Atmospheric Environment 232 (2020) 117507

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: http://www.elsevier.com/locate/atmosenv

Forecasting PM_{10} and $PM_{2.5}$ in the Aburrá Valley (Medellín, Colombia) via EnKF based data assimilation

Santiago Lopez-Restrepo [a,c,*], Andres Yarce [a,c], Nicolas Pinel [b], O.L. Quintero [a], Arjo Segers [d], A. W. Heemink [c]

[a] Mathematical Modelling Research Group at Universidad EAFIT, Medellín, Colombia
[b] Biodiversity, Evolution and Conservation Research Group at Universidad EAFIT, Medellín, Colombia
[c] Department of Applied Mathematics at TU Delft, Delft, the Netherlands
[d] TNO Department of Climate, Air and Sustainability, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Data assimilation
Air quality modelling
Chemical transport model
Ensemble kalman filter
Particulate matter

ABSTRACT

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

1. Introduction

The current exponential growth in world population heightens the importance of public health issues related to air quality (Akimoto, 2003; Gurjar et al., 2008). In developing countries, decision makers must cope with the environmental demands of expanding and overpopulated urban centers. Short term air quality forecasts and long term mitigation strategies for these centers are usually based on specialized assessments of particulate matter dynamics (Bell et al., 2011; Sallis et al., 2016). The Aburrá Valley houses the city of Medellín and its neighboring municipalities, constituting 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, with an elevation range of 1300–1750 m.a.s.l. From Northeast to South, and a height difference of up to 1800 m from the valley bottom to the canyon rim.

Air quality conditions deteriorate severely within the Aburrá Valley around the time of the arrival of the Intertropical Convergence Zone (March–April, and with lower intensity in October–November). During these periods, the atmospheric boundary layer persists below the rim of the canyon throughout the whole day, thus trapping the urban atmospheric contaminants within the lower atmosphere (Jiménez, 2016). Resulting concentrations of particulate matter below 10 μm (PM_{10}) and 2.5 μm ($PM_{2.5}$) remain at levels considered hazardous for the general population.

Understanding local and regional transport dynamics of atmospheric particulate matter thus becomes a top priority for urban valleys in the northern Andes. This efforts will enable the development of reliable air

* Corresponding author. Mathematical Modelling Research Group at Universidad EAFIT, Medellín, Colombia.
E-mail addresses: slopezr2@eafit.edu.co, s.lopezrestrepo@tudelft.nl (S. Lopez-Restrepo).

https://doi.org/10.1016/j.atmosenv.2020.117507
Received 11 December 2019; Received in revised form 4 April 2020; Accepted 11 April 2020
Available online 11 May 2020

quality forecasting systems for the Aburrá Valley (and similar locations) and strengthen decision making. In that vein, Chemical Transport Models (CTM) represent invaluable tools in understanding the dynamics of atmospheric contaminants, and as such have been widely used in monitoring air quality (Thunis et al., 2016; Lateb et al., 2016). In Colombia, studies have focused on the development of emission inventories and the characterization of pollutants (Toro et al., 2005, 2006; Zarate et al., 2007; Nedbor-Gross et al., 2018; Pachón et al., 2018). To date, however, the Colombian territory still lacks studies of data assimilation in CTM.

Previous atmospheric modelling efforts in the northern Andes have applied the Weather Research and Forecasting (WRF) model with its atmospheric chemistry module (WRF-CHEM), in conjunction with the EDGAR (Emissions Database for Global Atmospheric Research) global emissions inventory to simulate the behavior of PM_{10} over the Bogotá metropolitan area (Kumar et al., 2016b). In this case, results persistently underestimated by an order of magnitude the PM_{10} concentrations relative to the available measurements. The model WRF-CHEM has also been applied to study the behaviour of O_3 over the medium-size, mountain city of Manizales (Gonzalez et al., 2018). This work used high-resolution (1 km by 1 km) simulations to compare the performance of the model in two different emissions scenarios, using the EDGAR emission inventory and a high-resolution emission inventory previously developed (Gonzalez et al., 2017). The work showed a considerable improvement of the model under the high-resolution emissions inventory.

This study uses simulations of LOTOS-EUROS (LE) chemistry transport model (CTM) for studying the atmospheric contaminant dynamics within the Aburrá valley, spanning a set of 10 municipalities including the city of Medellín. LE is equipped with several Ensemble-based data assimilation applications focused in the reanalysis and forecasting of gasses over Europe (Manders et al., 2017). In (Barbu et al., 2009) the EnKF is used with the covariance localization technique for assimilating ground based observations to represent the dynamics of SO_4 and SO_2 over the European continent. In their work, two different sources of uncertainty were studied, the reaction rate in the production of SO_4 from SO_2 and the emissions of SO_2 and SO_4. The uncertainty was modelled as a colored noise process and estimated following the method presented in (Heemink and Segers, 2002) (explained in detail in Section 2.2). In the work of (Barbu et al., 2009) they concluded that by improving the error uncertainty representation, the emissions uncertainty and reaction rate uncertainty, the performance of the data assimilation was enhanced.

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

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

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

This paper is organized as follows. In Section 2 we present the materials and methods, including the theoretical framework for the ensemble-based data assimilation technique and the covariance localization that was used for improving the model results. Section 2.1 introduces the LE model and describes its main features. Section 3 presents the experimental set up, and the data assimilation calibration with different radii for covariance localization, and several factors for the stochastic processes in the LE model. It also presents the observation error covariance matrix estimation from ground-based sensor network data. Section 4 presents the main results of the paper in terms of investigating the ability to forecast particulate matter concentrations over a few days. Section 5 offers some concluding remarks and outlines the needed future work.

2. Ensemble-based data assimilation of the LOTOS-EUROS model

This section will briefly introduce the theoretical elements for data assimilation in the LE model, defining the Ensemble Kalman Filter, the covariance localization, and the stochastic process used for modelling emission corrections. All these elements play a fundamental role in the data assimilation strategy.

2.1. The LOTOS-EUROS model

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

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

$$\boldsymbol{c_k} = \boldsymbol{M}_{LE}(t_k, t_{k-1}, \boldsymbol{c}_{k-1}, \ldots) \tag{1}$$

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

2.2. Stochastic and uncertainty representation

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

$$\widehat{\boldsymbol{e}}_k = \boldsymbol{e}_k \left(1 + \boldsymbol{\delta e}_k \right) \tag{2}$$

In here, $\boldsymbol{e}_k$ is the nominal emission from the emission inventory.The emission deviation is modelled as an autoregressive model of order one (AR-1), following the structure of a colored noise process (Jazwinski, 1970):

$$\boldsymbol{\delta e}_k = \alpha_k \, \boldsymbol{\delta e}_{k-1} + \sigma\sqrt{1-\alpha^2}\, \boldsymbol{w}_k \tag{3}$$

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

$$\boldsymbol{w}_k \sim N(0,1) \tag{4}$$

Over an infinite number of samples, the stochastic factors are drawn out of a normal distribution with zero mean and standard deviation σ. The temporal correlation coefficient $\alpha_k \in [0,1]$ is used to describe the temporal variation, where the value should be set between two extremes: for $\alpha = 0$, the deviation is pure white noise with completely different values for every sample; for $\alpha = 1$ there is no temporal variation at all and the deviation factor is a single sample out of the normal distribution. In this study the correlation parameter is described using a temporal length scale τ following (Barbu et al., 2009):

$$\alpha_k = \exp(-|t_k - t_{k-1}| / \tau) \tag{5}$$

A stochastic model state is formed by augmenting the state vector (1) with the correction factor $\boldsymbol{\delta e}$:

$$\begin{bmatrix} \boldsymbol{c}_k \\ \boldsymbol{\delta e}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{M}_{LE}(\boldsymbol{c}_{k-1}, \boldsymbol{\delta e}_{k-1}) \\ \alpha_k \, \boldsymbol{\delta e}_{k-1} \end{bmatrix} + \begin{bmatrix} 0 \\ \sigma\sqrt{1-\alpha^2} \end{bmatrix} \boldsymbol{w}_k \tag{6}$$

or simply:

$$\boldsymbol{x}_k = M(\boldsymbol{x}_{k-1}) + G\, \boldsymbol{w}_{k-1} \tag{7}$$

With the augmented vector (6), it is possible to apply a sequential data assimilation scheme to estimate both the state and the emission correction factor. The non-linear operator M propagates the augmented state vector $\boldsymbol{x}$ in time, while G distributes the stochastic forcing $\boldsymbol{w}_k$ over the elements of the state. The description of the uncertainty sources and the stochastic model is presented in Section 3.

2.3. Ensemble Kalman Filter

The Ensemble Kalman filter (EnKF) is the most frequently used ensemble-based data assimilation method (Evensen, 2003). The EnKF is a Monte Carlo ensemble method, based on the representation of the probability density of the state estimates in an ensemble of N states:

$$\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N)} \tag{8}$$

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

The EnKF is initialized by generating a random ensemble $\boldsymbol{\xi}_0^{a(i)}$ to represent the uncertainty in the initial condition $\boldsymbol{x}_0$. Then, the forecast step of the EnKF propagates each ensemble member in time using the state-space operator from Eq. (7) and a random forcing:

$$\boldsymbol{\xi}_k^{f(i)} = M\left(\boldsymbol{\xi}_{k-1}^{a(i)}\right) + G\, \boldsymbol{w}_{k-1}^{(i)} \tag{9}$$

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

$$\boldsymbol{x}_k^f = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\xi}_k^{f(i)} \tag{10}$$

$$\boldsymbol{P}_k^f = \left[\boldsymbol{L}_k^f \left(\boldsymbol{L}_k^f\right)^T\right] \Big/ (N-1) \tag{11}$$

where the matrix $\boldsymbol{L}$ is formed by deviations of the ensemble members from the mean:

$$\boldsymbol{L}_k^f = \left[\boldsymbol{\xi}_k^{f(1)} - \boldsymbol{x}_k^f, \ldots, \boldsymbol{\xi}_k^{f(N)} - \boldsymbol{x}_k^f\right] \tag{12}$$

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

$$\boldsymbol{y}_k = \boldsymbol{H}_k\, \boldsymbol{x}_k + \boldsymbol{v}_k, \quad \boldsymbol{v}_k \sim N(0, \boldsymbol{R}_k) \tag{13}$$

The observation operator H describes how the observations are sampled from the concentration fields in the state. The observation representation error $\boldsymbol{v}_k$ describes the difference between the observations and the sampling, which are present due to both instrumental errors but also due to sampling errors. In this application the sampling errors are for example present since the state describes concentrations as averages in (large) grid boxes, while the observations concern point observations. The vectors $\boldsymbol{v}_k$ are assumed to be samples out of a random distribution with zero mean and covariance $\boldsymbol{R}_k$; how $\boldsymbol{R}_k$ is parameterized is described in section 3.

