BLUE-based NO2 data assimilation at urban scale

Anne Tilloy, Vivien Mallet, David Poulet, Céline Pesin, Fabien Brocheton

BLUE-based NO_2 data assimilation at urban scale

Anne Tilloy[1,2], Vivien Mallet[1,2], David Poulet[3], Céline Pesin[3], Fabien Brocheton[3]

Anne Tilloy, INRIA, Domaine de Voluceau, 78150 Rocquencourt, France. (Anne.Tilloy@inria.fr)

Vivien Mallet, INRIA, Domaine de Voluceau, 78150 Rocquencourt, France. (Vivien.Mallet@inria.fr)

David Poulet, Numtech, 6 allée Alan Turing, 63175 Aubière, France. (poulet@numtech.fr)

Céline Pesin, Numtech, 6 allée Alan Turing, 63175 Aubière, France. (pesin@numtech.fr)

Fabien Brocheton, Numtech, 6 allée Alan Turing, 63175 Aubière, France. (brocheton@numtech.fr)

[1]INRIA, Paris-Rocquencourt research center, France

[2]CEREA, joint laboratory École des Ponts ParisTech - EDF R&D, Université Paris-Est, Marne la Vallée, France

[3]Numtech, Aubière, France

Abstract. We aim at optimally combining air quality computations, from the Gaussian model ADMS Urban, and ground observations at urban scale. An ADMS simulation generated NO_2 concentration fields across Clermont-Ferrand (France) down to street level, every three hours for the full year 2008. A monitoring network composed of nine fixed stations provided hourly observations to be assimilated. Every three hours, we compute the so-called BLUE (best linear unbiased estimator), which is a concentration field merging ADMS outputs and ground observations. Its error variance is supposed to be minimal under given assumptions regarding the errors on observations and model simulations. A key step lies in the modeling of error covariances between the computed NO_2 concentrations across the city. We introduce a parameterized covariance which heavily relies on the road network. The covariance between two locations depends on the distance of each location to the road network and on the distance between the locations along the road network. Efficient parameters for the covariances are primarily chosen according to prior assumptions, χ^2 diagnosis and leave-one-out cross-validations. According to the cross-validations, the improvements due to the assimilation seem moderate far from the observation network, but the root mean square error roughly decreases by 30% to 50% in the main city where the station density is high. The method is computationally tractable for the generation of improved concentration fields over a long period, or for day-to-day forecasts.

1. Introduction

Drivers, cyclists and pedestrians are mainly exposed to nitrogen dioxide and particles, especially originating from traffic exhausts. The nitrogen dioxide is a strong oxidizer which can lead to harmful effects on airways. The exceedance of given thresholds can raise problems for asthmatics. The particles have short term and long term effects on respiratory and cardiovascular systems, especially on children, asthmatics and old people. In recent years, there has been a growing interest for the numerical simulation of air quality at urban scale, aiming at the estimation of atmospheric pollutant concentrations in all urban areas, down to street level. One motivation is to improve the evaluation of exposure of the considerable urban population.

In order to estimate the concentrations of main urban pollutants, one can rely on both field observations and model simulations. Air quality monitoring stations provide accurate information at a few locations over a city and for a few pollutants, while the numerical simulations deliver less accurate concentrations at virtually any outdoor place and for a range of pollutants. Data assimilation can be employed to combine these two sources of information in order to better estimate the chemical state of the atmosphere.

Data assimilation has been applied in the air quality community, mostly at large scale and with Eulerian chemistry-transport models [e.g., *Elbern and Schmidt*, 2001; *Segers*, 2002; *Chai et al.*, 2007; *Wu et al.*, 2008]. In this paper, we address the assimilation of observations of an urban monitoring network in order to correct the concentrations of nitrogen dioxide computed by a Gaussian urban air quality model (ADMS Urban). A key step of the assimilation procedure is to model the error variance of the NO_2 concentration

field. It means specifying the variance of the error at all computed locations and specifying the correlation between errors at different locations. The urban air quality model is static so that it is not possible to apply a filter, like a Kalman filter, that would propagate the error variance. In this paper, the error variance for the concentration fields is therefore prescribed, through a specific parameterization that takes into account the road network. The so-called best linear unbiased estimator (BLUE) is then computed for every available date of the model simulation.

The concentration fields for nitrogen dioxide are computed by ADMS across the city of Clermont-Ferrand, France, every three hours for the whole year 2008. The air quality monitoring network is composed of nine fixed stations — two traffic stations, four urban stations and three peri-urban stations. Details about the model, its computations and the case study may be found in Section 2. The assimilation method is described in Section 3, and the Section 3.3 details the parameterization of the error variance for the concentration fields. The choice of the assimilation parameters is discussed in Section 4.1. The results are analyzed in Sections 4.2 and 4.3.

2. Urban Air Quality Modeling over Clermont-Ferrand

2.1. ADMS Urban

ADMS Urban [*D.J. Carruthers and Singles*, 1998] is an air quality model for the dispersion in the atmosphere of continuous releases from the full range of emission sources including road traffic, industrial, commercial and domestic emissions. This static model estimates the stationary solution of the dispersion equation, using a three dimensional quasi-Gaussian formulation. It requires input meteorological data, background concen-

trations and detailed emission inventories. The output simulation mesh is subdivided in a coarse regular grid and a high-resolution mesh in the vicinity of main emission sources.

