Improvement of the PM_{10} estimations of the air quality model AURORA by means of Kalman filtering techniques

Oscar Mauricio Agudelo, Peter Viaene and Bart De Moor

***Abstract*— This paper presents the results of using two data assimilation techniques, the Ensemble Kalman Filter (EnKF) and the Deterministic Ensemble Kalman Filter (DEnKF), for improving the PM_{10} estimations of the air quality model AURORA. These filters are set up around a stochastic formulation of the model where some of its parameters are assumed uncertain. These uncertainties are considered the main cause of the discrepancies between the model estimations and the PM_{10} ground-based measurements. The measurements have been provided by IRCEL and the region on which the assimilation experiments are performed includes Belgium, Luxemburg, and some small parts of France, the Netherlands and Germany. The simulation results show that both filters manage to significantly reduce the error of the PM_{10} estimations, which indirectly leads to an improvement of the $PM_{2.5}$ field.**

I. INTRODUCTION

Particulate Matter (PM) is the term used for solid or liquid particles found in the air. Depending on the size of the particles we can distinguish two groups: $PM_{2.5}$ (up to 2.5 micrometers in aerodynamic diameter) and PM_{10} (up to 10 micrometers in aerodynamic diameter). The inhalation of these particles has harmful effects on human health (such as asthma, lung cancer, cardiovascular issues, birth defects, and premature death [1]) and therefore accurate estimation of PM is of vital importance. However, air quality models typically fail to provide accurate estimations of PM given the complexity of the mechanisms involved and the uncertainties in the model parameters and model inputs such as boundary conditions, meteorological patterns, emissions, etc. A way to tackle this problem is by using data assimilation techniques. These techniques combine the outputs of the numerical model with observational data in order to improve the quality of the estimations. Among the existing techniques, the Ensemble Kalman Filter (EnKF) [2] has been widely used due to its capability of handling nonlinear large-scale systems and its relative easy implementation when compared to other approaches. The EnKF is a sequential data assimilation technique that uses Monte Carlo or ensemble integration. This technique has been successfully used in meteorology, oceanography, space weather forecast and also in the air-quality field [3]–[6]. Several authors have pointed out that the use of perturbed observations in the traditional EnKF leads to a suboptimal filter behavior, especially when the number of ensembles is small. Hence, several ensemble-based filters that do not use perturbed observations have been proposed, like for example the Deterministic Ensemble Kalman Filter (DEnKF) [7]. Although in [7] this technique was successfully applied to three small models with different characteristics, the DEnKF still needs to be tested on large-scale nonlinear systems like for instance air quality models.

In a previous work [6], the authors successfully used the EnKF to improve the ozone estimations of the air quality model AURORA. However in that study, AURORA was set with a very coarse 3D grid ($11\times15\times15$ grid-cells) due to computational limitations and to speed up the test of different settings for the EnKF. The use of coarse grids may artificially diffuse the emissions, leading to significant errors in the pollutant concentrations, especially those that are formed via nonlinear chemical reactions. In addition, coarse grids may result in large numerical errors [8].

In this paper we apply the EnKF and DEnKF algorithms to an AURORA model setup with a much finer 3D grid ($59\times49\times22$ grid-cells) in order to improve its PM_{10} estimations over a period where a PM_{10} pollution episode took place. Unlike the scenario addressed in [6] where the model estimations compare relatively well to the ozone measurements, in this study there is a big gap between the PM_{10} observations and the estimations made by the model, which poses a serious challenge for the data assimilation methods. Furthermore in this work we evaluate the benefits of using the DEnKF over the EnKF when applied to a much larger and complex model than those used in [7].

This paper is organized in the following way. In Section II, we present a brief description of AURORA as well as the network of air quality stations used for assimilation and validation purposes. Section III presents the derivation of the stochastic model of AURORA on which the assimilation schemes are based. This section also describes the EnKF and DEnKF techniques. The details of the assimilation experiments as well as the simulation results are presented in Section IV, and in Section V some concluding remarks are given.

II. DESCRIPTION OF AURORA

AURORA (Air quality modeling in Urban Regions using an Optimal Resolution Approach) is a three-dimensional Eulerian model developed by VITO (Flemish Institute for Technological Research) for simulating air pollution in the lower troposphere at urban or regional scale [9]. AURORA includes modules that describe (i) the transport by diffusion and advection of a number of trace gas components

O. M. Agudelo and B. De Moor are with the Department of Electrical Engineering (ESAT), Research Group SCD-SISTA, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium. E-mail: {mauricio.agudelo, bart.demoor}@esat.kuleuven.be

P. Viaene is with the Environmental Modeling Unit, VITO - Flemish Institute for Technological Research, Boeretang 200, BE-2400 Mol, Belgium. E-mail: peter.viaene@vito.be