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

$$\boldsymbol{\xi}_k^{a(i)} = \boldsymbol{\xi}_k^{f(i)} + \boldsymbol{K}_k\left[\boldsymbol{y}_k - \boldsymbol{H}_k\boldsymbol{\xi}_k^{f(i)} + \boldsymbol{v}_k^{(i)}\right] \tag{14}$$

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

$$\boldsymbol{K}_k = \boldsymbol{P}_k^f \boldsymbol{H}_k^T \left[\boldsymbol{H}_k \boldsymbol{P}_k^f \boldsymbol{H}_k^T + \boldsymbol{R}_k\right]^{-1} \tag{15}$$

The Kalman gain is defined such that the sample covariance of the analysis ensemble is minimal with respect to l_2 matrix norm (Asch et al., 2016). Note that the sample covariance $\boldsymbol{P}_k^f$ cannot be computed in this application given its large size ($\sim\mathcal{O}(10^6) \times \mathcal{O}(10^6)$). However, for the actual implementation it is sufficient to store only the factorization $\boldsymbol{L}$ from Eq. (12).

2.4. Covariance localization

Due to the approximation of the state space covariance by a finite number of ensemble members, it is unavoidable the appearance of spurious correlations between elements of the state. These spurious correlations can be removed by a procedure called localization (Ott et al., 2004). The localization method used in this work is the *covariance localization* (Houtekamer and Mitchell, 2001). The covariance localization or Schur localization, focuses on the forecast error covariance matrix, cutting off longer range correlations in the error covariances at a specified distance (Houtekamer and Mitchell, 2001; Petrie, 2008). The pointwise multiplication is called a Schur product and denoted by ∘:

$$\left[f \circ \boldsymbol{P}^f\right]_{i,j} = \left[\boldsymbol{P}^f\right]_{i,j} [f]_{i,j} \tag{16}$$

The Schur product theorem ensures that if f and $\boldsymbol{P}^f$ are positive semidefinite, then the Schur product, $f \circ \boldsymbol{P}^f$, is positive semidefinite. A cutoff function would be defined by $r \in \mathrm{R}^+ \rightarrow G(r/\rho)$, where r is the Euclidean distance between two state members and ρ is a length scaling

called the localization radius (Sakov and Bertino, 2011). The localization radius is defined such that beyond this the correlation reduces from 1 and at a distance of more than 3.5ρ the correlation reduces to zero (Petrie, 2008). The cutoff function utilized in this work has the following form:

$$f_{i,j} = exp\left(-0.5\left(r_{i,j}/\rho\right)^2 \right) \tag{17}$$

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

$$\boldsymbol{K} = \left(f \circ \boldsymbol{P}^f\right)\boldsymbol{H}^T\left[\boldsymbol{H}\left(f \circ \boldsymbol{P}^f\right)\boldsymbol{H}^T + \boldsymbol{R}\right]^{-1}$$

An analysis of the DA scheme performance with different ρ values is presented in Section 3.

2.5. Performance metrics

In this work, the performance of the LOTOS-EUROS simulations and the assimilation scheme were calculated by comparison with a subset of the ground observations not used in the assimilation. As described in Section 3.2, the collection of observations available in this study is rather small, and therefore only two time series were used to quantify the performance. Three metrics were computed to compare the simulations (assimilations) with the validation data; in addition, diurnal cycles were also compared.

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

$$\text{MFB} = \frac{2}{M}\sum_{i=1}^{M}\frac{(\boldsymbol{H}(\boldsymbol{c}))_i - y_i}{(\boldsymbol{H}(\boldsymbol{c}))_i + y_i} \tag{18}$$

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

The *root mean square error* (RMSE) represents the sample standard deviation of the differences between predicted values and observed values (equation (19)). The RMSE penalizes a high variance as it gives errors with larger absolute values more weight than errors with smaller absolute values (Chai and Draxler, 2014):

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left((\boldsymbol{Hc})_i - y_i\right)^2} \tag{19}$$

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 (Yu et al., 2006):

$$\text{Corr} = \frac{\sum_{i=1}^{M}\boldsymbol{H}(\boldsymbol{c})_i - \overline{(\boldsymbol{H}(\boldsymbol{c})_i)}(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{M}(\boldsymbol{c})(\boldsymbol{c})(\boldsymbol{c})_i - \overline{(\boldsymbol{H}(\boldsymbol{c})_i)}^2}\ \sqrt{\sum_{i=1}^{M}(y_i - \bar{y})^2}} \tag{20}$$

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

3. Calibration of the data assimilation system

This section presents the results obtained from the data assimilation experiments with the LE model during a two-week episode. Simulations were conducted with the LE model in a nested domain configuration as described in Section 3.1. Default initial and boundary conditions were used, and data for assimilation was obtained and processed as described in Section 3.2. The goal of the experiments was to obtain insight into the sensitivity of the assimilation system for the configuration parameters such as the temporal correlation of the emission uncertainty (Section 3.5), the localization length scale (Section 3.4), and the observation representation errors (Section 3.6). Also the impact of other parameters such as the standard deviation of the parameter uncertainty was evaluated, but for the chosen configuration their impact was minor. To see the impact of each configuration parameter, these are calibrated and analyzed independently. Based on the results, the best values for the assimilation parameters were selected for use in subsequent assimilation experiments. A series of emission deviation factors were obtained during the two-week episode using the calibrated assimilation system and used as nominal emissions for the next two-week test period. The forecast skill of the calibrated assimilation system was evaluated throughout the episode as described in Section 4. The summarized experimental setup is presented in Fig. 1.

3.1. The LOTOS-EUROS model setup for Aburrá Valley

Simulations were conducted with the LE model, adopting a nested domain configuration as depicted in Fig. 2 and detailed in Table 1. Four nested domains were used to have a smooth transition on the dynamics from the regional scales (Caribbean and Northern part of South America) to the local conditions of the Aburrá Valley. The first Domain (D1) spans from the coast of Nicaragua in the West, to the Caribbean Dutch Islands and Venezuela in the East; model resolution was set to 0.27° (about 28 km). For this domain, meteorological data from ECMWF was used at a resolution of 0.14°; also the orography was obtained from this data set. The inner domain D2 is centered over the valley and includes the Northwest part of Colombia, encompassing most of the Colombian Andes; model resolution was set to 0.09°(about 9 km). For this and the following inner domains, meteorological data were obtained from ECMWF at 0.07°resolution, while for the elevation model was derived from the Global Multi Resolution Terrain Elevation Data set (GMTED2010) (Danielson and Gesch, 2011) at a resolution of 0.002° (approx. 220 m). The third inner domain D3 includes the department of Antioquia, at a model resolution of 0.03°(about 3 km). The innermost domain D4, the focus of the present study, includes primarily the region of the Aburrá Valley using model resolution of 0.01° (about 1 km).

The data sets used in the model are summarized in Table 2. For each of the domains, anthropogenic emissions were obtained from the global EDGAR inventory (Petrescu et al., 2012). Although previous works have shown that there is a considerable gap in the information in EDGAR for the Colombian territory, this database is the only one available with all the necessary species to run the model in the selected domains (Gonzalez et al., 2017; Pachón et al., 2018; Nedbor-Gross et al., 2018). The resolution of the EDGAR database is 10 × 10 km, which is approximately 10 times coarser than the resolution of the innermost domain. The low resolution of the emission data compared to the high resolution of the model simulation can produce an unrealistic spatial distribution of emissions and concentrations. This emphasizes the importance of considering emissions as a major source of uncertainty for the DA system. The behavior of our data assimilation scheme is studied using EnKF with 15 ensemble members ($N = 15$ in eqns. (8)-12) in Section 2.3) for both periods of assimilation. Previous experiments in related works and using LOTOS-EUROS model showed that using an greater ensemble members the performance of the algorithm did not increase significantly to justify the additional computational cost and 12–15 ensemble members are sufficient to describe the local covariance and to produce

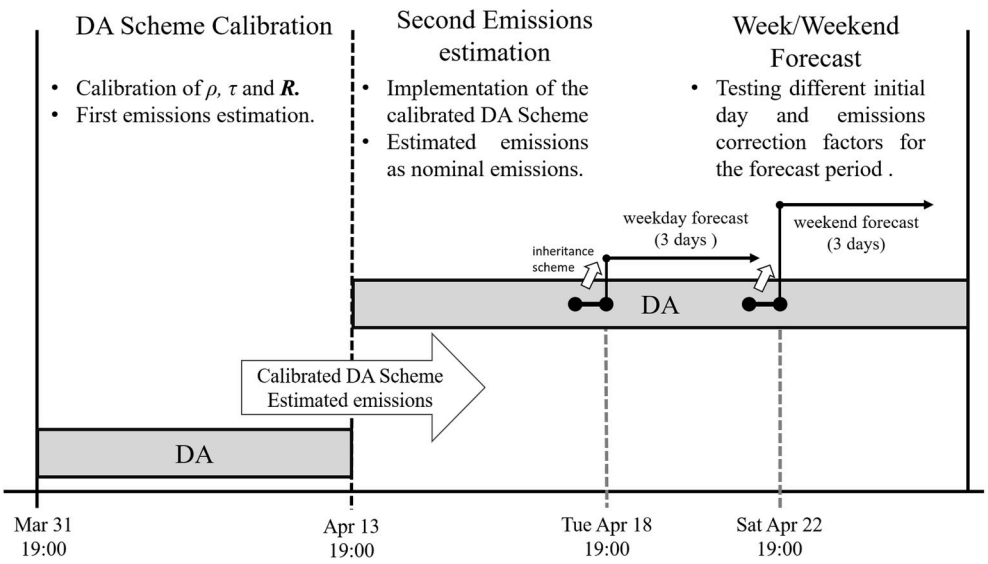

Fig. 1. Graphic representation of the experimental setup.