A meteorological pre-processor calculates the required boundary layer parameters from a variety of input data. The wind speed and the cloud cover enable to determine the surface heat flux through a surface radiation budget [*Holtslag and Ulden*, 1983]. A two-equation system, composed of a surface layer wind profile equation and a Monin Obukhov length equation, enables to estimate the friction velocity and the Monin Obukhov length. These two parameters are used to compute the boundary layer height in stable conditions as described by *Nieuwstadt* [1981]. In convective atmosphere, the boundary layer height evolves according to an unstationary integral model [*Tennekes*, 1973; *Tennekes and Driedonks*, 1981; *Driedonks*, 1982]. Different profiles of the boundary layer (mean wind, temperature, standard deviation of wind components, etc.) are then determined from surface similarity theory. A topography module manages the dispersion over hills and over regions with surface roughness changes. In neutral or convective conditions, the wind and turbulence fields are calculated using linearized analytical solutions of the momentum and continuity equations. In very stable conditions, the atmosphere is divided into two layers: in the layer just above the surface, the air flows around the relief; in the other layer, the air flows over the relief. For intermediate conditions, ADMS Urban relies on a weighted average of these two behaviors based on Froude number.

From the boundary layer profiles and the mean plume height, ADMS Urban determines the horizontal and vertical concentration distributions, which are always Gaussian except in convective conditions, where the non-Gaussian vertical concentration distributions depend on the skewed vertical velocity distributions. A street canyon model enables to

determine the concentration field in the streets whose buildings are higher than 0.5 m. This model is based on the Danish model Operational Street Pollution Model [*Hertel and Berkowicz*, 1989]. For this work, the chemistry is quite simple: the NO_2 concentration is determined from the NOx concentration as described in *Derwent and Middleton* [1996].

2.2. Meteorological Data, Topography and Land Use

The meteorological input data is measured at the Météo-France station Aulnat located in the Clermont-Ferrand airport. Wind speed, wind direction and temperature are required along with the cloud cover.

The Shuttle Radar Topography Mission (SRTM) data sets provide the topography data in case of activation of the topography module. The $3''$-resolution data base results from the collaboration between the NASA and the National Imagery and Mapping Agency, among others.

We consider homogeneous land cover with constant roughness length of 0.4 m but we use specific value (0.2 m) for the site of the meteorological station : the model adjusts wind speed measurements to take into account this difference.

2.3. Emissions

Emissions include main industrial sources, road sources and a grid source for poorly-defined sources like heating sources and minor roads. Location and width of roads and buildings heights are estimated from "Clermont Communauté" database.

For road sources, the emissions in g are computed as $E = AF$, where A is the vehicle activity in vehicles km^{-1} and F a unitary emission factor in g km vehicules^{-1}. The emissions are computed using COPERT IV, the COmputer Program to calculate Emission

from Road Transport (`http://www.emisia.com/copert/General.html`), which relies on a database of unitary emission factors. A unitary emission factor is attributed for each pollutant to each vehicle category. It depends on the carburetor mode, the engine size and the vehicle registration date. The emission factor also depends on the vehicle speed, imposed by road signs, and on the traffic conditions, which depend on the month and on the day type (weekday, Saturday and Sunday). Traffic conditions are determined from past observations of traffic counters over the city. Note that the real time traffic is not considered. The model COPERT IV takes into account the warm emissions, the cold emissions and the slope-induced emissions for the heavy transport. A few corrections are applied for old vehicles and for fuel improvements.

2.4. Case Study

A simulation at urban scale has been carried out over the city of Clermont-Ferrand for the whole year 2008. The output concentrations are computed at $30,971$ ADMS Urban receptors, all located at 1.5 m from the ground. The concentrations of nitrogen dioxide have been computed at these receptors every three hours. As depicted in Figure 1, the air quality monitoring network is composed of nine fixed stations, with two traffic stations (Gare and Roussillon), four urban stations (Lecoq, Delille, Jaude and Montferrand) and three peri-urban stations (Gerzat, Gravanches and Royat). The stations at Roussillon, Gerzat, Gravanches and Royat are rather far from the group of stations located in the center of the city.

The altitude of the stations varies while the computed concentrations are all located at 1.5 m height, so as to avoid modeling the error correlations along the vertical between simulated concentrations (see Section 3.3). However, in order to better evaluate the model

performance without assimilation, we add one ADMS receptor per station, located at the real stations altitudes. Note that these nine additional receptors are not used in the assimilation procedure.

The performance evaluation relies on the scores shown in Table 1 and on criteria introduced by *Chang and Hanna* [2004]: a normalized bias between -0.3 and 0.3 is recommanded and a normalized mean square error (NMSE) should be lower than 1.5. We prefer to define the limit NMSE as 0.5 and we target a correlation higher than 0.6. The actual values for our full-year simulation are given in Table 2. For all the stations, the normalized bias is between -0.3 and 0.25. The correlation and the NMSE are out of these criteria only for the station Royat, with a correlation of 0.59 and a NMSE of 1.03. At this station, the dispersion model overestimates the concentrations. Royat is located on the Clermont-Ferrand heights and the relief is rugged around this station, so the wind field is hard to simulate and ADMS Urban does not succeed in it. The scores at the other stations are significantly better, except for the station Jaude whose NMSE is almost equal to the limit value.

3. Assimilation Method

3.1. Problem Statement

The model produces the state vector c^b (b stands for background). The concentration field is observed at given locations, which gives an observation vector o. A data assimilation algorithm will produce a new state vector c^a (a stands for analysis) based on the model state c^b and the observation o.

Each observation location matches on the horizontal with an ADMS Urban receptor. We consider that the concentration simulated by ADMS at these receptors is our estimation

of the true concentration at the station location, even though there may be a difference of altitude between the station and the ADMS receptor. We introduce the so-called observation operator H which maps from the state space to the observation space, so that Hc^b is the simulated counterpart of o. The operator H is therefore a matrix in which each row i is full of zeros except at the the column j that corresponds to the receptor located at observation station i. The elements H_{ij} are equal to one if and only if the j-th receptor corresponds to the i-th observation station. The discrepancy between the observations and the simulated concentrations, $o - Hc^b$, is called the innovation.

