor network, in the assimilation of the LOTOS-EUROS model, enhanced the PM_{10} and $PM_{2.5}$ representation.

4. Urban air quality modeling using low-cost sensor network and data assimilation in the Aburrá Valley, Colombia

Public air quality monitoring networks frequently consist of fixed measuring stations equipped with expensive sensors and maintained under strict operational and calibration regimes. Because of the high costs of setting up and maintaining
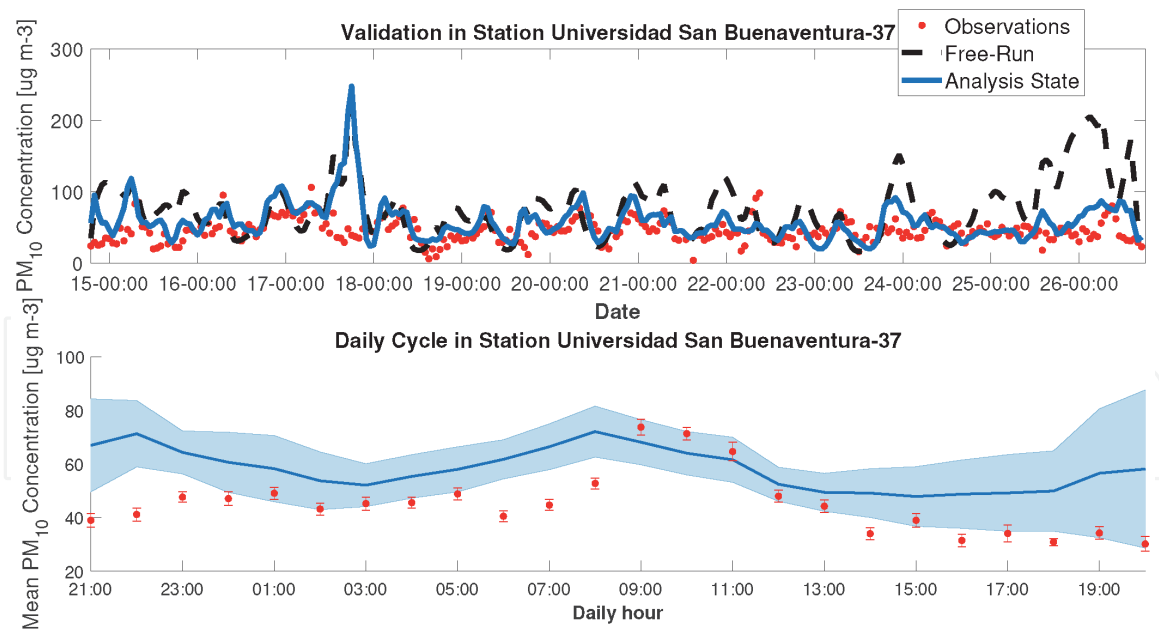
Figure 7.
PM_{10} validation for the second DA iteration. Estimated emissions were used as nominal emissions, the estimated observation error covariance is used in the assimilation step. Red points are observations, solid black line is the free run model and the solid blue line is the analysis step for the assimilated model. The diurnal cycles were obtained from 13 samples for each hour. The shadows and the bars represent the standard deviation of the 13 samples. The time axis corresponds with the local time zone UTC-5. Taken from [24].
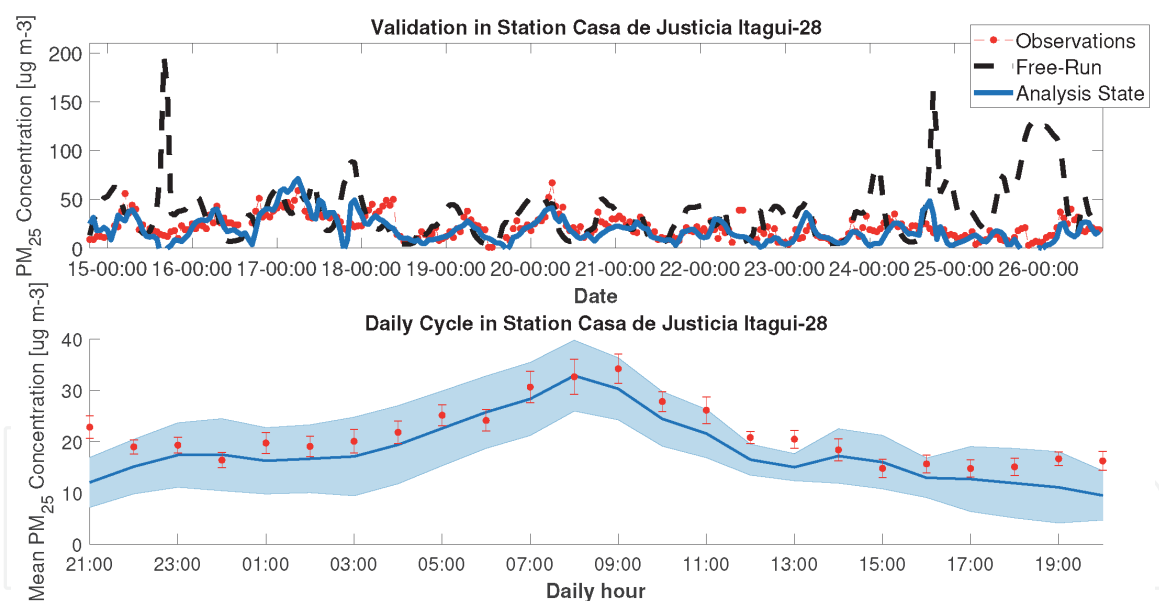
Figure 8.
$PM_{2.5}$ validation for the second DA iteration. Estimated emissions were used as nominal emissions, the estimated observation error covariance was used in the assimilation step. Red points are observations, solid black line the free run model and solid blue line the analysis step for the assimilated model. The diurnal cycles were obtained from 13 samples for each hour. The shadows and the bars represent the standard deviation of the 13 samples. The time axis corresponds with the local time zone UTC-5. Taken from [24].