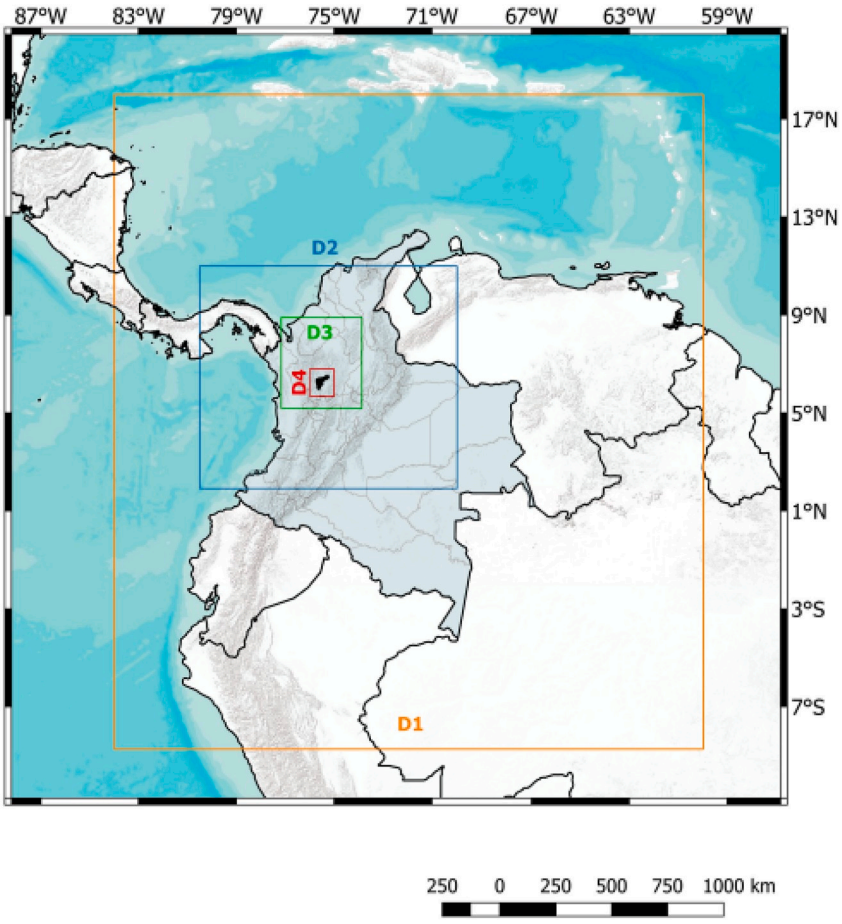

Fig. 2. Four nested domains for Metropolitan Area of Aburra Valley assesment.

assimilation with stable results Barbu et al. (2009); Curier et al. (2012).

3.2. Ground based data for assimilation

The *Sistema de Alerta Temprana del Valle de Aburrá* (SIATA) network of sensors provides high quality measurements for different air pollutants in the atmosphere over the Aburrá Valley region, monitoring species such as O_3, SO_2, PM_{10}, $PM_{2.5}$ and PM_1. The network is distributed in the five most populated Aburrá Valley's municipality, with the majority of the measuring stations located in the city of Medellín. The distribution of the observation sites is shown in Fig. 3.

In this work, only PM_{10} and $PM_{2.5}$ measurements were used for the

Table 1
Nested domain specifications.

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

Table 2
Data set used in the D4 domain.

Period	31-March-2016 to 25-April-2016
Metereology	ECMWF; Temp.res: 3h; spat.res: 0.07° × 0.07°
Initial and boundary conditions	LOTOS-EUROS (D3). Temp.res: 1h. 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°

assimilation experiments, obtained from 8 PM_{10} stations, 3 $PM_{2.5}$ stations, and 1 combined station on a total of 12 observation sites. The observation time series have an hourly temporal resolution, with full coverage for most days. Measurements for one station for each species (represented with a star in Fig. 3) were used for validation, taking two stations with a considerable distance between them to obtain a acceptable spatial representation, namely *Universidad San Buenaventura* (located within the city of Medellín, near the geographic center of the valley; station 37 in Fig. 3) for PM_{10}, and *Casa de la Justicia Itagui* (Southwest of the valley, in a mostly industrial zone; station 28 in Fig. 3) for $PM_{2.5}$. During air quality crises, these stations tend to reach some of the highest values measured during the year. The metrics from section 2.5 are calculated only over these two stations.

3.3. Standard model run results

In this section we described the simulated PM_{10} and $PM_{2.5}$ concentration for the two-week period between 31-March-2016 and 13-April-2016 using the model parameterization shown in Table 2 for domain D4. The evaluation against the validation stations are presented in Figs. 4 and 5. All the figures in this work are presented using the local time zone UTC-5. The statistical errors are shown in Table 3.

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

3.4. Calibration of covariance localization radius

The covariance localization as described in section 2.4 requires the definition of a localization radius ρ. In summary, the larger the radius chosen, the more observations are used to analyze a single element of the state. In this application, the state consists of concentrations and emission deviation factors, and the localization radius therefore has an impact on both. The influence of this parameter was evaluated by running the assimilation system with different values for ρ: 5, 10, 20, and 30 km. The temporal correlation length was fixed in $\tau = 3$ days for all the experiments. Fig. 6 shows maps of the average value over the 2 week assimilation window emission deviation factor δe.

Figs. 7 and 8 show time series of the average diurnal cycle of

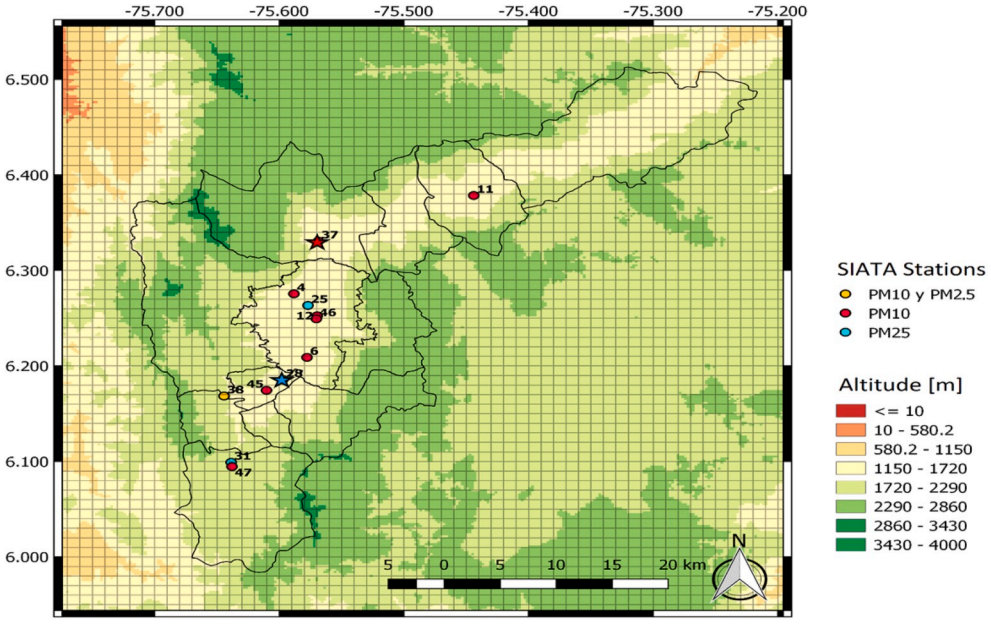

Fig. 3. SIATA sensor network for PM_{10} and $PM_{2.5}$. The stars represent observation points for validation and the circles represent observations points for assimilation.

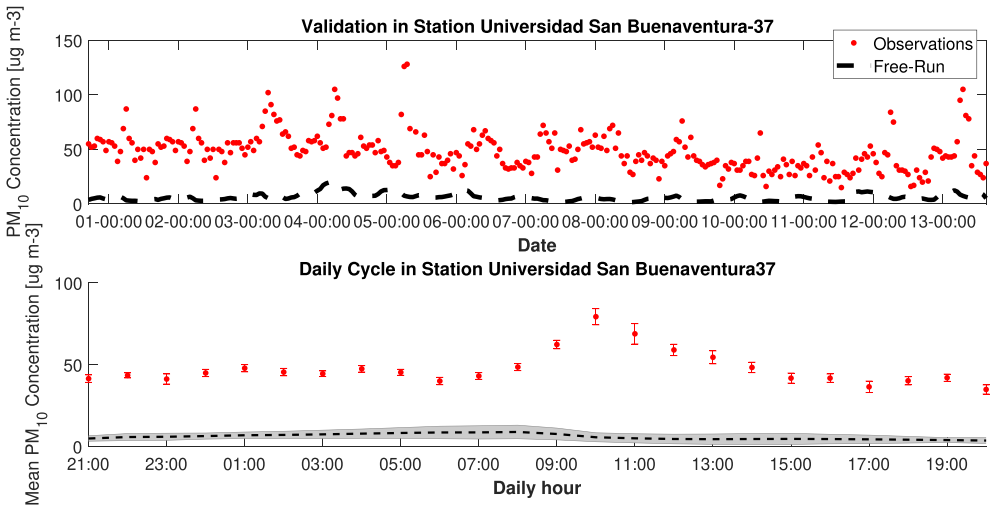

Fig. 4. Time series and diurnal cycle of PM_{10} in validation site 37 for a standard model Free-Run. The time axis corresponds with the local time zone UTC-5.

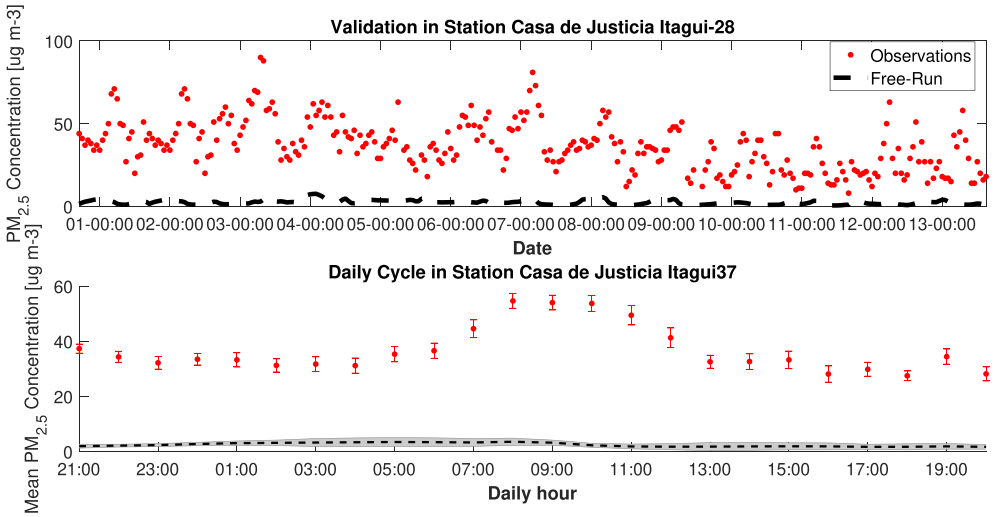

Fig. 5. Time series and diurnal cycle of $PM_{2.5}$ in validation site 28 for a standard model Free-Run. The time axis corresponds to the local time (UTC-5).

Table 3
Statistical error evaluation for PM_{10} and $PM_{2.5}$ via MFB, RMSE and CF for model standard Free-Run.

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

particulate matter concentrations in the two validation sites for the assimilation period.