Let c^t be the real atmospheric concentrations that at the ADMS receptors. We assume that the computed concentrations c^b have an unbiased error $c^b - c^t$ with variance B. We assume that the observation vector o has an unbiased error $o - Hc^t$ with variance R. Note that the observational error depends on H. If the true concentrations at the observed locations are o^t, the observational error $o - Hc^t$ can be decomposed in an instrumental error $o - o^t$ and a representativeness error $o^t - Hc^t$. In our case, the latter is due to altitude difference between the observation station and the ADMS receptor.

3.2. Best Linear Unbiased Estimator (BLUE)

Based on c^b, B, o and R, the analysis state vector is computed as the so-called "Best Linear Unbiased Estimator" which is linearly dependent on c^b and o, has unbiased error $c^a - c^t$ and has a variance with minimum trace [see, e.g., *Bouttier and Courtier*, 1999]. This estimator is uniquely defined as

$$c^a = c^b + K(o - Hc^b) \ ,$$

where

$$K = BH^T(HBH^T + R)^{-1} .$$

For data assimilation at larger scale, the state error covariances can be reasonably parameterized as a function of the geographical distance, e.g., with a decreasing exponential. At urban scale, our state error variances do not only depend on the distance, but also on the road network.

3.3. Modeling of the Covariance Matrices

The observational error covariance matrix is taken diagonal, hence assuming no correlation between the observational errors at two different stations. The observational errors covariance matrix is therefore

$$R = v_o I ,$$

where v_o is the observational error variance.

For nitrogen dioxide, we assume that an important part of the state errors originates from the traffic emissions. As a consequence, we assume high error correlations between receptors on the same road or on connected roads. Also, a receptor on a road should show a lower error correlation with a receptor in the background than with another (equally distant) receptor on the road.

We introduce the distance d_{ij} along the road between two receptors indexed by i and j. The distance along the road is defined as the smallest distance it takes to travel on the road network between the two receptors. If the two receptors i and j are not located on a road, they are first orthogonally projected on the road network, and d_{ij} is taken as the distance along the road between the projections. We also introduce the distance P_i of the receptor i to the road network, that is the geographic distance to the closest road.

We define B_{ij}, the covariance between the state errors at receptors i and j, as

$$B_{ij} = v_c \exp\left(-\frac{d_{ij}}{L_d}\right) \exp\left(-\frac{|P_i - P_j|}{L_p(i,j)}\right) ,$$

with

$$L_p(i,j) = L_p + \alpha \min(P_i, P_j) ,$$

where L_d and L_p are characteristic distances, strictly positive, respectively along the road network and transverse to the road network, α a scaling coefficient without dimension and v_c a variance. The covariance is assumed to decrease exponentially against the distance along the road and in the direction transverse to the road. The correction $\alpha \min(P_i, P_j)$ is added so that the decorrelation length is increased with the distance to the network: while the error correlation with a road receptor is assumed to decrease fast in the vicinity of the road, the errors correlation between two background receptors should remain significant across a larger distance. The Figure 2 illustrates the state error covariances modeling: the first figure shows the error correlations (B_{ij}/v_c) with a receptor located on the road network, whereas the second figure shows the error correlations with a receptor located out of the road network.

The error covariances are constant in time. In particular, they do not depend on traffic conditions. This is surely an approximation which should be addressed by uncertainty quantification studies on urban models. Such study would propagate in the model the uncertainties originating from traffic emissions. It would require prior uncertainty quantification on traffic assignment (and corresponding emissions), which would in turn require the availability of traffic observations for the evaluation of the traffic model. In this paper, the proposed covariance model is parameterized so that it can be applied in the absence of a reliable uncertainty quantification study.

3.3.1. Specific examples

Between two receptors on the road network ($P_i = P_j = 0$), the state error covariance is equal to $\frac{1}{2}v_c$ when the distance between the receptors is $d_{ij} = 0.7L_d$. Between two receptors on the same normal to the road network ($d_{ij} = 0$) and on the same side, the state error covariance is equal to $\frac{1}{2}v_c$ when the distance between the receptors is $0.7(L_p + \alpha \min(P_i, P_j))$. By definition, this distance increases for background receptors. Between two receptors so that $P_i = P_j$, not necessarily on the road network, the covariance highly depends on the distance along the road network. Their errors correlation is equal to 1 if $d_{ij} = 0$: we assume that these two receptors are subject to the same errors.

Note that state error covariance matrix B is a covariance matrix, hence symmetric and positive semi-definite. The matrix is not positive definite because we can found two distinct receptors with the same distance to the road network and the same projection on the road network; hence several columns (or rows) of B are identical.

3.3.2. Implementation

Computing B requires the evaluation of the distance along the road between all receptors projections on the road network. In order to carry out these computations, we represent the road network as a non-oriented graph: each road portion without any crossroad is an edge and each crossroad is a node. In the graph, we also add as new nodes the projections of the receptors on the road network. We then add the corresponding edges, which represent the road portions between all nodes (i.e., the projections and the crossroads). The weight of an edge is the length of the road portion.

The celebrated Dijkstra's algorithm may be applied to find the shortest path between two nodes in the graph. If V is the number of vertices and E is the number of edges, the

complexity of an efficient implementation of the algorithm is $\mathcal{O}(E+V\log V)$. This should be applied to each pair of nodes, hence resulting in a complexity of $\mathcal{O}(EV^2+V^3\log V)$. This is intractable in our case where $E=44,242$ and $V=35,413$.