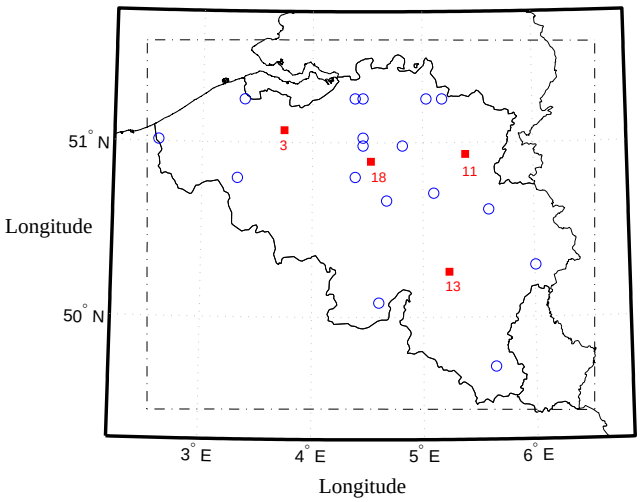

Fig. 1. Geographical location of the stations measuring PM_{10}. The blue circles and red squares represent assimilation and validation stations respectively. The dash-dot line delimits the area covered by AURORA.

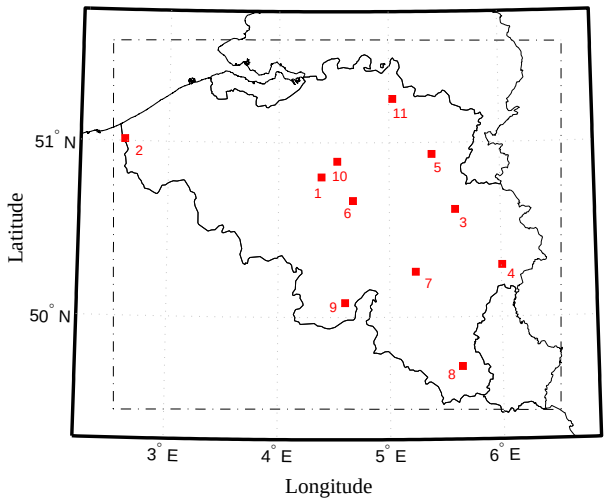

Fig. 2. Geographical location of the $PM_{2.5}$ stations (red squares). The dash-dot line delimits the area covered by AURORA.

emitted both by anthropogenic and biogenic sources (ii) the (photo)chemistry of this gas mixture resulting in formation of secondary pollutants and (iii) the loss of these components by deposition. The thermodynamic equilibrium between the gas phase and the inorganic aerosol is modeled using a module based on ISORROPIA [10]. Both the coarse (particles that have an aerodynamic diameter between 2.5 and 10 micrometers) and fine (particles with an aerodynamic diameter smaller than 2.5 micrometers) fractions of the species that form PM_{10} are modeled separately in AURORA. These species include both primary (emitted) and secondary (formed from others components) particles. The primary particles considered in AURORA are elementary carbon, and primary particulate matter (e.g., dust or sea salt). The secondary particles include inorganic particles consisting of ammonium, nitrate and sulphate, and secondary organic aerosols which are formed by condensation of oxidation products with a lower volatility.

In this study AURORA has been configured to cover a region that includes Belgium, Luxemburg, and some parts of the Netherlands, France and Germany (see Figure 1 or Figure 2). The horizontal domain was divided into 59×49 grid-cells with a resolution of 5 km. Regarding the vertical domain, 22 layers were used to span an altitude of 2000 m. The model was set up for simulating a period of 12 days starting on January 20th, 2010. This period is of special interest since within it a PM_{10} episode took place, that is, elevated PM_{10} concentrations were registered.

The air quality data for carrying out the assimilation experiments were provided by IRCEL, the Belgian Interregional Environment Agency [11]. Since the spatial resolution of the model is 5 km, only measurements from rural and urban-background stations were deemed representative and taken into account. In Figure 1 we can see the location of the PM_{10} air quality stations used in this work. These 21 stations have been split into two groups: (i) assimilation stations and (ii) validation stations. The first group is used in the assimilation process to obtain the optimal estimate of the state. The second group is not used in the assimilation, but only to verify the results. The split have been carried out in such a way that the validation stations are evenly distributed along the spatial domain and surrounded by neighboring assimilation stations. The number of assimilation and validation stations is 17 and 4 respectively.

Given that $PM_{2.5}$ is part of PM_{10}, we also evaluate the impact of assimilating PM_{10} measurements by checking the improvement in the estimation of the $PM_{2.5}$ concentrations. Again the data were provided by IRCEL, and due to the resolution of the model we only considered measurements coming from rural and urban-background stations. Figure 2 shows the 11 stations finally selected.