For small localization radii, the concentrations in the validation sites were less influenced by the analysis since the number of analyzed observations was limited for these locations. If the localization radius increases, the simulations become more in line with the observations, although even for a ρ of 30 km the simulations are lower than what is observed. In both stations, the assimilated model progressively approaches the observations towards the end of the assimilation window. The day cycles for both species show that the temporal dynamics are not significantly affected by the different values of ρ and the change is mostly reflected in the magnitude. For the available sensor network, the value of $\rho = 30$ km presented the best overall performance for both species. As shown in Tables 4–5, the improvement of the error statistics related to the absolute error (MFB and RMSE) was more significant than the change in correlation factor (CF). The lack of accurate emissions inventories seems to have had a similar impact on simulations in all sites of the network, and therefore the best performance was obtained by changing emissions in the same way over the entire domain. It is expected that when using a sensor network with a higher spatial density, smaller values for ρ will become beneficial.

3.5. Calibration of temporal correlation length

With a similar experiment as for the covariance localization, the temporal correlation parameter τ of the emission uncertainty described in section 2.2 was calibrated. The uncertainty on the emissions was modelled via equation (3). To evaluate the impact of the temporal correlation parameter, the assimilation experiment was performed with τ set to either 1, 2, 3, or 5 days. The localization radius was fixed in $\rho = 30$ km for all the experiments.

Figs. 9 and 10 show the time series and average diurnal cycles of PM_{10} or $PM_{2.5}$ in the two validation sites, obtained from the observations, a standard model run, and analyzed ensemble means from assimilation experiments with different τ. Compared with the results shown in Figs. 7 and 8 for variation of the localization radius, the impact of changes in the temporal correlation length are rather small. The

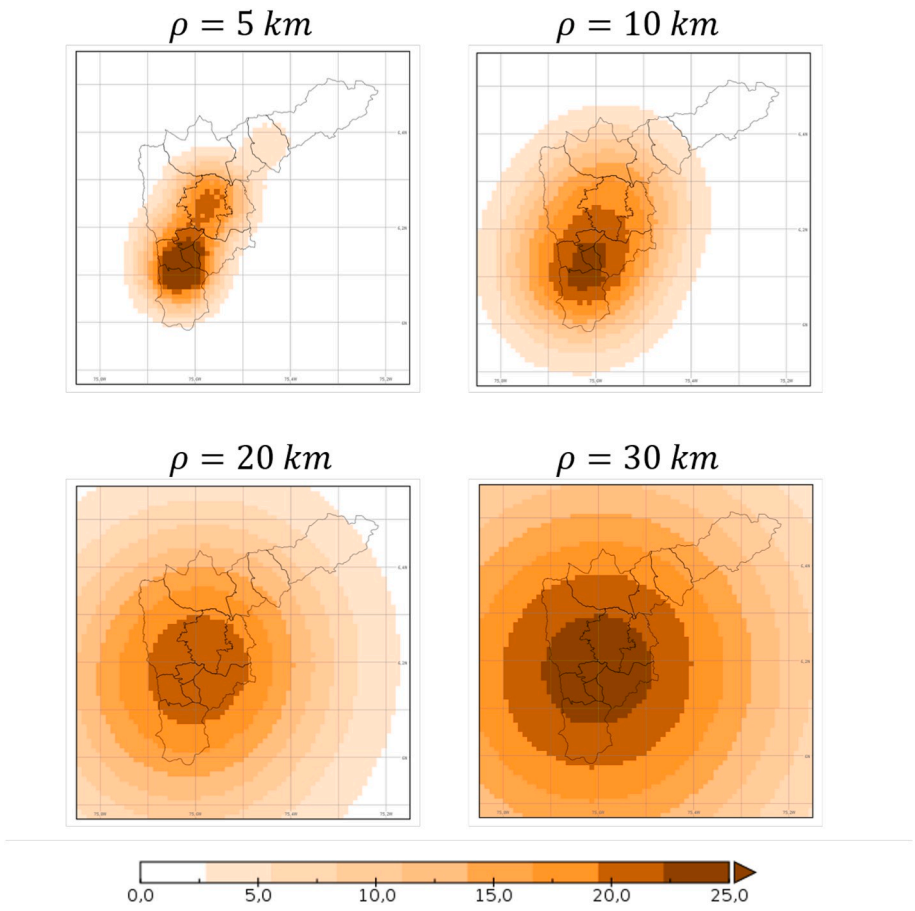

Fig. 6. Maps of mean emission deviation factors for particle matter emissions during assimilation experiments with different localization radii.

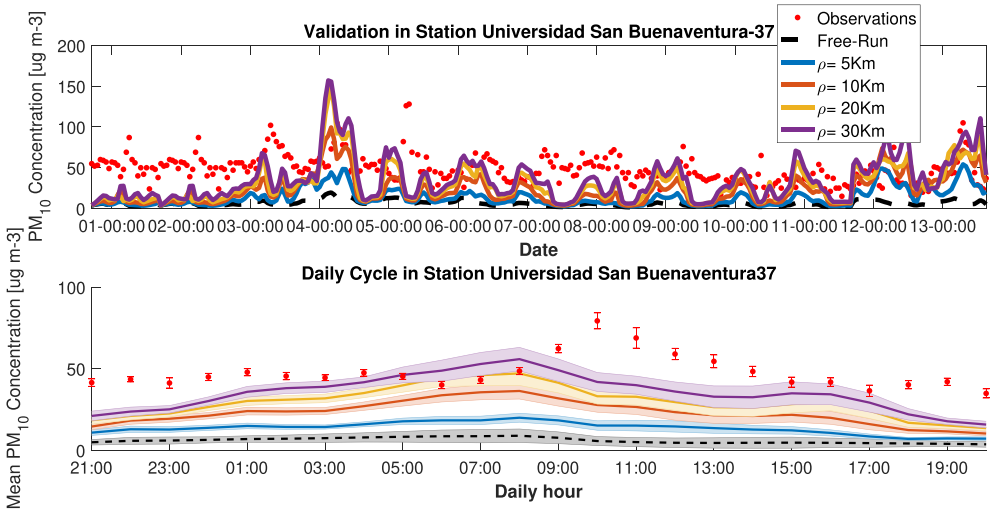

Fig. 7. Time series and diurnal cycle of PM_{10} in validation site 37. Dots denote observations, dashed black lines are simulations by the standard model, and solid lines are analyzed ensemble means from the assimilation for different localization radii. The diurnal cycles were obtained from 13 samples for each hour. The shadows and the bars represent the standard deviation of the 13 samples. The time axis corresponds with the local time zone UTC-5.

assimilation results hardly differ from each other when τ changes, indicating that in the current application this parameter is of minor importance.

Tables 6 and 7 show the values of the metrics defined in section 2.5 for the assimilation experiments with different τ. The MFB, RMSE and CF for both localization radius and correlation length showed good behavior in estimations for the PM_{10} and $PM_{2.5}$. Variations in the local analysis radius tended to diminish the MFB, RMSE and CF for the PM_{10} and $PM_{2.5}$ estimates in Figs. 6 and 7. The increase in correlation time does not seem to have improved statistical measurements and in general presents a smaller impact in the data assimilation performance than the localization radius.

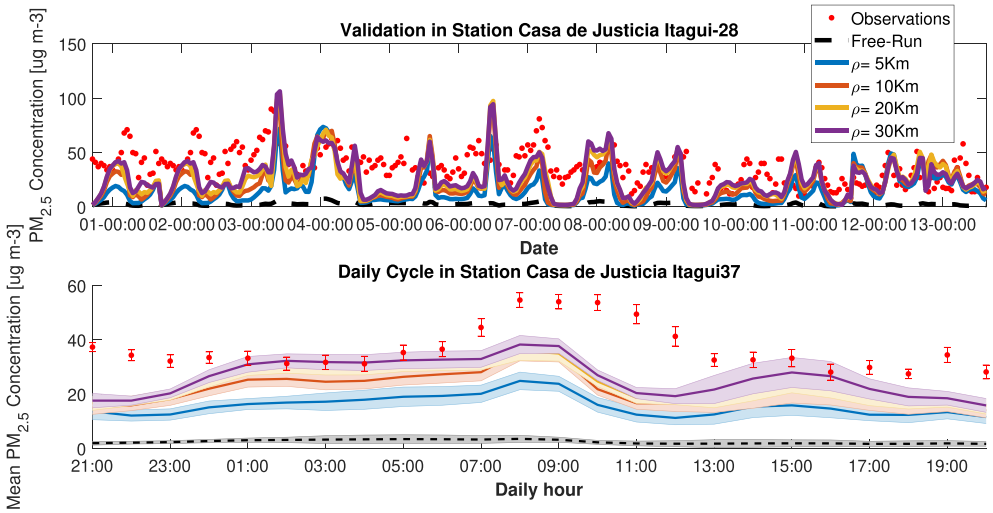

Fig. 8. Time series and diurnal cycle of $PM_{2.5}$ in validation site 28. Lines as in Fig. 7.

Table 4
Statistical error evaluation for PM_{10} via MFB, RMSE and CF for ρ variation.

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

Table 5
Statistical error evaluation for $PM_{2.5}$ via MFB, RMSE and CF for ρ variation.

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

3.6. Calibration of observation error covariance

Since the observation network described in section 3.2 has not been previously used for a data assimilation experiment, no suitable formulation for the observation error representation covariance was present yet. We implemented the method proposed in (Desroziers et al., 2005) to estimate the observation error covariance matrix **R**. Desroziers et al. (2005) showed that the relation:

$$\mathbf{E}\left[\mathbf{d}_a^o\left(\mathbf{d}_f^o\right)^T\right]=\mathbf{R} \tag{21}$$

is valid if the matrices specified in

$$\mathbf{HK}=\mathbf{HP}^f\mathbf{H}^T\left(\mathbf{HP}^f\mathbf{H}^T+\mathbf{R}\right)^{-1} \tag{22}$$

are the true covariances for background and observation error. Here **K** is the Kalman gain, $\mathbf{d}_f^o$ is the difference between observations and forecast state in observation space and $\mathbf{d}_a^o$ is the difference between observations and analysis state in observation space. One application of this relationship is that it can be used to diagnose the observation error variance after the analysis cycle has been completed (Li et al., 2009). In practice,

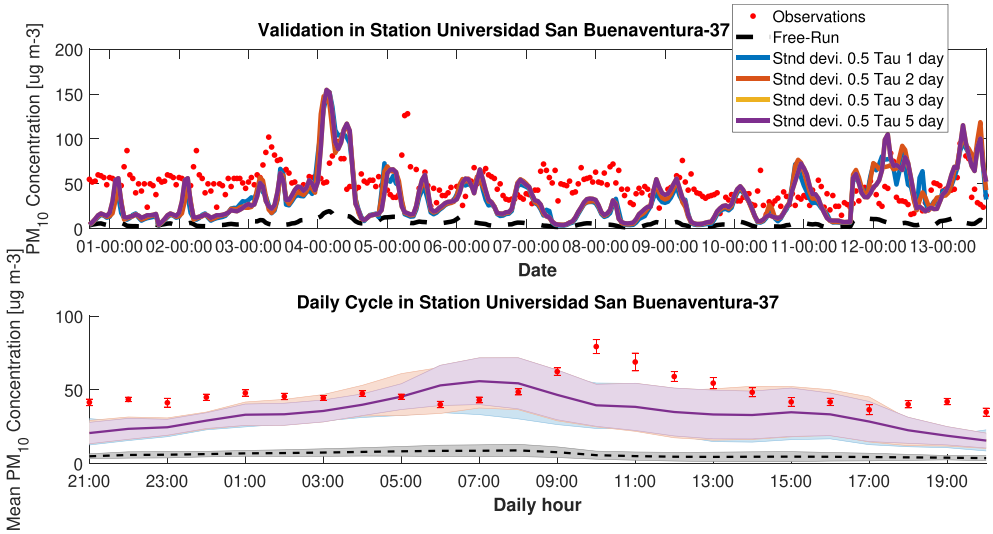

Fig. 9. Time series and diurnal cycle of PM_{10} in validation site 37. Dots denote observations, dashed black lines are simulations by the standard model, and solid lines are analyzed ensemble means from the assimilation for different temporal correlation lengths. The diurnal cycles were obtained from 13 samples for each hour. The shadows and the bars represent the standard deviation of the 13 samples.