such stations, not all cities in developing countries can afford monitoring networks with sufficient spatial coverage [14]. Even in developed cities, official air quality monitoring networks do not always provide information at the spatial and temporal resolution required to assess the impact of pollution sources on health, [15], due to the equipment's high cost. This has prompted the development and improvement of low-cost PM measurement systems and programs. According to [17], a small number of studies evaluating newer generations of low-cost PM2.5 sensors have found

that the most widely used low-cost sensors achieve high accuracy when compared to standard monitoring stations (R^2 values ranging from 0.93 to 0.95). The data collected by these sensors can be used to supplement that collected by traditional systems, increasing data resolution and allowing studies of human exposure [15, 18].

Using techniques like data fusion or data assimilation to integrate observations from dense networks of low-cost sensors into mathematical models allows for a spatially continuous representation of concentration fields with significantly reduced bias citeLahoz2014. By spatially interpolating between monitoring locations and constraining the model with observations, these techniques add value to the sensor observations while also adding value to the model [17, 18, 25]. Both sources of information can thus be combined in a mathematically objective manner to reduce the uncertainty inherent in both sources [12]. Although data assimilation is a more complex family of methods than data fusion or interpolation techniques, it is the most versatile and robust of these approaches. The goal of evaluating the data from the low-cost sensor network as an alternative to monitoring PM2.5 concentrations in developing countries is to see if it is viable.

4.1 Material and methods

The SIATA project operates the official high-end air quality monitoring network (henceforth *official network*, and a hyper-dense, low-cost air quality network developed within the Citizen Scientist program (henceforth *low-cost network*).