III. Data Assimilation Schemes

The deterministic model AURORA can be compactly written as follows:

$$\mathbf{c}(k+1) = \mathbf{M}_{\mathrm{A}}(\mathbf{c}(k)) \tag{1}$$

where $\mathbf{c}(k)$ is the state vector containing the concentrations of different pollutants at every grid-cell, and $\mathbf{M}_{\mathrm{A}}(\cdot)$ is the nonlinear state-space operator that describes the time evolution of $\mathbf{c}(k)$ from k to $k+1$.

In order to use the EnKF and DEnKF algorithms with AURORA, we first need to define a stochastic model of the system that accounts for the model and measurement errors. The stochastic model used in this study is the one that was derived in [6], where ozone observations were assimilated. In order to keep this paper self contained, in the next section we present the derivation of this stochastic model, and afterwards we describe the data assimilation algorithms.

A. Stochastic representation of AURORA

AURORA was extended so that uncertain inputs such as boundary conditions and emissions, and uncertain model pa-

rameters such as deposition among others can be multiplied by nonnegative correction factors. Each of these factors is modeled in stochastic way as follows:

$$f(k) = \max\left(0, 1 + \lambda(k)\right) \tag{2}$$

where $f(k)$ is the correction factor and $\lambda(k)$ is a colored noise process whose dynamics is described by

$$\lambda(k+1) = \alpha\lambda(k) + \left(\sigma\sqrt{1-\alpha^2}\right) w(k), \tag{3}$$

where $w(k) \sim N(0,1)$, $\alpha \in [0\ 1]$ is the time correlation parameter and σ is the standard deviation. If we set α to zero, then we get a white noise sequence with zero mean and standard deviation σ. Now, if we set α to one, then the colored noise process is reduced to a constant value. The covariance function of $\lambda(k)$ is parameterized by $\alpha = e^{(-T_\mathrm{s}/\tau)}$ [12], where τ is the time correlation length and T_s is the time between two consecutive samples, which in our case is equal to 1 hour. We use colored noise for modeling the factors in order to avoid the introduction of rapid fluctuations that take place when only white noise is used.

In a similar fashion as it was done in [12] and [3], the stochastic model of AURORA is built by augmenting the state vector of (1) with the vector $\boldsymbol{\lambda}(k) = [\lambda_1(k), \lambda_2(k) \ldots, \lambda_n(k)]^T$ containing the colored noise process of every uncertain factor. Thus this model is given by

$$\begin{bmatrix} \mathbf{c}(k+1) \\ \boldsymbol{\lambda}(k+1) \end{bmatrix} = \begin{bmatrix} \mathbf{M}_\mathrm{A}\left(\mathbf{c}(k), \boldsymbol{\lambda}(k)\right) \\ \boldsymbol{\Psi}\boldsymbol{\lambda}(k) \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\Sigma} \end{bmatrix} \mathbf{w}(k) \tag{4}$$

with

$$\begin{aligned} &\mathbf{w}(k) \in \mathbb{R}^n \sim N(\mathbf{0}, \mathbf{I}_n), \\ &\boldsymbol{\Psi} \in \mathbb{R}^{n\times n} = \mathrm{diag}\left([\alpha_1, \alpha_2, \ldots, \alpha_n]\right), \\ &\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n} = \mathrm{diag}\Big(\Big[\sigma_1\sqrt{1-\alpha_1^2}, \ldots, \sigma_n\sqrt{1-\alpha_n^2}\Big]\Big), \end{aligned}$$

where n is the number of uncertain factors, $\alpha_1, \alpha_2, \ldots, \alpha_n$ and $\sigma_1, \sigma_2, \ldots, \sigma_n$ are the time correlation parameters and the standard deviations respectively of the colored noise processes associated to the factors. By defining $\mathbf{x}(k) = \left[\mathbf{c}(k)^T, \boldsymbol{\lambda}(k)^T\right]^T$ as the new augmented state vector, we can write Equation (4) more compactly as

$$\mathbf{x}(k+1) = \mathbf{M}\left(\mathbf{x}(k)\right) + \mathbf{G}\mathbf{w}(k) \tag{5}$$

where $\mathbf{M}(\cdot)$ is the nonlinear operator that describes the time evolution of $\mathbf{x}(k)$ from time k to $k+1$ and $\mathbf{G}\mathbf{w}(k)$ is the stochastic forcing term. The relation between the model state $\mathbf{x}(k)$ and the measurements $\mathbf{y}(k)$ is explained by means of the following expression:

$$\mathbf{y}(k) = \mathbf{C}(k)\mathbf{x}(k) + \mathbf{v}(k), \quad \mathbf{v}(k) \sim N\left(\mathbf{0}, \mathbf{R}(k)\right). \tag{6}$$

Here, $\mathbf{C}(k)$ is a linear observation operator that maps the states that form PM_{10} (coarse and fine fractions of primary and secondary particles) at the bottom layer of AURORA to the ground-based measurements from IRCEL, and $\mathbf{v}(k)$ is a Gaussian white noise vector with covariance $\mathbf{R}(k)$ that accounts for the errors of the measurements. Given that the stations are not providing valid measurements all the time, $\mathbf{R}(k)$, $\mathbf{C}(k)$ and $\mathbf{v}(k)$ must be properly adjusted within the assimilation algorithms for coping with each particular situation.