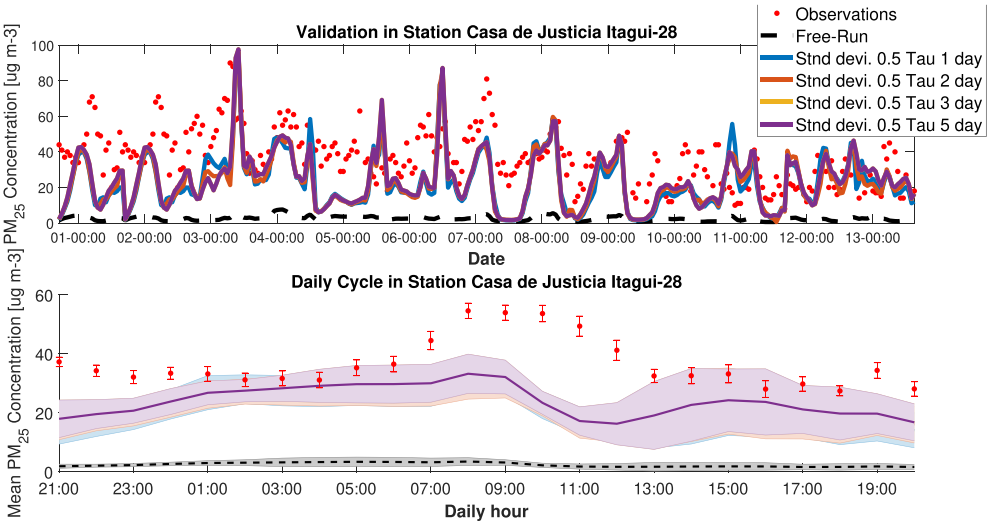

Fig. 10. Time series and diurnal cycle of $PM_{2.5}$ in validation site 28. Lines as in Fig. 9.

Table 6
Statistical error evaluation for PM_{10} via MFB, RMSE and CF for both localization radius variation and correlation length.

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

Table 7
Statistical error evaluation for $PM_{2.5}$ via MFB, RMSE and CF for both localization radius variation and correlation length.

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

the requirements for the relationship in Eq. (22) are never fully satisfied because the background covariance matrix is only an approximation of the real one. Nevertheless, the relationship could be used to obtain a useful first estimate of the observation error covariance matrix. For any subset of observations i with M observations, it is possible to compute an estimate of the error variance

$$\left(\widehat{\sigma}_o^2\right)_i = \frac{\left(\mathbf{d}_a^o\right)_i^T \left(\mathbf{d}_f^o\right)_i}{M} \tag{23}$$

$$= \sum_{j=1}^{M} \frac{\left(y_j - \boldsymbol{H}(\mathbf{x}^{\mathbf{a}})_j\right)\left(y_j - \boldsymbol{H}(\mathbf{x}^{\mathbf{f}})_j\right)}{M}$$

where $\widehat{\sigma}_o^2$ correspond with the diagonal of the matrix **R**.

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

For the subsequent experiment (test period, April 13–25), the off-line estimated matrix was used in the assimilation exercise. Once the DA scheme was calibrated, the estimated values for emissions correction factors were applied to the emissions inventory in EDGAR V4.2 and taken as the nominal emissions inventory for the test period. The new DA iteration was done using the estimated values for **R**, the radius size in the covariance localization scheme and τ.

4. Emissions estimation and particulate matter forecasting

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

4.1. Model data assimilation with a calibrated scheme

Once the calibration process was completed, a new model run was conducted using the corrected emissions as nominal emission values, and a new 12-day data assimilation exercise was performed using the chosen fixed radius (for local analysis), time correlation length τ and the estimated observation error covariance **R** as was shown in Fig. 1. In this second period the emissions were again estimated, but starting for the estimated emissions in the first period. Thus, the emissions were updated twice. It is important to note that experiments with other combinations of ρ and τ were performed but, in all the cases the results using the selected values in the previous section presented the best performance.The comparison between the nominal emissions of PM_{10} (from EDGAR data base) used in the first experiments and the estimated emissions of PM_{10} used as nominal emissions for this new model runs is presented in Fig. 11. Beyond the clear different in terms of magnitude and resolution between the estimated emissions and EDGAR that allow a better spacial representation of the emissions, the location of the hotspots in EDGAR appears in rural zones of the valley with minimal influence of human activity. The estimated emissions try to correct this behavior centring most of the emission in the urban zone of the valley, giving a more realistic representation.

The results of the assimilation for PM_{10} follow closely the measurements from validation station *Universidad San Buenaventura* (center of the valley) from April 13 at 19:00 UTC-5 through April 25 at 11:00 UTC-5 (see Fig. 12). The peak near 18–00:00 (and in general almost all the day close to that hour) can be caused by an incorrect time profile in the emissions factors from EDGAR database, that does not reflect the real temporal dynamics of the emissions. Additionally, the meteorological fields can cause and increment of the concentrations levels. Note that the daily cycle for the assimilated model remains closer to the observations than the model without assimilation and with the previously estimated emissions.

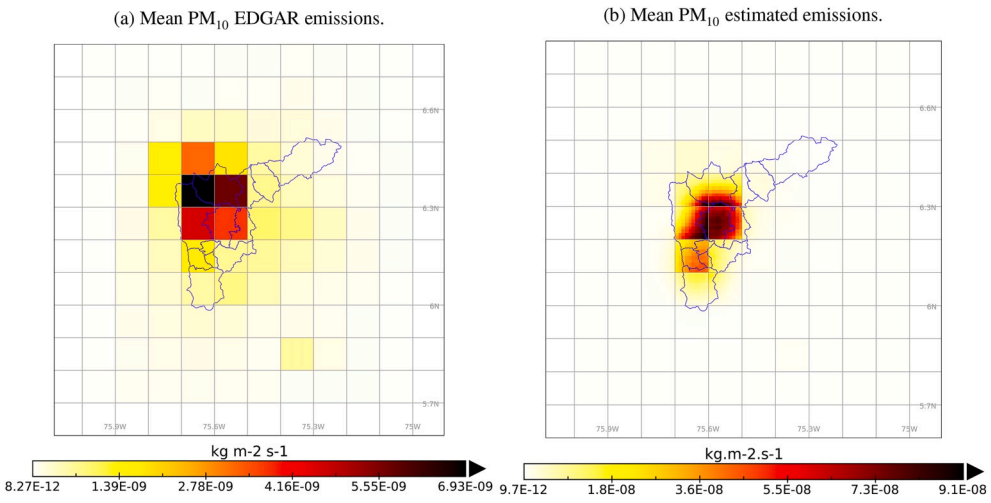

Fig. 11. Comparison between EDGAR PM_{10} and estimated PM_{10} emissions.

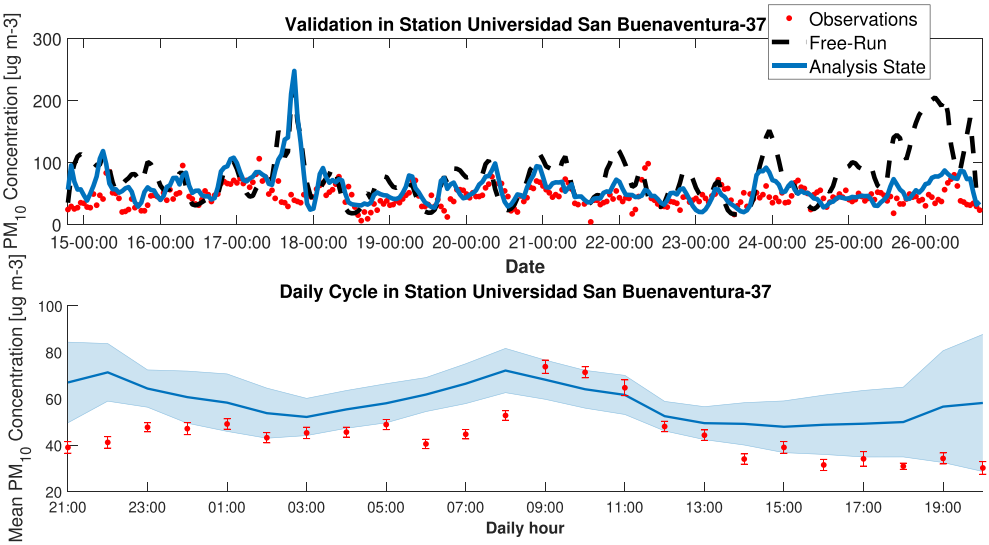

Fig. 12. PM_{10} validation for the second DA iteration. Estimated emissions were used as nominal emissions, the estimated observation error covariance is used in the assimilation step. Red points are observations, solid black line is the free run model and the solid blue line is the analysis step for 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. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