The low-cost network was created with the aim of engaging the community in issues surrounding air quality, and as an extension of the official network. The low-cost network consists of 255 real-time $PM_{2.5}$ (**Figure 9**, panel b).The measuring equipment was developed by SIATA based on the well-known low-cost Shinyei PPD42NS, NOVA SDS011, and Bjhike HK-A5 sensors [27]. Each low-cost sensor is calibrated individually against BAM-1020 measurements [27]. The calibration process showed the measurements of 91% of the low-cost sensors with correlation values above 0.6 against the official measurements, and 67% with values above 0.8.
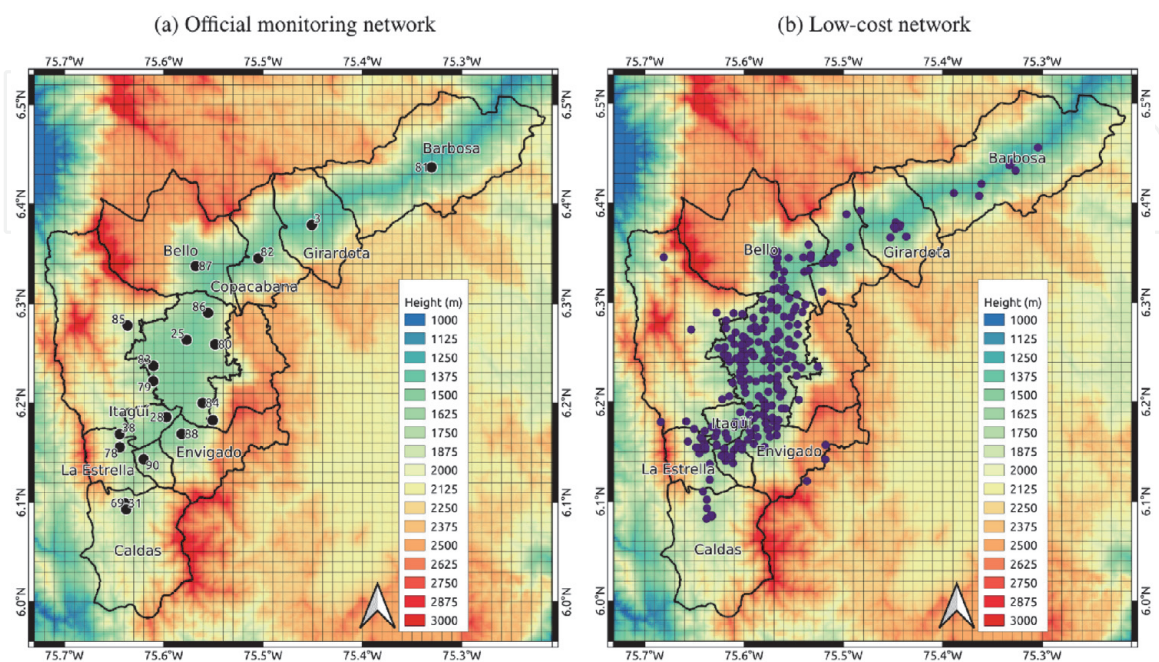
Figure 9.
Spatial distribution of the hyper-dense low-cost network citizen scientist and official monitoring air-quality network for PM$_{2.5}$. The gray raster represent the LOTOS-EUROS model grid. Taken from [26].
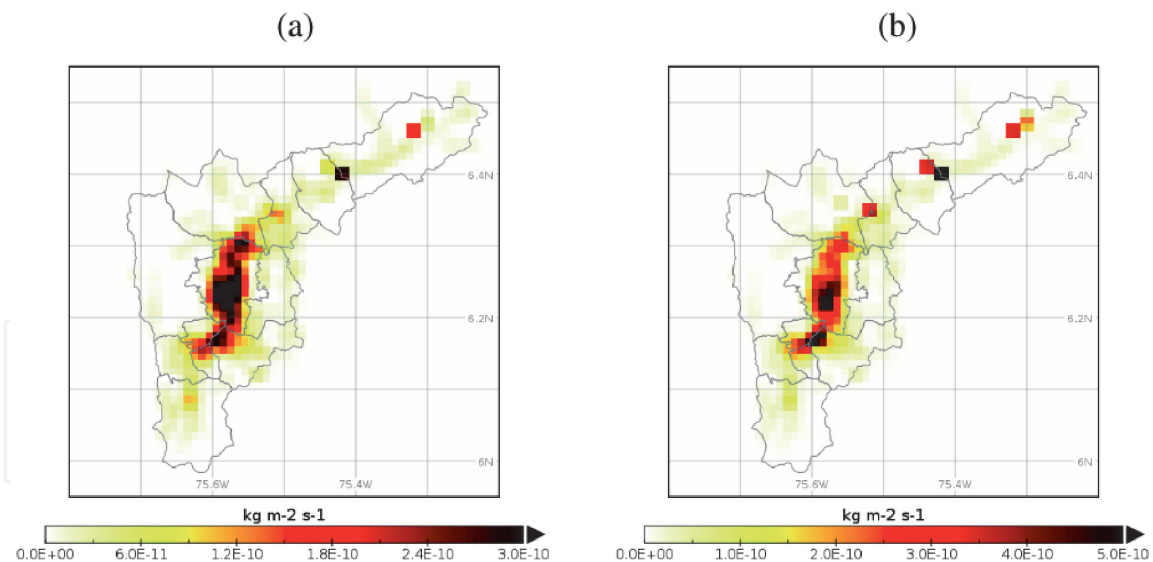
Figure 10.
Local particulate matter emission inventories for the Aburrá Valley: (a) PM$_{2.5}$, and (b) PM$_{10}$. The values correspond with the estimated annual emissions. Taken from [26].

The median of the root mean square error showed a value of 6.2 $\mu g/m^3$, with a tendency to decrease for higher concentrations [27]. The low-cost network thus represents satisfactorily the dynamics of PM$_{2.5}$ concentrations in the Valley's atmosphere.

An anthropogenic urban emissions inventory for 2016 specific to Medellín and the other nine municipalities of the Aburrá Valley was used for the simulations on the D4 domain. The construction of the inventory followed a bottom-up methodology, combining activity data (traffic intensities, industrial production) with emission factors. Only traffic and industrial point sources were considered, without accounting for neither household nor commercial emissions [28].

The emission inventory was disaggregated over the Aburrá Valley (76°W-75°W and 5.7°N-6.8°N) at a resolution of 0.01° × 0.01° (approximately 1 km × 1 km), using a method based on road density as in [29]. The road network map was obtained from the OpenStreetMap database [30], and simplified by removing segments classified as residential, as recommended in [31, 32]. The simplification of the road network can reduce errors in the spatial disaggregation since residential roads correspond to a high portion of the road network length but carry a low percentage of total vehicular traffic. The point-source emissions were distributed on the grid using their known location [28]. **Figure 10** shows the resulting emissions maps for PM$_{2.5}$ and PM$_{10}$.

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

We performed four different LOTOS-EUROS simulations:

1. a LOTOS-EUROS model simulation without data assimilation (henceforth *LE*);

2. a simulation with assimilation of data (observations) from the 14 stations of the official network (henceforth *LE-official*;

3. a simulation with assimilation of the data from the entire low-cost network (henceforth *LE-lowcost*)

4. a simulation with assimilation only of high-quality data from the low-cost network (henceforth *LE-lowcost-HQ*).