B. Ensemble Kalman Filter

The Ensemble Kalman Filter (EnKF) is a Monte Carlo method where the probability density of the state estimate is represented by a finite ensemble of N possible states, $\mathbf{x}_1(k), \mathbf{x}_2(k), \ldots, \mathbf{x}_N(k)$. This representation makes possible the application of the filter to large-scale systems where the explicit storage and manipulation of the state error covariance are impossible or impractical. In the EnKF formulation, it is assumed that every ensemble member is a single sample taken from the probability distribution of the true state. Statistical information such as mean and error covariances can be approximated with sample statistics whenever such information is required. The EnKF algorithm for the model (5)-(6) can be summarized as follows [2]:

Initialization

First, generate an initial ensemble $\mathbf{x}_1^\mathrm{a}(k-1), \mathbf{x}_2^\mathrm{a}(k-1), \ldots, \mathbf{x}_N^\mathrm{a}(k-1)$ that properly represent the error statistics of the initial guess $\mathbf{x}(k=0)$ for the model state.

for $k = 1, \ldots$

Forecast step

- Update every ensemble member using (5),

$$\mathbf{x}_i^\mathrm{f}(k) = \mathbf{M}\left(\mathbf{x}_i^\mathrm{a}(k-1)\right) + \mathbf{G}\mathbf{w}_i(k-1)$$

$$\mathbf{w}_i(k-1) \sim N(\mathbf{0}, \mathbf{I}), \quad \forall i = 1, 2, \ldots, N$$

- Calculate the ensemble of anomalies or perturbations,

$$\mathbf{A}^\mathrm{f}(k) = \left[\boldsymbol{\xi}_1^\mathrm{f}(k), \boldsymbol{\xi}_2^\mathrm{f}(k), \ldots, \boldsymbol{\xi}_N^\mathrm{f}(k)\right]$$

$$\boldsymbol{\xi}_i^\mathrm{f}(k) = \mathbf{x}_i^\mathrm{f}(k) - \bar{\mathbf{x}}^\mathrm{f}(k), \quad \bar{\mathbf{x}}^\mathrm{f}(k) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i^\mathrm{f}(k)$$

Analysis step

- Compute the Kalman gain

$$\mathbf{K}(k) = \mathbf{P}^\mathrm{f}(k)\mathbf{C}(k)^T \left(\mathbf{C}(k)\mathbf{P}^\mathrm{f}(k)\mathbf{C}(k)^T + \mathbf{R}(k)\right)^{-1}$$

$$\mathbf{P}^\mathrm{f}(k) = \left(1/(N-1)\right)\mathbf{A}^\mathrm{f}(k)\left(\mathbf{A}^\mathrm{f}(k)\right)^T$$

- Update $\mathbf{x}_i^\mathrm{f}(k)$ to $\mathbf{x}_i^\mathrm{a}(k)$, for $i = 1, 2, \ldots, N$

$$\mathbf{x}_i^\mathrm{a}(k) = \mathbf{x}_i^\mathrm{f}(k) + \mathbf{K}(k)\left(\mathbf{y}(k) + \mathbf{v}_i(k) - \mathbf{C}(k)\mathbf{x}_i^\mathrm{f}(k)\right)$$

$$\mathbf{v}_i(k) \sim N\left(\mathbf{0}, \mathbf{R}(k)\right)$$

- Calculate the state estimation (the analysis ensemble mean)

$$\bar{\mathbf{x}}^\mathrm{a}(k) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i^\mathrm{a}(k)$$

end for

where $\mathbf{y}(k)$ is a vector comprising the available measurements at time k.

C. Deterministic Ensemble Kalman Filter

It is well known that the fact of using perturbed observations ($\mathbf{y}(k) + \mathbf{v}_i(k)$) in the traditional EnKF leads to a

suboptimal filter behavior, especially when the number of ensembles is small. In [7], a simple modification of the EnKF that does not use perturbed observations is proposed. This modified algorithm is referred to as Deterministic Ensemble Kalman Filter (DEnKF). This filter is based on the recognition that in the case of small corrections to the forecast, the traditional EnKF without perturbed observations reduces the forecast error covariance matrix $\mathbf{P}^{\mathrm{f}}(k)$ by an amount that is roughly twice as large as the amount that it should be. In other words, it underestimates the analysis error covariance matrix $\mathbf{P}^{\mathrm{a}}(k)$ by a factor of two. Hence in the analysis step of the DEnKF, the ensemble mean and the ensemble of anomalies are updated separately; the mean is updated by the standard analysis equation and the anomalies by means of the same equation but using half of the Kalman gain. The DEnKF algorithm for the model (5)-(6) is as follows [7]:

Initialization

First, generate an initial ensemble $\mathbf{x}_1^{\mathrm{a}}(k-1), \mathbf{x}_2^{\mathrm{a}}(k-1), \ldots, \mathbf{x}_N^{\mathrm{a}}(k-1)$ that properly represent the error statistics of the initial guess $\mathbf{x}(k=0)$ for the model state.

for $k = 1, \ldots$

Forecast step

- Update every ensemble member using (5),

$$\mathbf{x}_i^{\mathrm{f}}(k) = \mathbf{M}\left(\mathbf{x}_i^{\mathrm{a}}(k-1)\right) + \mathbf{G}\mathbf{w}_i(k-1)$$

$$\mathbf{w}_i(k-1) \sim N(\mathbf{0}, \mathbf{I}), \quad \forall i = 1, 2, \ldots, N$$

- Calculate the ensemble of anomalies or perturbations,

$$\mathbf{A}^{\mathrm{f}}(k) = \left[\boldsymbol{\xi}_1^{\mathrm{f}}(k), \boldsymbol{\xi}_2^{\mathrm{f}}(k), \ldots, \boldsymbol{\xi}_N^{\mathrm{f}}(k)\right]$$

$$\boldsymbol{\xi}_i^{\mathrm{f}}(k) = \mathbf{x}_i^{\mathrm{f}}(k) - \bar{\mathbf{x}}^{\mathrm{f}}(k), \quad \bar{\mathbf{x}}^{\mathrm{f}}(k) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i^{\mathrm{f}}(k)$$

Analysis step

- Compute the Kalman gain

$$\mathbf{K}(k) = \mathbf{P}^{\mathrm{f}}(k)\mathbf{C}(k)^T \left(\mathbf{C}(k)\mathbf{P}^{\mathrm{f}}(k)\mathbf{C}(k)^T + \mathbf{R}(k)\right)^{-1}$$

$$\mathbf{P}^{\mathrm{f}}(k) = \left(1/(N-1)\right)\mathbf{A}^{\mathrm{f}}(k)\left(\mathbf{A}^{\mathrm{f}}(k)\right)^T$$

- Calculate the state estimation (analysis)

$$\bar{\mathbf{x}}^{\mathrm{a}}(k) = \bar{\mathbf{x}}^{\mathrm{f}}(k) + \mathbf{K}(k)\left(\mathbf{y}(k) - \mathbf{C}(k)\bar{\mathbf{x}}^{\mathrm{f}}\right)$$

- Compute the analyzed anomalies or perturbations

$$\mathbf{A}^{\mathrm{a}}(k) = [\boldsymbol{\xi}_1^{\mathrm{a}}(k), \ldots, \boldsymbol{\xi}_N^{\mathrm{a}}(k)] = \mathbf{A}^{\mathrm{f}}(k) - \frac{1}{2}\mathbf{K}(k)\mathbf{C}(k)\mathbf{A}^{\mathrm{f}}(k)$$

- Calculate the analyzed ensemble

$$\mathbf{x}_i^{\mathrm{a}}(k) = \boldsymbol{\xi}_i^{\mathrm{a}}(k) + \bar{\mathbf{x}}^{\mathrm{a}}(k), \quad \forall i = 1, 2, \ldots, N$$

end for

where $\mathbf{y}(k)$ is the vector that contains the available measurements at time k.

One of the advantages of the assimilation schemes used along this work is that it is not required to linearize the model, since the ensembles are propagated by using the original nonlinear model operator. In addition, it is not necessary to explicitly compute the forecast error covariance matrix $\mathbf{P}^{\mathrm{f}}(k)$ in order to calculate the Kalman gain. This fact allows these techniques to be applied to large-scale systems.

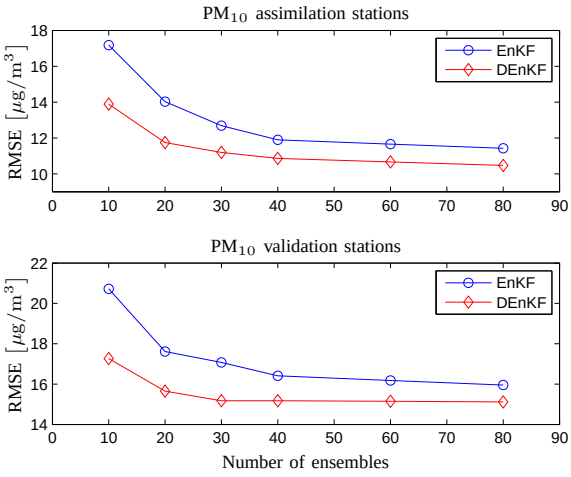

Fig. 3. Effect of the number of ensembles on the performance of the data assimilation schemes. RMSE calculated by using Equation (7).