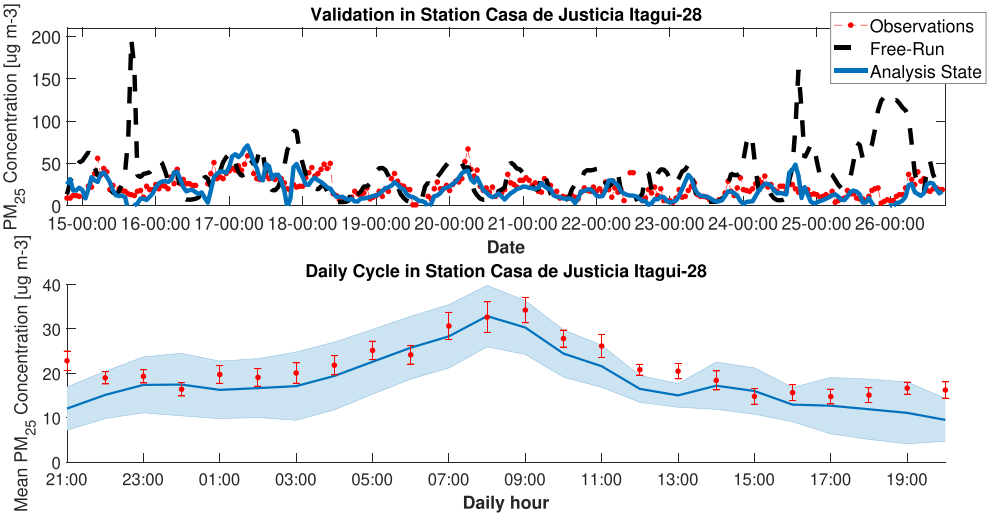

Fig. 13. $PM_{2.5}$ validation for the second DA iteration. Estimated emissions were used as nominal emissions, the estimated observation error covariance was used in the assimilation step. Red points are observations, solid black line the free run model and solid blue line the analysis step for the assimilated model. The diurnal cycles were obtained from 13 samples for each hour. The shadows and the bars represent the standard deviation of the 13 samples. The time axis corresponds with the local time zone UTC-5. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Fig. 13 shows a similar comparison for the $PM_{2.5}$ station. The model in a free run tends to over estimate the $PM_{2.5}$ concentrations (see peaks in 15 April at 23:00 UTC-5, 24 April at 22:00 and 25 April at 23:00 UTC-5). The results of the assimilation process offer a better average estimation. The daily cycle of $PM_{2.5}$ within the Aburrá valley is related to the industrial and mobile sources emissions profile and the meteorological conditions inside the valley.

The second period of assimilation provides a good representation not only of the dominant dynamics, but also offer an opportunity to forecast taking into account the profiles of emission sources. The next section will address the results from forecasts for the assimilated model with different radii in local analysis and correlation time lengths for the emissions.In the Appendix is presented a validation of the model for O_3 and NO_2 concentrations for the second study period.

4.2. Forecasting PM profiles during weekdays and weekends

Using the second assimilation period (twice estimated emissions and the analysis state as initial condition, see Fig. 1) three forecasts experiments were performed for up to three days, under the following scenarios: *i)* forecast starting on a Saturday night (19:00 UTC-5), with an assimilation window of the nine (9) days prior; *ii)* forecast starting on Tuesday night (19:00 UTC-5), with an assimilation window of the five (5) days prior; and *iii)* as in *ii)*, but using a localization radius of 5 km instead of 30 km.

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

1. Forecast default: Retaining only the state values from the end of the assimilation window. The correction factors estimated in the assimilation window are not used in the forecast.
2. Forecast hourly: Starting from the state values of the end of the assimilation window and using the hourly profile from the last 24 h in the assimilation window.
3. Forecast average: Starting from the state values of the end of the assimilation window and using the average state values from the entire assimilation window.

For both PM species, forecasts starting on a weekend failed to reflect the observed dynamics. Forecasts initiating on a weekday were able, in general, to replicate the observed dynamics, having better performance in reproducing the dynamics of PM_{10} (Fig. 15). All three inheritance schemes used in the PM_{10} weekday forecast (Fig. 15) reproduced the observed dynamics, but the hourly scheme tracked the measured concentrations the closest. For the $PM_{2.5}$ weekday forecast, again the hourly profile tracked the measured concentrations more closely than the other two during the forecast period.

Fig. 16 presents a comparison of the forecasts for PM_{10} and $PM_{2.5}$ under different local analysis radii. The smaller 5 km radius for local analysis, does not improve either forecast. We can interpret this phenomenon if we look at the forecast with 30 km radius in local analysis, taking into account that including more sensors in the assimilation it is possible to improve the correction factors in emissions. Consequently, if the emissions correction factors are higher for the latter two weeks in local analysis with 30 km radius, it can reduce the real emissions at a higher rate than the local analysis of 5 km radius via DA.

In order to provide quantitative measurements of forecast performance under various scenarios, the following error statistics were calculated and presented in Fig. 17: Mean Factoral Bias(MFB), Root Square Mean Error (RSME); and Correlation Factor (CF). The error statistics are calculated over a single forecast for each scenario and over the validations stations described in Fig. 3. Since there were no considerable changes in the error statistics between the forecast days, the presented values correspond to the three-day average. Only PM_{10} are presented; the behavior of $PM_{2.5}$ is very similar. Weekday forecasts (initiating on Tuesday) under a 30 km local analysis ratio scenario, presented the best error statistics. Independent of the inheritance scheme and the localization radius, weekend forecasts performed worse than weekday forecasts. For weekday forecasts, scenarios with localization radius of 5 km tended to perform worse than scenarios using a localization radius of 30 km.

5. Conclusions

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

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

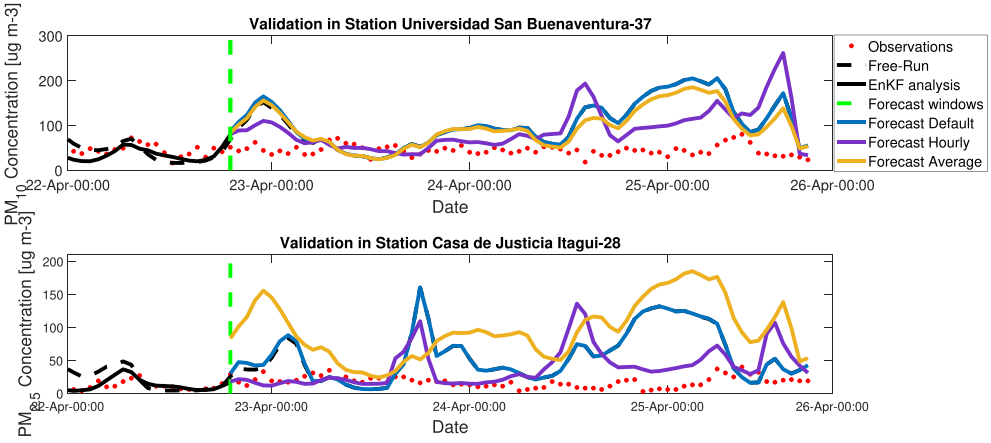

Fig. 14. PM Forecast Starting on Saturday 19:00 UTC-5. The red points indicate observations; the dotted black line indicates the Free-Run; the solid black line shows the analysis of the EnKF; blue, purple and yellow lines show the forecasts under the different scenarios; vertical green line indicates the start of the forecasts. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

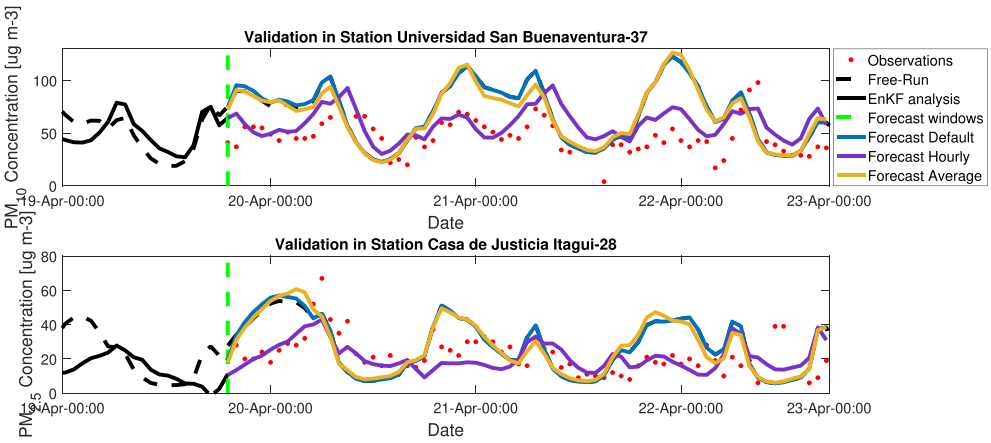

Fig. 15. Pm forecast starting on tuesday 19:00 UTC-5. Lines as in Fig. 14.

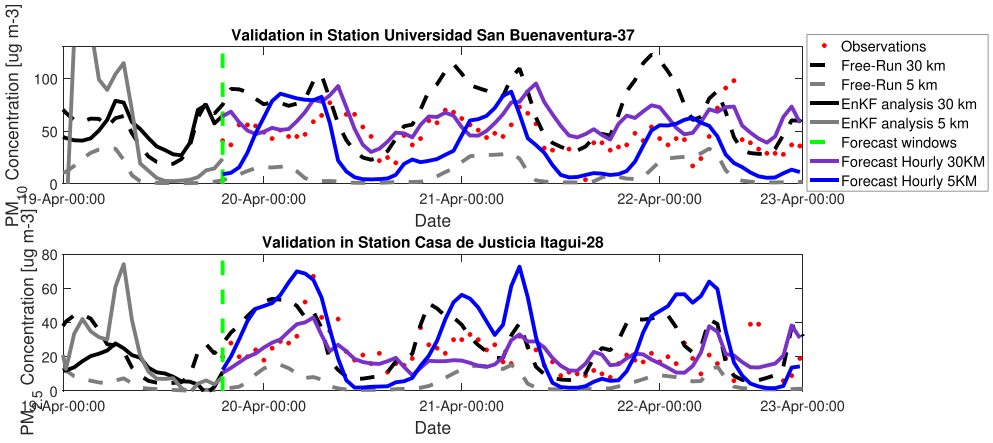

Fig. 16. PM Weekday Forecast Comparisons for Different Radii. Red points depicts the observations, dotted black line the Free-Run using a 30 km radius, dotted grey line the Free-Run using a 5 km radius, solid black line the analysis of the EnKF using a 30 km radius, solid grey line the analysis of the EnKF using a 5 km radius, and purple and blue lines the forecast scenarios for 30 km and 5 km radius, respectively. Vertical green line indicates the beginning of the forecast window. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

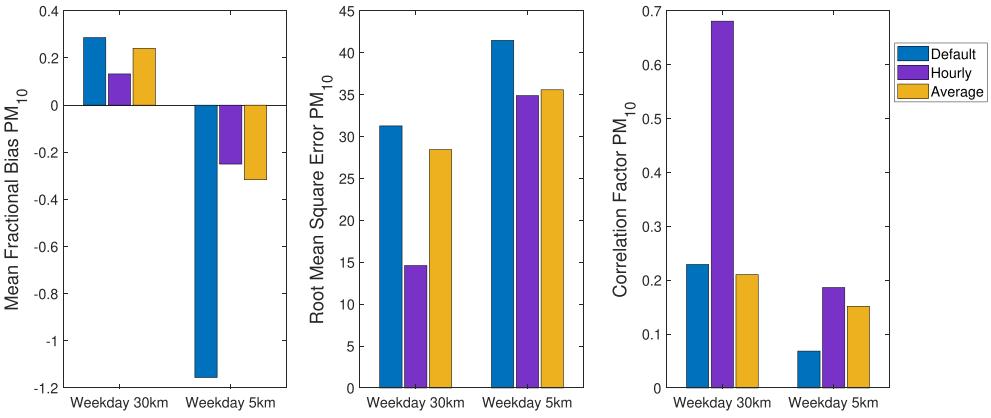

Fig. 17. PM_{10} Forecast Error Statistics. Blue bars represent forecasts under the default inheritance scheme. Purple bars indicate forecasts under the hourly inheritance scheme. Yellow bars show forecasts under the average inheritance scheme. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

and the estimation of their atmospheric concentrations within the Aburrá Valley.