4.2 Results

The concentration fields were evaluated using seven of the official monitoring stations (*validation stations*. **Figure 11** shows the temporal series for the simulated and observed $PM_{2.5}$ concentrations at four of the validation stations. The four selected stations represent downtown Medellín (station 25), residential areas (station 86), areas with high vehicular flow (station 88), and a peri-urban area in the outskirts of the city (station 85). Those stations summarize the behavior of all seven validation stations. The LE simulation consistently underestimated the concentrations observed at stations 85 and 88. At stations 25 and 86, the LE simulation results were close in magnitude between February 24 and March 3 and March 10 to March 15; between March 3 and March 10, the model presented values much lower than those observed. The day-to-day variability was reduced for this same period, as seen in stations 85 and 86. This inconsistent behavior suggests a poor representation of the meteorological dynamics that govern the dispersion and accumulation of $PM_{2.5}$ within the valley. Simulations using data assimilation showed noisier behaviors than the LE simulation. This process is commonly observed when applying the EnKF and obeys the stochastic nature and the handling of uncertainty inherent to the method [21]. However, those simulations managed to correct the large discrepancies present
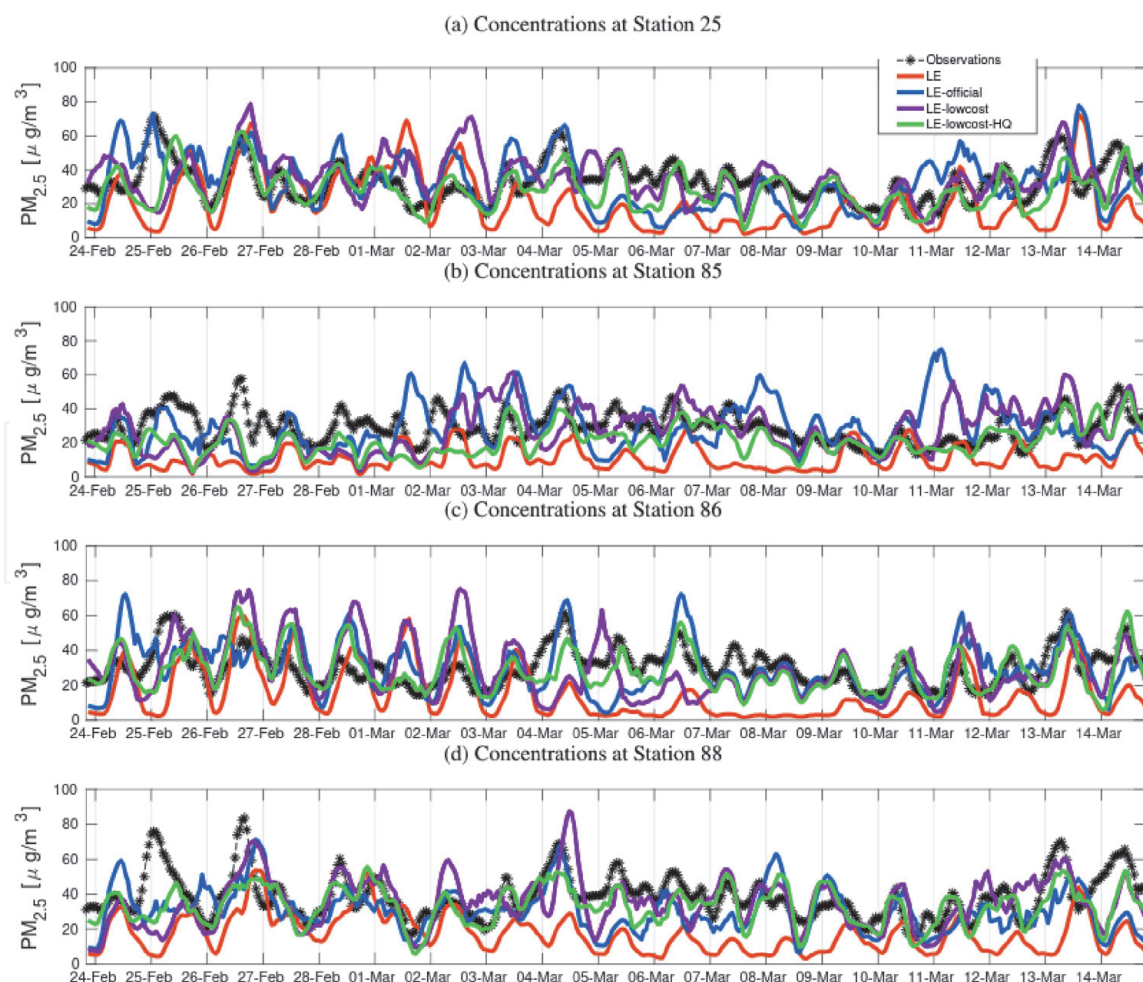
Figure 11.
Temporal series of $PM_{2.5}$ concentrations from selected validation stations of the official network, LOTOS-EUROS without assimilation, LE-official, LE-lowcost and LE-lowcost-HQ. Time stamps are valid for local time (UTC-5). A spin-up of 5 previous days was taken for each simulation. Taken from [26].

in the LE simulation. Both LE-official, LE-lowcost, and LE-lowcost-HQ represented more accurately the day-to-day variability of the observations than LE. In general terms, there was no evidence of a sizeable and persistent difference among the simulations with data assimilation throughout the entire period. Nevertheless, the LE-lowcost-HQ simulation reproduced with greater accuracy the concentrations observed in different periods, such as between February 26 and March 4 in station 25, between March 9 and March 14 in stations 85 and 86.