We thus apply Johnson's algorithm which is designed to efficiently compute the shortest paths between all pairs of nodes. This algorithm uses Dijkstra's algorithm, but its overall time complexity is $\mathcal{O}(VE\log(V))$ in the Boost implementation. The shortest path algorithm is fully described on the page `http://www.boost.org/doc/libs/1_40_0/libs/graph/doc/johnson_all_pairs_shortest.html`.

4. Application

4.1. Determination of Assimilation Parameters

4.1.1. Observations and their Error Variances

We do not have access to detailed information on observation errors over Clermont-Ferrand, but we have access to the mean observation variance over the monitoring network of Paris metropolitan area. Based on *Airparif* [2007], the air quality association for Paris area, Airparif, evaluates the uncertainty of the observations of its monitoring network. The uncertainty is computed as a sum of variances which correspond to different error sources (instrument calibration, temperature and pressure conditions, data processing, etc.). We analyzed the uncertainties evaluated by Airparif for the full year 2009. On average, the uncertainty decreases with the concentration. For nitrogen dioxide, the mean concentration measured over Paris is $40.7\,\mu g\,m^{-3}$ whereas it is only $25.4\,\mu g\,m^{-3}$ over Clermont-Ferrand. Consequently, the mean uncertainty value obtained over Paris cannot be directly applied to Clermont-Ferrand. A way around the problem is to remove

the highest concentrations from the database in order to reduce the mean concentration down to $25.4\,\mu g\,m^{-3}$. The corresponding error variance is then $5.96\,\mu g^2\,m^{-6}$.

The concentrations are simulated at 1.5 m from the ground, but they are measured at higher altitude. This difference is taken into account in the representativeness error, which is part of the observational error. We approximate the representativeness error based on model simulations which are available both at the station height and at 1.5 m. The mean empirical variance of the differences between the simulated concentrations at the two altitudes is $1.75\,\mu g^2\,m^{-6}$. The observational error variance is roughly estimated by the sum of the measure error variance and the representativeness error variance; we finally set it to $8\,\mu g^2\,m^{-6}$.

4.1.2. State Error Variance: χ^2 Diagnosis

The state error variance is determined using a χ^2 diagnosis. The diagnosis enables to check the consistency between the available innovations

$$o_n - H_n c_n^b$$

and their variances

$$S_n = R_n + H_n B_n H_n^T \ ,$$

where n represents the time step. The scalar

$$\chi_n^2 = (o_n - H_n c_n^b)^T S_n^{-1} (o_n - H_n c_n^b)$$

is expected to be equal to the number F_n of observations. And therefore, we should have

$$\sum_{n=1}^{T} \frac{\chi_n^2}{F_n} \simeq T \ .$$

Hereafter, we consider the value

$$A = \frac{1}{T} \sum_{n=1}^{T} \frac{\chi_n^2}{F_n} \, ,$$

where T is the total number of steps. This value of A should be 1.

The χ^2 diagnosis is carried out for several values of (v_c, L_d, L_p, α). The Table 3 reports a few tests, and supports the choice $(v_c, L_d, L_p, \alpha) = (220\,\mu\mathrm{g}^2\,\mathrm{m}^{-6}, 3000\,\mathrm{m}, 200\,, 1\,)$, which we define as the reference configuration. The impact on the value of A of the decorrelation length transverse to the road network and of α is lower than the impact of the state error variance and of the decorrelation length along the road network.

4.2. Results

The assimilation is carried out every three hours, when new simulated concentrations are available.

The analyzed concentration at a station location is almost equal to the observation (see Figure 3), which is partly expected because the ratio between the state error variance and the observation error variance is very low.

Before assimilation, the model often computes too low concentrations at urban stations. The assimilation of the observations efficiently corrects this problem, as depicted in Figure 3. After assimilation, the road network remains clearly visible and the concentrations are higher in the immediate vicinity of the road. At peri-urban stations, the model may simulate too high concentrations, which is also corrected by data assimilation. The analyzed values lead to a reduced background pollution in a large perimeter around the peri-urban stations while the pollution over the roads in this area is almost not impacted.

As the data assimilation strongly corrects the concentrations in the vicinity of the stations and may not correct the concentrations further, the concentration maps can

show some spatial inconsistencies, even if every point concentration of these maps is likely to be closer to its true value. The main scenarios, when inconsistencies can occur, are of two kinds. In the first scenario, the model overestimates the concentrations in urban area. The assimilation of the observations at traffic stations decreases the concentrations on the road network, while the background concentrations may remain essentially unchanged, and possibly with higher values. In this case, the concentrations in one road may be lower than the concentrations in the background. In the second scenario, the observations at a peri-urban stations are strongly higher than the simulated concentrations. Then again, the corrected concentrations in the background can become higher than the concentrations along the roads. However these scenarios seldom occur.

Note that the reference values $(v_c, L_d, L_p, \alpha) = (220\,\mu g^2\,m^{-6}, 3000\,m, 200\,m, 1)$ were selected not only on the basis of the χ^2 diagnosis (which can be satisfied with other values), but also on the basis of the output maps. The physical inconsistencies previously mentioned especially occur when the value chosen for L_p is too low compared to L_d and when α is lower than 1.

4.3. Performance Evaluation with Leave-One-Out Cross-Validation

The leave-one-out cross-validation consists in removing the observations of a given station from the data assimilation process. Only the observations from the other stations are used to correct the concentrations. This procedure is carried out for all stations, one by one: only one station is removed at a time. At the removed station, the model performance at 1.5 m height is compared to the performance after assimilation of the observations of the other stations. This enables to check whether the assimilation properly distributes

in space the corrections that originate from the observed locations. The cross-validation evaluates the effects of the data assimilation method at locations without any observation.