For most practical applications, the most time consuming part of the EnKF and DEnKF algorithms is the evaluation of the ensembles in the model operator $\mathbf{M}(\cdot)$. Hence, the computational load is approximately N model evaluations. The errors in the state estimate are of statistical nature and decrease slowly with the number of ensemble members.

IV. Assimilation Results

As it was discussed in Section III-A, the stochastic model of AURORA is built by considering some correction factors uncertain, and consequently the parameters and/or inputs that are multiplied by them. In this work, we have considered uncertain the factors associated to the north, south, west, and east boundary conditions of PM_{10}, the deposition of PM_{10}, and the PM_{10}, NH_3 and SO_2 emissions. All these factors have a significant impact on the PM_{10} concentration field. For all the uncertain factors, the standard deviation σ of the colored noise process was set to 0.4 and the time correlation length τ was fixed to 6 hours.

Regarding the PM_{10} observations provided by IRCEL, we considered an error of 20% of the measured concentration with a minimum value of $1\mu\mathrm{g/m}^3$.

Covariance inflation is a heuristic technique which is used for avoiding filter divergence due to underestimation of the forecast error covariance matrix $\mathbf{P}^{\mathrm{f}}(k)$ [13]. It consists of multiplying the ensemble of anomalies $\mathbf{A}^{\mathrm{f}}(k)$ by an inflation factor r, which is slightly larger than 1. This mechanism was implemented in the data assimilation algorithms by adding the following line at the end of the forecast step:

$$\mathbf{A}^{\mathrm{f}}(k) \leftarrow r\mathbf{A}^{\mathrm{f}}(k),$$

where $\leftarrow$ represents the replacement of the previous value. In general, the choice of the inflation factor for a given experiment is to some extend heuristic in nature. After

TABLE I

PERFORMANCE OF THE DATA ASSIMILATION TECHNIQUES WHEN PM_{10} OBSERVATIONS ARE ASSIMILATED

		PM_{10} stations		$PM_{2.5}$ stations
		Assimilation	Validation	
RMSE of AURORA		33 $\mu g/m^3$	34.2636 $\mu g/m^3$	29.5165 $\mu g/m^3$
EnKF	RMSE	11.4212 $\mu g/m^3$	15.9505 $\mu g/m^3$	15.3891 $\mu g/m^3$
	Error reduction	65.39 %	53.45 %	47.86 %
DEnKF	RMSE	10.4703 $\mu g/m^3$	15.1164 $\mu g/m^3$	13.8644 $\mu g/m^3$
	Error reduction	68.27 %	55.88 %	53.03 %

RMSE calculated by using Equation (7). Number of ensembles: 80.

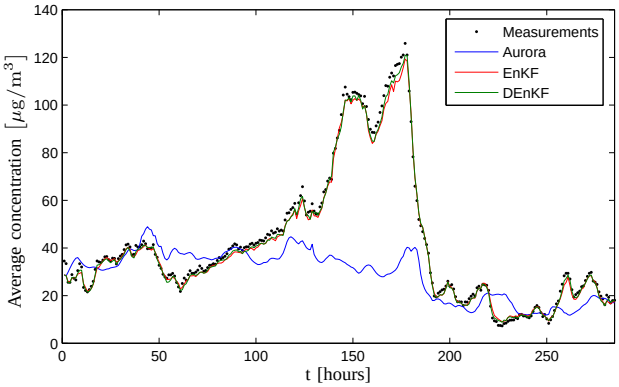

Fig. 4. Average of the PM_{10} concentration over the assimilation stations. Starting date: January 20th, 2010.

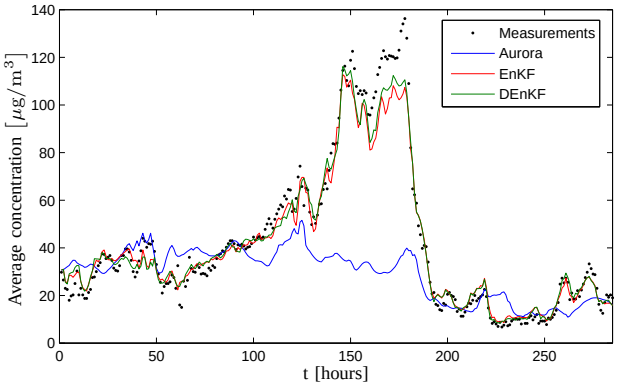

Fig. 5. Average of the PM_{10} concentration over the validation stations. Starting date: January 20th, 2010.

several experiments we fixed the value of r to 1.08.