Figure 12 shows the diurnal cycles during the simulation period in the four selected validations stations. The diurnal cycle of the LE simulation differed from the observations in both magnitude and temporal behavior. The highest concentration peak that appears around 09:00 in all the stations is mainly due to traffic dynamics. In stations 25 and 88, the LE morning peak corresponded in time but not in magnitude with the observations; in stations 85 and 86, said peak appeared later in the simulations than in the observations. This time lag suggests a poor spatial representation of mobile emissions by the emissions inventory; or a deficiency it the wind fields in reproducing the valley dynamics, showing a late transport of the particulate material to these areas. The LE simulation did not capture the evening peak shown by the observations around 21:00 hours. The simulations using data assimilation presented diurnal cycles closer to the observations than did the LE simulation. The LE-official simulation captured the time and magnitude of the morning peak in stations 85 and 86. In station 88, LE-official corrected the time lag in the morning peak seen in LE, and improved the estimated magnitudes albeit still falling short of the observed values. A different behavior was seen for station 25,
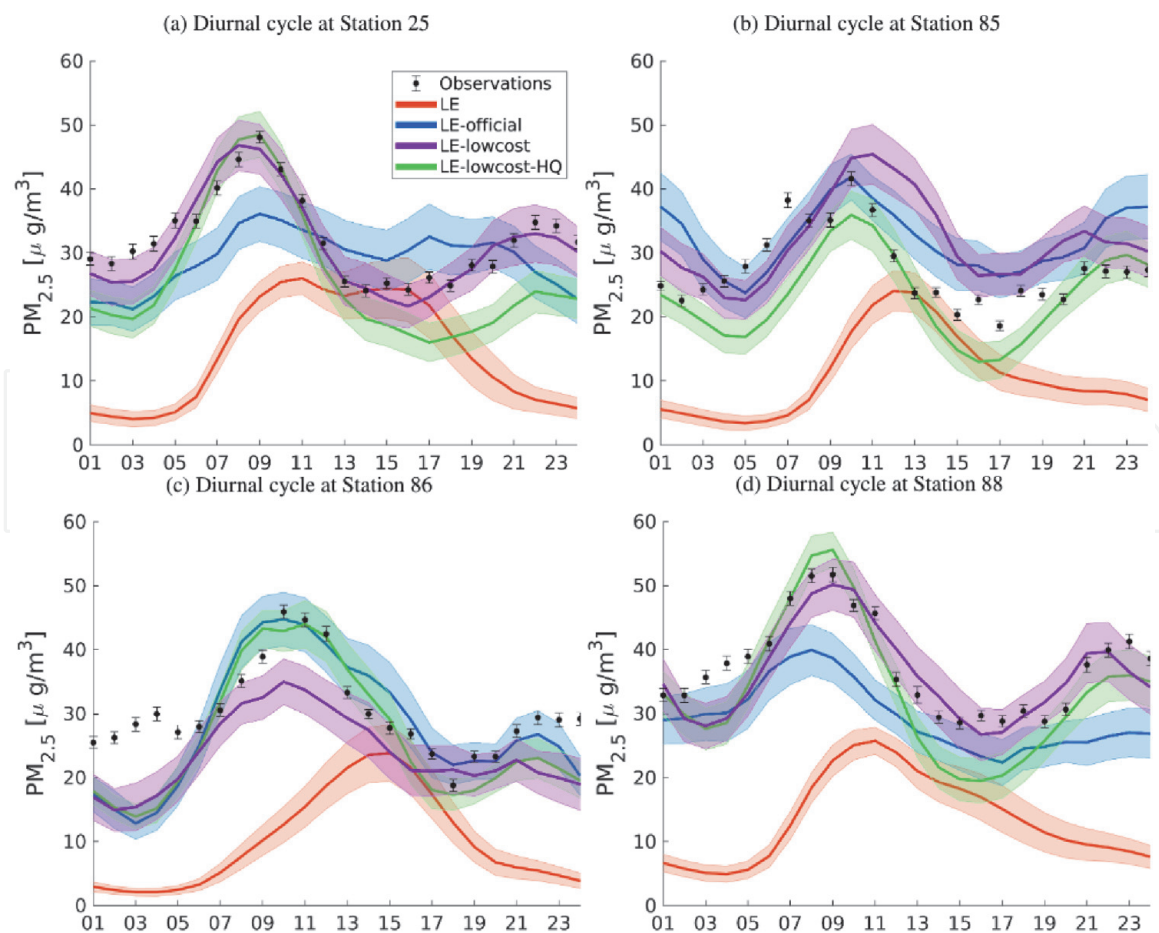
Figure 12.
Diurnal cycle of PM$_{2.5}$ concentrations from selection stations of the official network, LOTOS-EUROS without assimilation, LE-official, LE-lowcost and LE-lowcost-HQ. The bars and the shadows represent the standard deviation over the simulation period. The time stamps are valid for local time (UTC-5). Taken from [26].