4.3.1. Scores

The cross-validation was carried out for the reference values $(v_c, L_d, L_p, \alpha) = (220\,\mu g^2\,m^{-6}, 3000\,m, 200\,m, 1)$ from Section 4.1.2. The performance before assimilation is given in Table 4. The results after assimilation are given in Table 5. The largest improvements occur in urban area (at the station Jaude, the improvement is of 46 %), compared to peri-urban area (at the stations Gravanches and Royat, the improvements are respectively of 17 % and 5 %). It is likely that the distance between the peri-urban stations makes it difficult to obtain enough information to compute strong and reliable corrections from one station to the other. Another possible explanation may be an unsatisfactory modeling of the error covariances between peri-urban receptors or between urban and peri-urban receptors. In some cases, the absolute bias increases but remains inside the interval recommanded in *Chang and Hanna* [2004] (see the first two columns of Table 5).

Figure 4 shows the RMSE for the months of the year, at all stations and at Jaude. Note that the largest improvements are found at Jaude (see Figure 5), which is close to the road network and in the vicinity of three other stations. The distance to the other stations plays an important role, as shown in Figure 5. The largest improvements are found at stations close to the rest of the network.

We finally consider all discrepancies between observations and simulated concentrations. Figure 6 shows the relative frequency distribution of the discrepancies, before and after

assimilation. After assimilation, the discrepancy distribution is significantly narrowed around 0. The largest discrepancies have a much lower frequency after assimilation.

4.3.2. Sensibility to the Parameters of the State Error Covariance Matrix

First, several values of the scalar α are tested, whereas the characteristic decorrelation lengths L_d and L_p remain constant and equal respectively to 3000 m and 200 m. The Table 6 shows that the global RMSE decreases when α increases, but the sensitivity is very low. This parameter essentially plays a role in the vicinity of peri-urban stations, but there is no pair of close peri-urban stations that could help to evaluate the real impact of α. It is set to 1 in the rest of the study.

The assimilation performance significantly increases with the characteristic decorrelation length along the road network, L_d. Table 7 reports the performance for several values of L_d, with L_p set to 200 m. The best performance is achieved for $L_d = 5000$ m and slight performance variations occur for lengths greater than 4000 m. As the values v_c that satisfy the χ^2 diagnosis increase with L_d, the value of the characteristic length is limited by the range of variances v_c which are consistent with the model performance. Finally, we selected the intermediate value $L_d = 3000$ m, for which the correlation between errors drops down to 0.5 at a distance of 2100 m along the road network. It gives good results for a moderate decorrelation length and variance. There is a clear need for research on uncertainty estimation at urban scale in order to decide which values may be more adapted.

The impact of the decorrelation length transverse to the road network, L_p, is much more limited. The optimal value of L_p is not clearly determined by the Figure 7. With $L_d = 3000$ m, the RMSE is almost identical for L_p equal to 200 m or 300 m. The RMSE at peri-

urban stations is better with $L_p = 300\,\mathrm{m}$ than with $L_p = 200\,\mathrm{m}$. On contrary, at traffic and urban stations, the RMSE is lower with $L_p = 200\,\mathrm{m}$ than with $L_p = 300\,\mathrm{m}$. Finally, we recommend a moderate length $L_p = 200\,\mathrm{m}$ which leads to same global performance.

5. Conclusions

The paper demonstrates the efficiency of data assimilation at urban scale for the improvement of NO_2 concentration fields using fixed monitoring stations. Computing the best linear unbiased estimator (BLUE) has proved to be efficient for the correction of prior concentrations computed by the urban Gaussian model ADMS. Despite the low number of stations available in the simulation domain, strong improvements (30% to 50%) were found at urban monitoring stations excluded from the assimilation procedure, in a leave-one-out cross-validation. This shows that, in the part of the domain where the station density is high, large improvements are likely to occur at non-observed locations.

However, in the background, far from the monitoring network, the improvements are low. It is not clear whether these low improvements at rural locations is due to lack of information from the observation network or to shortcomings in the error covariance modeling. In the algorithm, a key variable is indeed the error covariance matrix B that determines the spatial distribution of the corrections. The proposed covariance matrix is motivated by the prominent role of traffic emissions in urban NO_2 concentrations, but it surely misses significant error correlations.

The parameters of the error covariance matrix are constant in time and in space, whereas the characteristic lengths can depend on traffic and the variance surely depends on the concentration levels. A future work on traffic model evaluation, using observations from traffic counters, is essential to improve the parameterization. Involving the concentration

field in the variance v_c or more generaly in the covariance formula is also the next step. One option is to follow *Riishøjgaard* [1998] to model the term transverse to the road network.

Future work on uncertainty estimation at urban scale should be a key step for better uncertainty estimation, and therefore a better modeling of the error covariance matrix B. There is a need for the generation of ensembles of urban simulations that would properly sample the concentrations uncertainties. Classical approaches based on Monte Carlo simulations or multimodel ensembles should be investigated at urban scale, although they require so tremendous computational resources that model reduction may be needed.

Uncertainty estimation for the concentrations after assimilation should also be investigated. The error covariance matrix for the analysis, i.e., $(I - KH)B$ for BLUE, should show much lower eigenvalues than B. For instance, one objective would be to provide some confidence interval on the population exposure.

Another direction is inverse modeling. One may want to correct the input emissions which are known to be an important source of uncertainty. Such approach often has high computational costs. It is however difficult to anticipate whether the resulting air concentrations would be closer to the real concentrations than those of our current approach.