A 12 days period starting on January 20th, 2010 at midnight was selected to carry out the assimilation experiments. As it was stated in Section II, this period was chosen because during that time a serious PM_{10} episode took place and AURORA was not able to describe it properly.

In order to evaluate the performance of the filters, the Root Mean Square Error (RMSE) between the observations and the estimations made by each of them is computed,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_\mathrm{s} N_\mathrm{h}} \sum_{i=1}^{N_\mathrm{s}} \sum_{k=1}^{N_\mathrm{h}} \left(y_i^\mathrm{M}(k) - y_i^\mathrm{O}(k) \right)^2} \qquad (7)$$

where N_s and N_h are the number of stations and the number of hours respectively, and $y_i^\mathrm{M}(k)$ and $y_i^\mathrm{O}(k)$ are the estimated and measured concentrations of the ith station at time k. In Figure 3 we can observe the effect of the ensemble size on the performance of the data assimilation methods. It is clear that the estimation errors of the DEnKF are smaller than the ones of the EnKF, especially when the number of ensembles is small. From the plots shown in this figure we can see that 30 or 40 ensembles offer a good trade-off between accuracy and computational burden. Despite this and looking for presenting the most accurate estimations, in the remainder of this section we will show some of the results obtained when 80 ensemble members were used.

The assimilation experiments were performed by using the infrastructure of the Flemish Supercomputer Center (VSC). 10 Nehalem compute nodes were used, each one with 2 quad-core Xeon 5560 processors and 24 GB of RAM memory. Thus, the evaluation of every ensemble member in $\mathbf{M}(\cdot)$ was carried out in parallel; each ensemble member was assigned to a different processor core. In average, every assimilation experiment took 8 hours and a half.

Table I shows the RMSEs between the observations and the estimations made by both the model and the data assimilation schemes. It is clear that a large error reduction is obtained by using the EnKF and DEnKF algorithms. For the case of the EnKF, the error of the PM_{10} estimations is decreased by 65.39% in the assimilation stations and by 53.45% in the validation stations. For the DEnKF the error in the assimilation and validation stations is reduced by 68.27% and 55.88% respectively. Furthermore, from this table we can see that the assimilation of PM_{10} observations leads to a notably error reduction in the $PM_{2.5}$ estimations.

In Figures 4 and 5 we can observe the average of the PM_{10} concentration over the assimilation and validation stations respectively. Notice that AURORA is not able to properly estimate the data. Despite the big gap between AURORA and the measurements, both data assimilation techniques manage to notably reduce this gap. Figure 6 shows the RMSE of the PM_{10} validation stations along the period of interest. This RMSE has been computed in the following manner:

$$\mathrm{RMSE}_\mathrm{station} = \sqrt{\frac{1}{N_\mathrm{h}} \sum_{k=1}^{N_\mathrm{h}} \left(y^\mathrm{M}(k) - y^\mathrm{O}(k) \right)^2} \qquad (8)$$

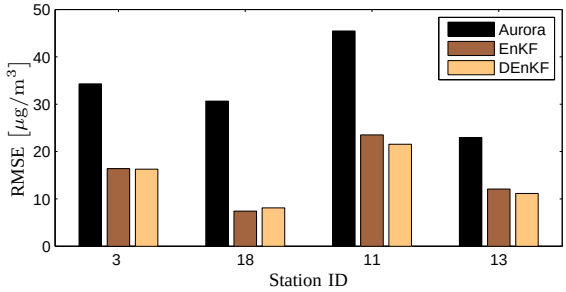

Fig. 6. Root mean square error of AURORA and the data assimilation techniques at the PM_{10} validation stations.

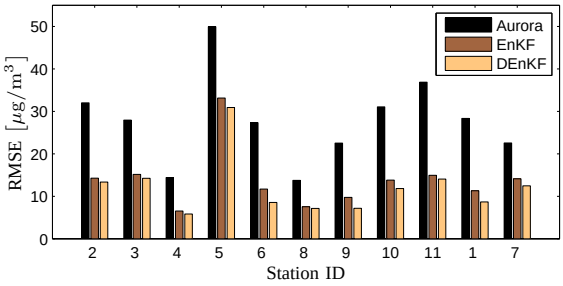

Fig. 7. Root mean square error of AURORA and the data assimilation schemes at the $PM_{2.5}$ stations.

where N_{h} is the number of hours, and $y_i^{\mathrm{M}}(k)$ and $y_i^{\mathrm{O}}(k)$ are the estimated and measured concentrations respectively of a given station at time k. Keep in mind that we have used the validation stations only for verifying the assimilation results and evaluating the impact of the data assimilation on the covered region. From Figure 6, it is evident that both assimilation techniques lead to more accurate PM_{10} estimations. Finally from Figure 7 we can evaluate the impact of assimilating PM_{10} measurements on the $PM_{2.5}$ estimations. Both the EnKF and the DEnKF remarkably reduce the estimation error at the $PM_{2.5}$ air quality stations.