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

Table 3.
Mean fractional bias, root mean square error and Pearson correlation coefficient for simulated $PM_{2.5}$. Values are averaged over all the validation stations for the simulation period.

where LE-official had low diurnal variability, with a slight underestimation in the morning, and an overestimation in the afternoon. The LE-lowcost and LE-lowcost-HQ simulations results resembled closely the diurnal behavior of the observations, especially the temporal component. In all the stations, both the morning and the evening peaks matched the observations. The observed concentrations for stations 25 and 88 fell inside the standard deviation range for the LE-lowcost simulation; the same simulation overestimated the concentrations between 11:00 and 19:00 for station 85, and underestimated the concentrations between 01:00 and 13:00 for station 86. The LE-lowcost-HQ simulation results were overall the closest to observations.

The averaged evaluation statistics among all the validation station are shown in **Table 3**. The simulation results without data assimilation (LE) underestimated the observed concentrations in all the validation stations. This was also seen in previous related works [24, 33]. The RMSE value reflected a low correspondence between the observed and simulated concentrations when using the model without data assimilation. The correlation coefficient was low, meaning that the model was not able to capture the variations in diurnal and day-to-day concentrations. In contrast, the three simulations using data assimilation had MFB values close to 0, without a significant difference among them. The data assimilation was thus effective in reducing between the model and reality. The RMSE also improved when using data assimilation, decreasing by 24.4% in the LE-official, 32.8% in the LE-lowcost, and 36.2% in the LE-lowcost-HQ simulations relative to the RMSE of the LE simulation. The *R* values were all above the criteria of good performance according with [34] **Table 2**, and based in [35, 36]. Assimilation of either data set from the low-cost network resulted in improved error statistics when compared to the LE-official simulation.

4.3 Conclusions

We present a data assimilation application of a hyper-dense low-cost PM network and the chemical transport model LOTOS-EUROS in a urban setting. The low-cost network provided high quality data comparable to those provided by the official monitoring network. The performance of the model with assimilation of the spatially-dense data from the low-cost network improved both in terms of its representation of the observed dynamics, as well as in its forecast capabilities, highlighting its value as an air-quality management tool. Our results support the idea than with the current advances in the low-cost sensors, it is possible to use low-cost networks and data assimilation to model and predict air quality in urban areas.

Jointly with previous work [15, 18, 25, 37–39], our results can support and motivate the development of future low-cost networks and their integration in data fusion applications. According to the literature, North America, Europe, and China

concentrate most of the current low-cost implementations, with experimental, citizen, and data dissemination purposes [14, 40]. In developing countries, a low-cost network, together with a CTM and data assimilation can provide a valuable first approach to monitoring PM without the high cost of an official air quality network.

Although one of the main advantages of a low-cost networks is the possibility of implemented hyper-dense networks with relative low costs, it is recommended to prioritize in the quality of the data (sensor quality, calibration, maintenance) and the study of optimal localization. High quality and the correct number and localization of sensors improve the data assimilation process and minimizes operational and computational costs.

Acknowledgements

Conflict of interest

The authors declare no conflict of interest.

Author details

Santiago Lopez-Restrepo[1,2,3]*, Andrés Yarce Botero[1,2,3], Olga Lucia Quintero[2], Nicolás Pinel[4], Jhon Edinson Hinestroza[1], Elias David Niño-Ruiz[5], Jimmy Anderson Flórez[6], Angela Maíra Rendón[7], Monica Lucia Alvarez-Laínez[8], Andres Felipe Zapata-Gonzalez[8], Jose Fernando Duque Trujillo[9], Elena Montilla[10], Andres Pareja[11], Jean Paul Delgado[12], Jose Ignacio Marulanda Bernal[13], Bibiana Boada[1], Juan Ernesto Soto[6], Sara Lorduy[6], Jaime Andres Betancur[12], Arjo Segers[14] and Arnold Heemink[2]

1 Mathematical Modelling Research Group, Universidad EAFIT, Medellín, Colombia

2 Department of Applied Mathematics, Delft University of Technology, The Netherlands

3 SimpleSpace, Medelliín, Colombia

4 Department of Biological Sciences, Research Group on Biodiversity, Evolution and Conservation at Universidad EAFIT, Medellín, Colombia

5 Applied Math and Computer Science Laboratory, Department of Computer Science, Universidad del Norte, Colombia

6 Comando Aereo de Combate N 5 CACOM 5, Centro Tecnologico Aeroespacial para la Defensa CETAD, Fuerza Aérea Colombiana, Rionegro, Colombia

7 Grupo GIGA, Universidad de Antioquia, Medellín, Colombia

8 GRID Nano-Fiber Group, Universidad EAFIT, Medellín, Colombia

9 Environmental Magnetism, Universidad EAFIT, Medellín, Colombia

10 Applied Optics Research Group, Universidad EAFIT, Medellín, Colombia

11 Unidad de Toxicidad in vitro, Universidad CES, Sabaneta, Colombia

12 Grupo, genérica, regeneración y cáncer, Universidad de Antioquia, Medellín, Colombia

13 GEMA, Universidad EAFIT, Medellín, Colombia

14 Department of Climate, Air and Sustainability, TNO, The Netherlands

*Address all correspondence to: slopezr2@eafit.edu.co

References

[1] J. Green and S. Sánchez, "Air Quality in Latin America: An Overview," tech. rep., Clean air Institute, Washington D. C., USA, 2012.