At the time this paper is written, the assimilation as previously detailed has been applied operationally for a year on the prototype "Votre Air" (operated by Airparif; see `http://votreair.airparif.fr/`). The prototype computes in near real-time the air quality over a part of Paris, and it assimilates the observations from eight fixed stations [*Pradelle et al.*, 2011]. This justifies that the approach, proved to be computationally tractable even for real-time computations, is currently integrated in the platform Urban

Air System [*Pradelle et al.*, 2010]. With the deployment of such systems, new questions will arise, such as the assimilation of observations from mobile sensors (e.g., embedded in public buses).

Acknowledgments. We would like to thank the air quality association Atmo Auvergne (`http://www.atmoauvergne.asso.fr/`) that provided us with the observations and the ground data (especially, the emissions) for the simulation over Clermont-Ferrand. We are also grateful to Cap Digital (`http://www.capdigital.com/`), the French business cluster for digital content and services of Île-de-France, for its financial support on the project "Votre Air" during which part of the assimilation approach was developed.

References

Airparif, Guide pratique d'utilisation pour l'estimation de l'incertitude de mesure des concentrations en polluants dans l'air ambiant, *Tech. Rep. Version 9*, AIRPARIF, 2007.

Bouttier, F., and P. Courtier, Data assimilation concepts and methods, *Meteorological training course lecture series*, ECMWF, 1999.

Chai, T., et al., Four-dimensional data assimilation experiments with International Consortium for Atmospheric Research on Transport and Transformation ozone measurements, *Journal of Geophysical Research*, *112*(D12S15), doi:10.1029/2006JD007763, 2007.

Chang, J. C., and S. R. Hanna, Air quality model performance evaluation, *Meteorology and Atmospheric Physics*, *87*, 167–196, 2004.

Derwent, R. G., and D. R. Middleton, An empirical function for the ratio NO_2:NO_x, *Clean Air*, *26*, 57–60, 1996.

411 D.J. Carruthers, H. E. C. M., and R. Singles, Development of ADMS-urban and com-
412 parison with data for urban areas in the UK. Proc. of Air Pollution Modelling and its
413 Application XII, *Tech. rep.*, CERC, 1998.

414 Driedonks, A., Models and observations of the growth of the atmospheric boundary layer,
415 *Boundary-Layer Meteorology*, *23*, 283–306, 1982.

416 Elbern, H., and H. Schmidt, Ozone episode analysis by four-dimensional variational chem-
417 istry data assimilation, *Journal of Geophysical Research*, *106*(D4), 3,569–3,590, 2001.

418 Hertel, O., and R. Berkowicz, Operational street pollution model (OSPM). Évaluation of
419 the model on data from st olavs street in oslo, *Tech. rep.*, DMU Luft, 1989.

420 Holtslag, A., and A. V. Ulden, A simple scheme for daytime estimates of the surface
421 fluxes from routine weather data, *Journal of Applied Meteorology and Climatology*, *22*,
422 517–529, 1983.

423 Nieuwstadt, F., The steady-state height and resistance laws of the nocturnal boundary
424 layer : Theory compared with cabauw observations, *Boundary-Layer Meteorology*, *20*,
425 3–17, 1981.

426 Pradelle, F., A. Armengaud, C. Pesin, M. N. Rolland, J. Virga, G. Luneau, C. Schillinger,
427 and D. Poulet, Urban air system: an operational modelling system for survey and
428 forecasting air quality at urban scale, 13th international conference on harmonisation
429 within atmospheric dispersion modelling for regulatory purposes, Paris, France, 2010.

430 Pradelle, F., F. Brocheton, B. Chabanon, C. Honoré, F. Dugay, K. Léger, F. Dambre,
431 V. Mallet, and A. Tilloy, The "Votre Air" project: development of a modelling tool to
432 assess the real atmospheric exposure in Paris, 14th international conference on harmon-
433 isation within atmospheric dispersion modelling for regulatory purposes, Kos Island,

Greece, 2011.

Riishojgaard, L. P., A direct way of specifying flow-dependent background error correlations for meteorological analysis systems, *Tellus*, *50A*, 42–57, 1998.

Segers, A., Data assimilation in atmospheric chemistry models using Kalman filtering, Ph.D. thesis, Delft University, 2002.

Tennekes, H., A model for the dynamics of the inversion above a convective boundary layer, *Journal of the Atmospheric Sciences*, *30*, 558–567, 1973.

Tennekes, H., and A. Driedonks, Basic entrainment equations for the atmospheric boundary layer, *Boundary-Layer Meteorology*, *20*, 515–229, 1981.

Wu, L., V. Mallet, M. Bocquet, and B. Sportisse, A comparison study of data assimilation algorithms for ozone forecasts, *Journal of Geophysical Research*, *113*(D20310), 2008.