V. Conclusions

In this paper we have presented the results of using the EnKF and the DEnKF for improving the quality of the PM_{10} estimations of the air quality model AURORA. The assimilation process has been driven by ground-based measurements provided by IRCEL. We have described the stochastic formulation of AURORA on which the EnKF and the DEnKF are based. This stochastic model has been constructed by assuming uncertain some model parameters/inputs that have a significant impact on the PM_{10} field. Both data assimilation techniques managed to significantly reduce the error in the PM_{10} estimations. Given that $PM_{2.5}$ is highly correlated with PM_{10}, the $PM_{2.5}$ field was indirectly improved by the data assimilation techniques. From this study, it is clear that the DEnKF provides better estimations than the EnKF, especially when the number of ensemble members is small.

Future work will be focused on performing assimilation experiments for longer periods, e.g., 1 year, while keeping the current spatial domain and the resolution of 5 km.

VI. Acknowledgments

For performing the assimilation experiments we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government - department EWI. This research has been supported by: • Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC) and PFV/10/002 (OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; • Flemish Government: ○ FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC); ○ IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; • Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); • IBBT • EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940), FP7-SADCO (MC ITN-264735), ERC HIGHWIND (259 166) • Contract Research: AMINAL • Other: ○ Helmholtz: viCERP ○ ACCM. Dr. Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. The scientific responsibility is assumed by its authors.

References

[1] C. A. Pope III, R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito, and G. D. Thurston, "Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution," *J. Am. Med. Assoc. (JAMA)*, vol. 287, no. 9, pp. 1132–1141, 2002.

[2] G. Burgers, P. Jan van Leeuwen, and G. Evensen, "Analysis scheme in the Ensemble Kalman Filter," *Monthly Weather Review*, vol. 126, no. 6, pp. 1719–1724, 1998.

[3] A. Heemink and A. Segers, "Modelling and prediction of environmental data in space and time using Kalman filtering," *Stochastic Environmental Research and Risk Assessment*, vol. 16, no. 3, pp. 225–240, 2002.

[4] M. van Loon, P. Builtjes, and A. Segers, "Data assimilation of ozone in the atmospheric transport chemistry model Lotus," *Environmental Modelling & Software*, vol. 15, pp. 603–609, 2000.

[5] G. Candiani, C. Carnevale, G. Finzi, E. Pisoni, and M. Volta, "Off-line data assimilation to provide the best estimate of tropospheric ozone concentrations by means of EnKF," in *49th IEEE Conference on Decision and Control*, USA, December 2010, pp. 3493–3498.

[6] O. M. Agudelo, O. Barrero, P. Viaene, and B. De Moor, "Assimilation of ozone measurements in the air quality model AURORA by using the Ensemble Kalman Filter," in *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC 2011)*, Orlando, FL, USA, December 2011, pp. 4430–4435.

[7] P. Sakov and P. R. Oke, "A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters," *Tellus*, vol. 60A, pp. 361–371, 2008.

[8] M. T. Odman, M. N. Khan, R. K. Srivastava, and D. S. McRae, "Initial application of the adaptive grid air pollution model." in *25th NATO/CCMS International Technical Meeting on Air Pollution Modelling and Its Application*, Louvain-la-Neuve, Belgium, October 2001, pp. 15–19.

[9] C. Mensink, K. De Ridder, N. Lewyckyj, L. Delobbe, L. Janssen, and P. Van Haver, "Computational aspects of air quality modelling in urban regions using an optimal resolution approach (AURORA)," in *Large-Scale Scientific Computing, Lecture Notes in Computer Science*, S. Margenov, J. Wasniewski, and P. Yamalov, Eds., vol. 2179. Springer Berlin / Heidelberg, 2001, pp. 299–308.

[10] A. Nenes, S. N. Pandis, and C. Pilinis, "ISORROPIA: A new thermodynamic equilibrium model for multiphase multicomponent inorganic aerosols," *Aquatic Geochemistry*, vol. 4, no. 1, pp. 123–152, 1998.

[11] IRCEL, "Intergewestelijke cel voor het leefmilieu (Belgian interregional environment agency)," 2011, [online] http://www.irceline.be/.

[12] A. Barbu, A. Segers, M. Schaap, A. Heemink, and P. Builtjes, "A multi-component data assimilation experiment directed to sulphur dioxide and sulphate over Europe," *Atmospheric Environment*, vol. 43, pp. 1622–1631, 2009.

[13] J. L. Anderson and S. L. Anderson, "A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts," *Monthly Weather Review*, vol. 127, pp. 2741–2758, 1999.