[2] C. Borrego, M. Coutinho, a. M. Costa, J. Ginja, C. Ribeiro, a. Monteiro, I. Ribeiro, J. Valente, J. H. Amorim, H. Martins, D. Lopes, and a. I. Miranda, "Challenges for a New Air Quality Directive: The role of monitoring and modeling techniques," *Urban Climate*, vol. 14, pp. 328–341, 2015.

[3] H. Akimoto, "Global air quality and pollution," *Science*, vol. 302, no. 5651, pp. 1716–1719, 2003.

[4] B. Gurjar, T. Butler, M. Lawrence, and J. Lelieveld, "Evaluation of emissions and air quality in megacities," *Atmospheric Environment*, vol. 42, no. 7, pp. 1593–1606, 2008.

[5] M. L. Bell, L. A. Cifuentes, D. L. Davis, E. Cushing, A. G. Telles, and N. Gouveia, "Environmental health indicators and a case study of air pollution in latin american cities," *Environmental Research*, vol. 111, no. 1, pp. 57–66, 2011.

[6] J. F. Sallis, F. Bull, R. Burdett, L. D. Frank, P. Griffiths, B. Giles-Corti, and M. Stevenson, "Use of science to guide city planning policy and practice: how to achieve healthy and sustainable future cities," *The Lancet*, vol. 388, no. 10062, pp. 2936–2947, 2016.

[7] J. F. Jiménez, *Altura de la Capa de Mezcla en un área urbana montañosa y tropical. Caso de estudio: Valle de Aburrá (Colombia)*. Doctoral thesis, Universidad de Antioquia, Medellín, 2016.

[8] P. Thunis, *A. Miranda*, J. M. Baldasano, N. Blond, J. Douros, A. Graff, S. Janssen, K. Juda-Rezler, N. Karvosenoja, G. Maffeis, A. Martilli, M. Rasoloharimahefa, E. Real, P. Viaene, M. Volta, and L. White, "Overview of current regional and local scale air quality modeling practices: Assessment and planning tools in the EU," *Environmental Science & Policy*, vol. 65, pp. 13–21, 2016.

[9] M. Lateb, R. Meroney, M. Yataghene, H. Fellouah, F. Saleh, and M. Boufadel, "On the use of numerical modelling for near-field pollutant dispersion in urban environments: A review," *Environmental Pollution*, vol. 208, pp. 271–283, 2016.

[10] M. Berardi, A. Andrisani, *L. Lopez*, and M. Vurro, "A new data assimilation technique based on ensemble Kalman filter and Brownian bridges: An application to Richards' equation," *Computer Physics Communications*, vol. 208, pp. 43–53, 2016.

[11] M. Van Loon, P. J. H. Builtjes, and a. J. Segers, "Data assimilation of ozone in the atmospheric transport chemistry model LOTOS," *Environmental Modelling and Software*, vol. 15, no. 6–7 SPEC. ISS, pp. 603–609, 2000.

[12] W. A. Lahoz and P. Schneider, "Data assimilation: Making sense of Earth Observation," *Frontiers in Environmental Science*, vol. 2, no. MAY, pp. 1–28, 2014.

[13] M. Bocquet, H. Elbern, H. Eskes, M. Hirtl, R. Aabkar, G. R. Carmichael, J. Flemming, A. Inness, M. Pagowski, J. L. Pérez Camaño, P. E. Saide, R. San Jose, M. Sofiev, J. Vira, A. Baklanov, C. Carnevale, G. Grell, and C. Seigneur, "Data assimilation in atmospheric chemistry models: Current status and future prospects for coupled chemistry meteorology models," *Atmospheric Chemistry and Physics*, vol. 15, pp. 5325–5358, may 2015.

[14] A. Kumar and B. R. Gurjar, "Low-Cost Sensors for Air Quality Monitoring

in Developing Countries -A Critical View," *Asian Journal of Water, Environment and Pollution*, vol. 16, no. 2, pp. 65–70, 2019.

[15] F. E. Ahangar, F. R. Freedman, and A. Venkatram, "Using low-cost air quality sensor networks to improve the spatial and temporal resolution of concentration maps," *International Journal of Environmental Research and Public Health*, vol. 16, no. 7, 2019.

[16] 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*, vol. 75, pp. 199–205, 2015.

[17] H. Y. Liu, P. Schneider, R. Haugen, and M. Vogt, "Performance assessment of a low-cost PM 2.5 sensor for a near four-month period in Oslo, Norway," *Atmosphere*, vol. 10, no. 2, 2019.

[18] P. Schneider, N. Castell, M. Vogt, F. R. Dauge, W. A. Lahoz, and A. Bartonova, "Mapping urban air quality in near real-time using observations from low-cost sensors and model information," *Environment International*, vol. 106, no. June, pp. 234–247, 2017.