Calibration of the radius for local analysis, the correlation time length τ and the estimation of the observation covariance error matrix **R**, led to a better tuned DA scheme with improved performance, approaching more closely the available observations. The estimation of an emission correction factor via data assimilation compensated for the scarcity in accurate and detailed emissions inventories for the region, enabling more accurate simulation results. Due to the coarse resolution of the emission inventory, and the rather low density of the sensor network available for 2016 within the area of interest, a large localization radius (30 km) performance better than a small radius (5 km).

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

yield improved performance and increase the resolution and usability of air quality forecast systems. Further researches are needed with better inputs fields of the CTM as a local and more detailed emission inventory and meteorology with better resolution capable to represent the transport dynamics into the valley. Improvement of emission inventories and meteorological input is subject of current studies. Additionally, a data assimilation scheme that consider uncertainty in the meteorological variables and different emissions correction factors for each component can help to improve the presented results.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Santiago Lopez-Restrepo: Conceptualization, Methodology, Software, Writing - original draft. **Andres Yarce:** Methodology, Software. **Nicolas Pinel:** Conceptualization, Methodology, Writing - review & editing. **O.L. Quintero:** Conceptualization, Methodology, Writing - original draft. **Arjo Segers:** Methodology, Software, Writing - review & editing. **A.W. Heemink:** Conceptualization, Supervision.

Appendix

O_3, NO_x SO_2 are crucial for the secondary aerosol formation and the PM modelling (Barbu et al., 2009; Manders et al., 2009). In this Appendix (Figs. 18 and 19) is shown a comparison of the model concentrations for O_3 and NO_2 for the second period of Data Assimilation, using the calibrated DA scheme and the estimated emissions (see Fig. 1). Unfortunately, there are not available data from SIATA networks to evaluate the concentrations of SO_2 in the period of interest, and there are not others sources of quality data over the region. The Figs. 18 and 19 show that in general the LE model tend to underestimate the O_3 and NO_2 concentrations, and no all the cycles are well captured by the model. These results support the idea that a improvement in the emission inventory and the meteorological fields are required to improve both, the gases and the aerosol representation in the Aburrá Valley.

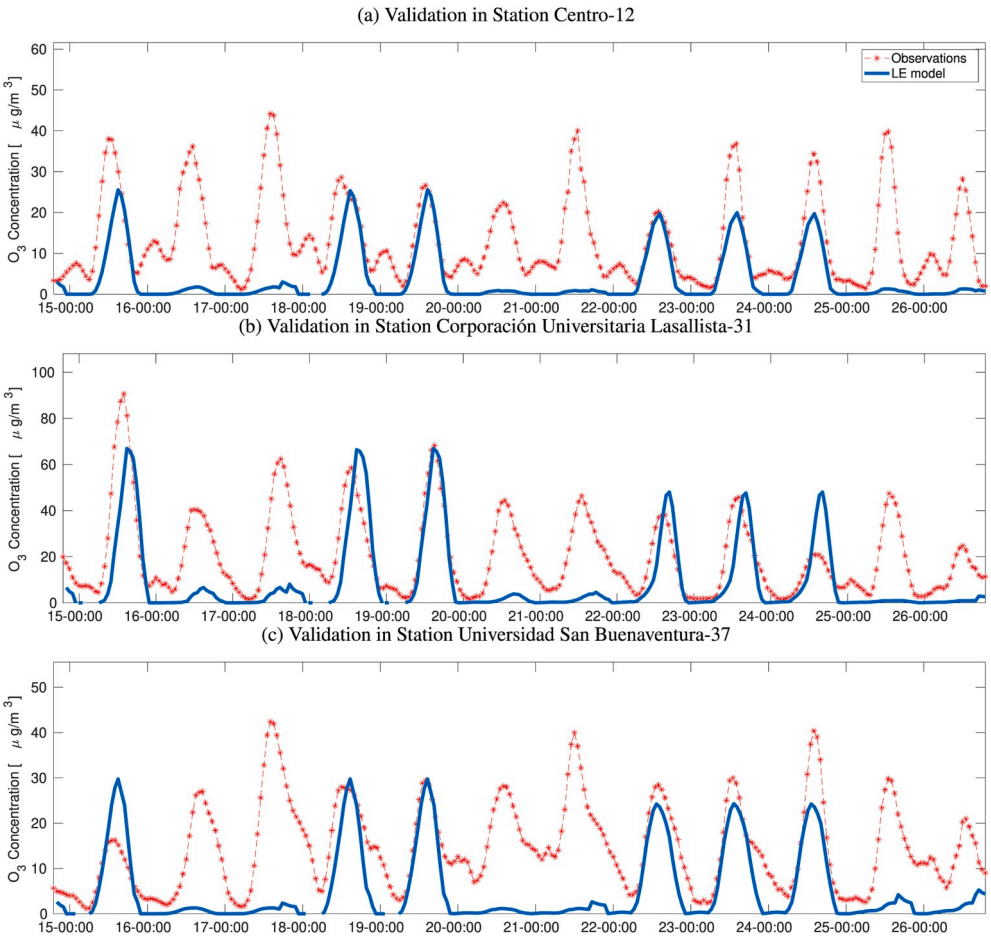

Fig. 18. Comparison of LOTOS-EUROS O_3 concentration and SIATA observations. The time axis corresponds with the local time zone UTC-5.

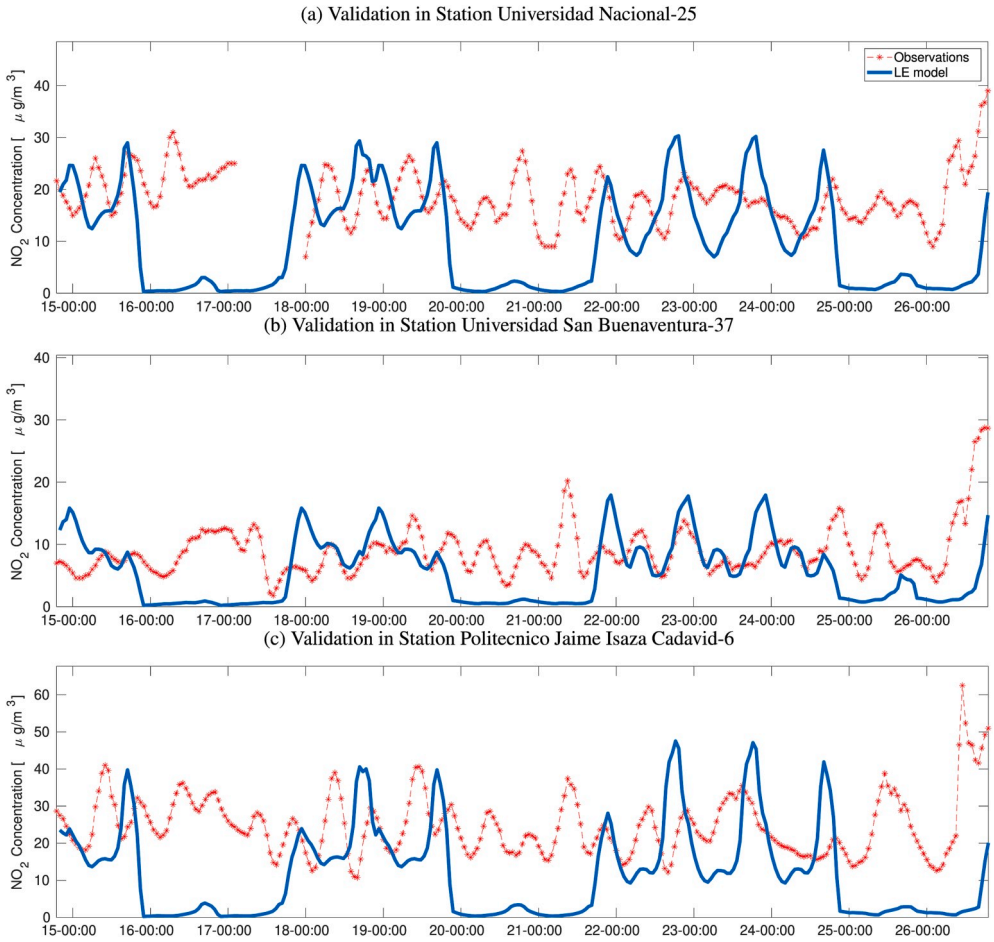

Fig. 19. Comparison of LOTOS-EUROS NO_2 concentration and SIATA observations. The time axis corresponds with the local time zone UTC-5.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2020.117507.

References

Akimoto, H., 2003. Global air quality and pollution. Science 302, 1716–1719. https://doi.org/10.1126/science.1092666.

Asch, M., Bocquet, M., Nodet, M., 2016. Data Assimilation, 1 ed. Society for Industrial and Applied Mathematics, Philadelphia, PA. https://doi.org/10.1137/1.9781611974546.

Barbu, A.L., Segers, A.J., Schaap, M., Heemink, A.W., Builtjes, P.J.H., 2009. A multi-component data assimilation experiment directed to sulphur dioxide and sulphate over Europe. Atmos. Environ. 43, 1622–1631. https://doi.org/10.1016/j.atmosenv.2008.12.005.

Bell, M.L., Cifuentes, L.A., Davis, D.L., Cushing, E., Telles, A.G., Gouveia, N., 2011. Environmental health indicators and a case study of air pollution in Latin american cities. Environ. Res. 111, 57–66. https://doi.org/10.1016/j.envres.2010.10.005. http://www.sciencedirect.com/science/article/pii/S0013935110001751.

Boylan, J.W., Russell, A.G., 2006. Pm and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models. Atmos. Environ. 40, 4946–4959. https://doi.org/10.1016/j.atmosenv.2005.09.087 special issue on Model Evaluation: Evaluation of Urban and Regional Eulerian Air Quality Models.