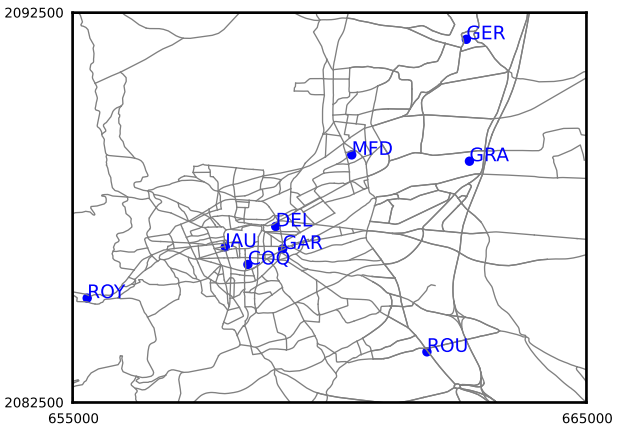
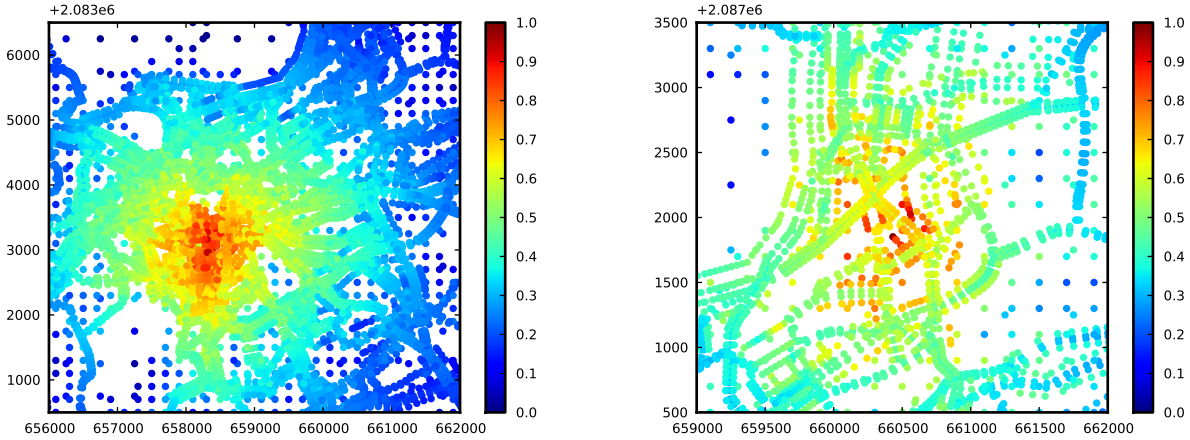
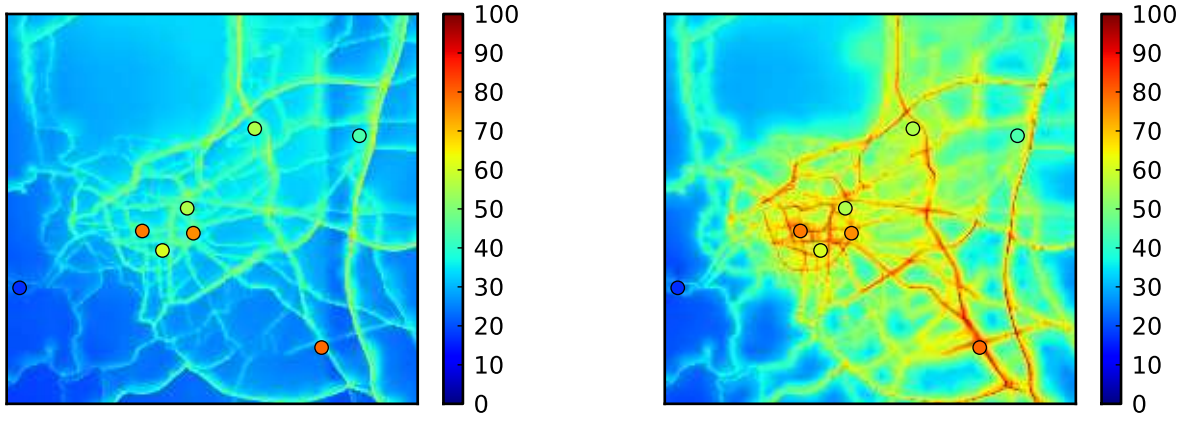


Figure 1. The modeled road network of the city of Clermont-Ferrand and the location of the observation stations: GAR stands for Gare, ROU for Roussillon, COQ for Lecoq, DEL for Delille, JAU for Jaude, MFD for Montferrand, GER for Gerzat, GAR for Gravanches and ROY for Royat. The coordinate projection system is the Lambert II extended.



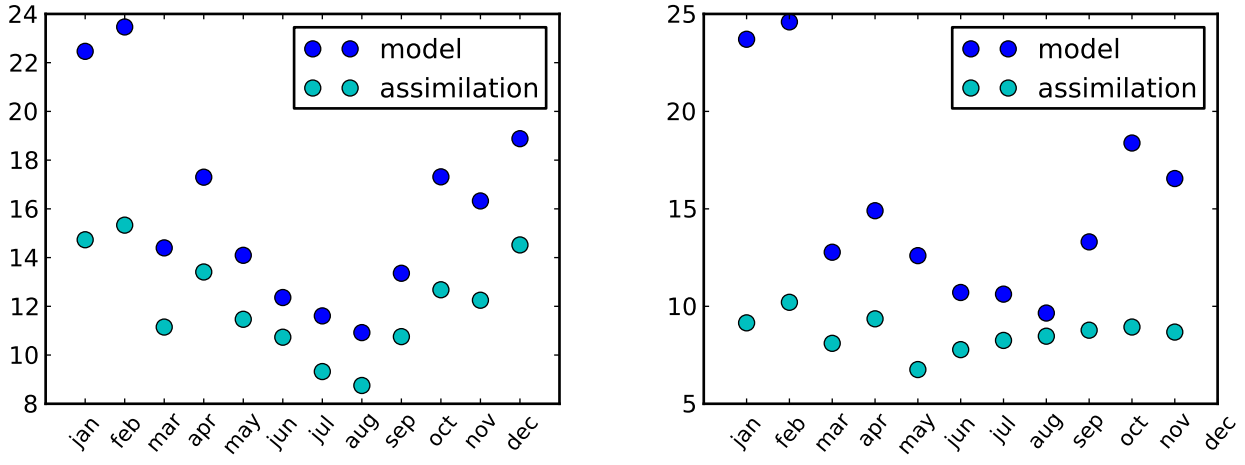
(a) (b)

Figure 2. The state error correlations (B_{ij}/v_c) between the receptor located at (a) the station Gare or at (b) the station Montferrand and the other receptors. This corresponds to one row of the state error covariance matrix divided by v_c. Notice that the figures do not correspond to the same domain areas.