[19] N. Castell, F. R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, and A. Bartonova, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?," *Environment International*, vol. 99, pp. 293–302, 2017.

[20] G. Fu, F. Prata, H. Xiang Lin, A. Heemink, A. Segers, and S. Lu, "Data assimilation for volcanic ash plumes using a satellite observational operator: A case study on the 2010 Eyjafjallajökull volcanic eruption," *Atmospheric Chemistry and Physics*, vol. 17, no. 2, pp. 1187–1205, 2017.

[21] G. Evensen, "The Ensemble Kalman Filter: Theoretical formulation and practical implementation," *Ocean Dynamics*, vol. 53, no. 4, pp. 343–367, 2003.

[22] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[23] G. Fu, *Improving volcanic ash forecasts with ensemble-based data assimilation*. PhD thesis, Delf University of Technology, 2017.

[24] S. Lopez-Restrepo, A. Yarce, N. Pinel, O. L. Quintero, A. Segers, and A. W. Heemink, "Forecasting PM10 and PM2.5 in the Aburrá Valley (Medellín, Colombia) via EnKF based data assimilation," *Atmospheric Environment*, vol. 232, no. April, p. 117507, 2020.

[25] O. A. Popoola, D. Carruthers, C. Lad, V. B. Bright, M. I. Mead, M. E. Stettler, J. R. Saffell, and R. L. Jones, "Use of networks of low cost air quality sensors to quantify air quality in urban settings," *Atmospheric Environment*, vol. 194, no. February, pp. 58–70, 2018.

[26] S. Lopez-restrepo, A. Yarce, N. Pinel, O. Quintero, A. Segers, and A. W. Heemink, "Urban Air Quality Modeling Using Low-Cost Sensor Network and Data Assimilation in the Aburrá Valley, Colombia," *Atmosphere*, vol. 12, no. 91, pp. 1–19, 2021.

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

[28] 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.

[29] M. Ossés de Eicker, R. Zah, R. Triviño, and H. Hurni, "Spatial accuracy of a simplified disaggregation method for traffic emissions applied in seven mid-sized Chilean cities," *Atmospheric Environment*, vol. 42, no. 7, pp. 1491–1502, 2008.

[30] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.

[31] D. Tuia, M. Ossés de Eicker, R. Zah, M. Osses, *E. Zarate*, and A. Clappier, "Evaluation of a simplified top-down model for the spatial assessment of hot traffic emissions in mid-sized cities," *Atmospheric Environment*, vol. 41, pp. 3658–3671, 2007.

[32] C. D. Gómez, C. M. González, M. Osses, and B. H. Aristizábal, "Spatial and temporal disaggregation of the on-road vehicle emission inventory in a medium-sized Andean city. Comparison of GIS-based top-down methodologies," *Atmospheric Environment*, vol. 179, no. February, pp. 142–155, 2018.

[33] 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*, no. January, pp. 0–1, 2020.

[34] C. Mogollón-sotelo, L. Belalcazar, and S. Vidal, "A support vector machine model to forecast ground-level PM 2. 5 in a highly populated city with a complex terrain," *Air Quality, Atmosphere & Health*, 2020.

[35] EPA, "Meteorological Monitoring Guidance for Regulatory Modeling Applications," tech. rep., U.S. ENVIRONMENTAL PROTECTION AGENCY, 2000.

[36] 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*, vol. 40, no. 26, pp. 4946–4959, 2006. Special issue on Model Evaluation: Evaluation of Urban and Regional Eulerian Air Quality Models.

[37] S. J. Johnston, P. J. Basford, F. M. Bulot, M. Apetroaie-Cristea, N. H. Easton, C. Davenport, G. L. Foster, M. Loxham, A. K. Morris, and S. J. Cox, "City scale particulate matter monitoring using LoRaWAN based air quality IoT devices," *Sensors (Switzerland)*, vol. 19, no. 1, pp. 1–20, 2019.

[38] V. Isakov, S. Arunachalam, R. Baldauf, M. Breen, P. Deshmukh, A. Hawkins, S. Kimbrough, S. Krabbe, B. Naess, M. Serre, and A. Valencia, "Combining dispersion modeling and monitoring data for community-scale air quality characterization," *Atmosphere*, vol. 10, no. 10, 2019.

[39] S. Moltchanov, I. Levy, Y. Etzion, U. Lerner, D. M. Broday, and B. Fishbain, "On the feasibility of measuring urban air pollution by wireless distributed sensor networks," *Science of the Total Environment*, vol. 502, pp. 537–547, 2015.

[40] L. Morawska, P. K. Thai, X. Liu, A. Asumadu-Sakyi, G. Ayoko, A. Bartonova, A. Bedini, F. Chai, B. Christensen, M. Dunbabin, J. Gao, G. S. Hagler, R. Jayaratne, P. Kumar, A. K. Lau, P. K. Louie, M. Mazaheri, Z. Ning, N. Motta, B. Mullins, M. M. Rahman, Z. Ristovski, M. Shafiei, D. Tjondronegoro, D. Westerdahl, and R. Williams, "Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone?," *Environment International*, vol. 116, no. April, pp. 286–299, 2018.