Chai, T., Draxler, R.R., 2014. Root mean square error (rmse) or mean absolute error (mae): arguments against avoiding rmse in the literature. Geosci. Model Dev. (GMD) 7, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

Curier, R.L., Timmermans, R., Calabretta-Jongen, S., Eskes, H., Segers, A., Swart, D., Schaap, M., 2012. Improving ozone forecasts over Europe by synergistic use of the LOTOS-EUROS chemical transport model and in-situ measurements. Atmos. Environ. 60, 217–226. https://doi.org/10.1016/j.atmosenv.2012.06.017.

Danielson, J., Gesch, D., 2011. Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010). U.S. Geological Survey Open-File Report 2011–1073 2010, p. 26. citeulike-article-id:13365221.

Desroziers, G., Berre, L., Chapnik, B., Poli, P., 2005. Diagnosis of observation, background and analysis-error statistics in observation space. Q. J. R. Meteorol. Soc. 131, 3385–3396. https://doi.org/10.1256/qj.05.108.

Evensen, G., 2003. The ensemble kalman filter: theoretical formulation and practical implementation. Ocean Dynam. 53, 343–367. https://doi.org/10.1007/s10236-003-0036-9.

Fountoukis, C., Nenes, A., 2007. Isorropiaii: a computationally efficient thermodynamic equilibrium model for k+-ca2+-mg2+-nh4 +-na+-so4 2–no3 –cl–h2o aerosols. Atmos. Chem. Phys. 7, 4639–4659.

Fu, G., 2017. Improving Volcanic Ash Forecasts with Ensemble-Based Data Assimilation. Ph.D. thesis. TU Delft. https://doi.org/10.4233/uuid:97d82967-998a-413d-b1f8-8c46f3e064ccImportant.

Gonzalez, C.M., Gomez, C.D., Rojas, N.Y., Acevedo, H., Aristizabal, B.H., 2017. Relative impact of on-road vehicular and point-source industrial emissions of air pollutants in a medium-sized Andean city. Atmos. Environ. 152, 279–289. https://doi.org/10.1016/j.atmosenv.2016.12.048.

Gonzalez, C.M., Ynoue, R.Y., Vara-Vela, A., Rojas, N.Y., Aristizabal, B.H., 2018. High-resolution air quality modeling in a medium-sized city in the tropical Andes: Assessment of local and global emissions in understanding ozone and PM10 dynamics. Atmospheric Pollution Research 1–15. https://doi.org/10.1016/j.apr.2018.03.003.

Gurjar, B., Butler, T., Lawrence, M., Lelieveld, J., 2008. Evaluation of emissions and air quality in megacities. Atmos. Environ. 42, 1593–1606. https://doi.org/10.1016/j.atmosenv.2007.10.048. http://www.sciencedirect.com/science/article/pii/S1352231007009697.

Heemink, A.W., Segers, A.J., 2002. Modeling and prediction of environmental data in space and time using Kalman filtering. Stoch. Environ. Res. Risk Assess. 16, 225–240. https://doi.org/10.1007/S00477-002-0097-1.

Heemink, A.W., Verlaan, M., Segers, A.J., 2001. Variance reduced ensemble kalman filtering. Mon. Weather Rev. 129, 1718–1728. https://doi.org/10.1175/1520-0493(2001)129<1718:VREKF>2.0.CO;2.

Houtekamer, P., Mitchell, H., 2001. A sequential ensemble kalman filter for atmospheric data assimilation. American Meteorological Society 129, 123–137. https://doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2.

Jazwinski, A., 1970. Stochastic Processes and Filtering Theory. Number 64 in Mathematics in science and engineering. Acad. Press, New York, NY [u.a.]).

Jiménez, J.F., 2016. 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.

Kumar, A., Jimenez, R., Belalcazar, L.C., Roja, N.Y., 2016. Application of wrf-chem model to simulate pm10 concentration over bogota. Aerosol and Air Quality Research 16, 1206–1221. https://doi.org/10.4209/aaqr.2015.05.0318, 10.4209/aaqr.2015.05.0318.

Kumar, A., Jimenez, R., Belalcazar, L.C., Rojas, N.Y., 2016. Application of WRF-chem model to simulate PM10 concentration over bogota. Aerosol and Air Quality Research 16, 1206–1221. https://doi.org/10.4209/aaqr.2015.05.0318.

Lateb, M., Meroney, R., Yataghene, M., Fellouah, H., Saleh, F., Boufadel, M., 2016. On the use of numerical modelling for near-field pollutant dispersion in urban environments: a review. Environ. Pollut. 208, 271–283. https://doi.org/10.1016/j.envpol.2015.07.039.

Li, H., Kalnay, E., Miyoshi, T., 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble kalman filte. Q. J. R. Meteorol. Soc. 135, 4523–4533. https://doi.org/10.1002/qj.371.

Manders, A.M., Schaap, M., Hoogerbrugge, R., 2009. Testing the capability of the chemistry transport model LOTOS-EUROS to forecast PM10 levels in The Netherlands. Atmos. Environ. 43, 4050–4059. https://doi.org/10.1016/j.atmosenv.2009.05.006.

Manders, A.M.M., Builtjes, P.J.H., Curier, L., Denier Van Der Gon, H.A.C., Hendriks, C., Jonkers, S., Kranenburg, R., Kuenen, J.J.P., Segers, A.J., Timmermans, R.M.A., Visschedijk, A.J.H., Kruit, R.J.W., Addo, W., Van Pul, J., Sauter, F.J., Van Der Swaluw, E., Swart, D.P.J., Douros, J., Eskes, H., Van Meijgaard, E., Van Ulft, B., Van Velthoven, P., Banzhaf, S., Mues, A.C., Stern, R., Fu, G., Lu, S., Heemink, A., Van Velzen, N., Schaap, M., 2017. Curriculum vitae of the LOTOS-EUROS (v2.0) chemistry transport model. Geosci. Model Dev 10, 4145–4173. https://doi.org/10.5194/gmd-10-4145-2017.

Manders-Groot, A.M.M., Segers, A.J., Jonkers, S., Schaap, M., Timmermans, R., Hendriks, C., Sauter, F., Kruit, R.W., Swaluw, E.V.D., Eskes, H., Banzhaf, S., 2016. LOTOS-EUROS v2.0 Reference Guide. TNO 2016 R10. TNO., Utrecht, The Netherlands. Technical Report.

Nedbor-Gross, R., Henderson, B.H., Pérez-Peña, M.P., Pachón, J.E., 2018. Air quality modeling in Bogotá Colombia using local emissions and natural mitigation factor adjustment for re-suspended particulate matter. Atmospheric Pollution Research 9, 95–104. https://doi.org/10.1016/j.apr.2017.07.004.

Ott, E., Hunt, B.R., Szunyogh, I., Zimin, A.V., Kostelich, E., Corazza, M., Kalnay, E., Patil, D., Yorke, J.A., 2004. A local ensemble Kalman filter for atmospheric data assimilation. Tellus 56, 415–428.

Pachón, J.E., Galvis, B., Lombana, O., Carmona, L.G., Fajardo, S., Rincón, A., Meneses, S., Chaparro, R., Nedbor-Gross, R., Henderson, B., 2018. Development and evaluation of a comprehensive atmospheric emission inventory for air quality modeling in the megacity of Bogotá. Atmosphere 9, 1–17. https://doi.org/10.3390/atmos9020049.

Petrescu, A.M.R., Abad-Viñas, R., Janssens-Maenhout, G., Blujdea, V.N.B., Grassi, G., 2012. Global estimates of carbon stock changes in living forest biomass: EDGARv4.3 - time series from 1990 to 2010. Biogeosciences 9, 3437–3447. https://doi.org/10.5194/bg-9-3437-2012. https://www.biogeosciences.net/9/3437/2012/.

Petrie, R.E., 2008. Localization in the Ensemble Kalman Filter. Master. University of Reading.

Sakov, P., Bertino, L., 2011. Relation between two common localisation methods for the EnKF. Comput. Geosci. 15, 225–237. https://doi.org/10.1007/s10596-010-9202-6.

Sallis, J.F., Bull, F., Burdett, R., Frank, L.D., Griffiths, P., Giles-Corti, B., Stevenson, M., 2016. Use of science to guide city planning policy and practice: how to achieve healthy and sustainable future cities. Lancet 388, 2936–2947. https://doi.org/10.1016/S0140-6736(16)30068-X. http://www.sciencedirect.com/science/article/pii/S014067361630068X.

Sauter, F., der Swaluw, E.V., Manders-groot, A., Kruit, R.W., Segers, A., Eskes, H., 2012. TNO Report TNO-060-UT-2012-01451. TNO, Utrecht, Netherlands. Technical Report.

Thunis, P., Miranda, A., Baldasano, J.M., Blond, N., Douros, J., Graff, A., Janssen, S., Juda-Rezler, K., Karvosenoja, N., Maffeis, G., Martilli, A., Rasoloharimahefa, M., Real, E., Viaene, P., Volta, M., White, L., 2016. Overview of current regional and local scale air quality modelling practices: Assessment and planning tools in the EU. Environ. Sci. Pol. 65, 13–21. https://doi.org/10.1016/j.envsci.2016.03.013.

Toro, M.V., Cremades, L., Ramírez-Echeverry, J., 2005. Inventario de emisiones biogenicas en el valle de aburrá. Revista ingeniería y gestión ambiental 1, 31–39.

Toro, M.V., Cremades, L.V., Calbó, J., 2006. Relationship between VOC and NOxemissions and chemical production of tropospheric ozone in the Aburrá Valley (Colombia). Chemosphere 65, 881–888. https://doi.org/10.1016/j.chemosphere.2006.03.013.

van Velzen, N., Segers, A.J., 2010. A problem-solving environment for data assimilation in air quality modelling. Environ. Model. Software 25, 277–288. https://doi.org/10.1016/j.envsoft.2009.08.008.

Verlaan, M., Heemink, A.W., 1997. Tidal flow forecasting using reduced rank square root filters. Stoch. Hydrol. Hydraul. 11, 349–368. https://doi.org/10.1007/BF02427924.

Whitaker, J.S., Hamill, T.M., 2002. Ensemble data assimilation without perturbed observations. Mon. Weather Rev. 130, 1913–1924. https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.

Yu, S., Eder, B., Dennis, R., Chu, S.H., Schwartz, S.E., 2006. New unbiased symmetric metrics for evaluation of air quality models. Atmos. Sci. Lett. 7, 26–34. https://doi.org/10.1002/asl.125 arXiv: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/asl.125 https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/asl.125.

Zarate, E., Carlos Belalcazar, L., Clappier, A., Manzi, V., Van den Bergh, H., 2007. Air quality modelling over Bogota, Colombia: combined techniques to estimate and evaluate emission inventories. Atmos. Environ. 41, 6302–6318. https://doi.org/10.1016/j.atmosenv.2007.03.011.