(a) modeled (b) after data assimilation

Figure 3. Maps of nitrogen dioxide concentrations over the city of Clermont-Ferrand on 10 July 2008 at 6 UTC, in $\mu g\,m^{-3}$. The data assimilation parameters are $L_d = 3000\,m$, $L_p = 200\,m$ and $\alpha = 1$. The disks represent the concentrations measured at stations.



(a) all stations (b) the station Jaude

Figure 4. Monthly RMSE in $\mu g\,m^{-3}$ of the model in blue and after data assimilation in cyan, for (a) all stations and at (b) Jaude.

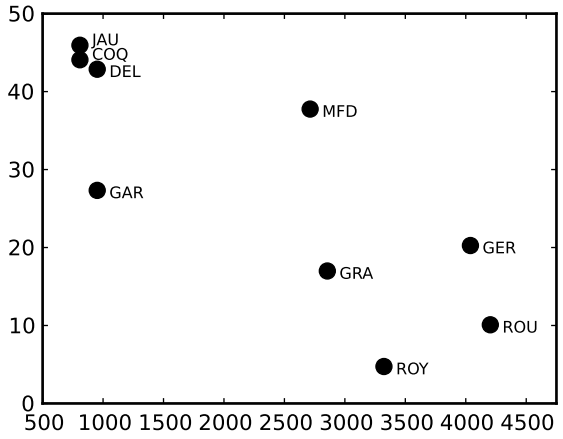


Figure 5. Improvement of stations RMSE in % (see Table 5), against the distance (m) to the rest of the network. See Figure 1 for the position of the stations.

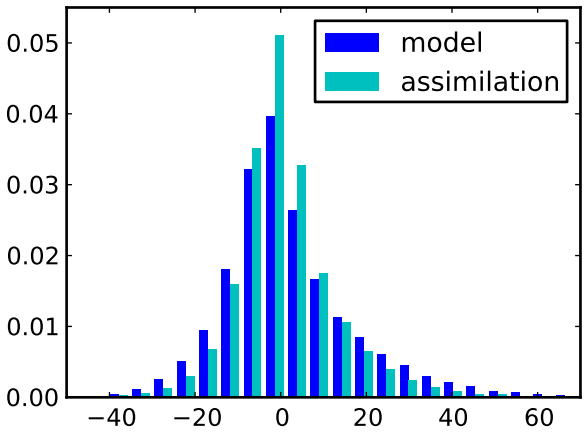


Figure 6. In blue the dispersion of the innovations, in cyan the dispersion of the discrepancy to observations after data assimilation (in leave-one-out cross-validation). The abscissa is a concentration discrepancy in $\mu g\,m^{-3}$ and the ordinate is the relative occurrence frequency.

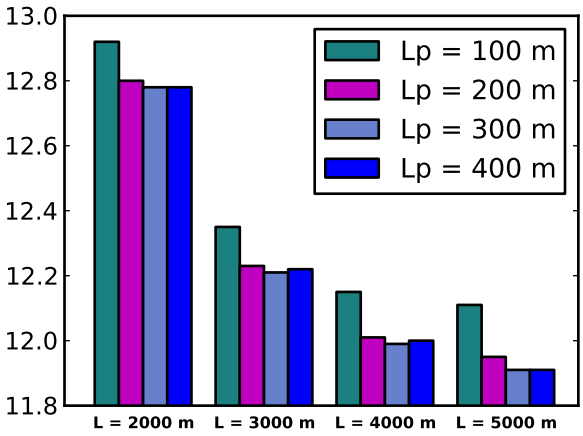


Figure 7. The RMSE in $\mu g\,m^{-3}$ over all stations for several pairs of parameters L_d and L_p.

Table 1. Scores for the performance evaluation of a model. $(c_i)_i$ is the simulated temporal sequence. $(o_i)_i$ is the corresponding observed sequence. n is the total number of elements in the sequence. $\bar{c}$ and $\bar{o}$ are respectively the mean of $(c_i)_i$ and $(o_i)_i$.

Score	Formula
Bias	$\frac{1}{n}\sum_{i=1}^{n}(c_i - o_i)$
Normalized bias	$\frac{1}{n}\sum_{i=1}^{n}\frac{(c_i - o_i)}{\bar{o}}$
Correlation	$\frac{\sum_{i=1}^{n}(c_i - \bar{c})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n}(c_i - \bar{c})^2}\sqrt{\sum_{i=1}^{n}(o_i - \bar{o})^2}}$
Mean absolute error	$\frac{1}{n}\sum_{i=1}^{n}\lvert c_i - o_i\rvert$
Normalized mean absolute error	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert c_i - o_i\rvert}{\bar{o}}$
Normalized mean square error	$\frac{1}{n}\sum_{i=1}^{n}\frac{(c_i - o_i)^2}{\bar{c}\bar{o}}$
Root mean square error	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(c_i - o_i)^2}$

Table 2. Model performance. The simulation values are computed at measurement height. The concentration, the bias and the MAE are in $\mu g\,m^{-3}$, the other indicators are without units. All the indicators formulae are defined in Table 1.

	Observation mean	Simulation mean	Normalized bias	Correlation	MAE	Normalized MAE	RMSE	NMSE
Traffic stations								
Gare	49.5	37.0	-0.3	0.69	18.0	0.42	25.5	0.36
Roussillon	37.6	29.1	-0.26	0.69	14.6	0.44	19.8	0.36
Urban stations								
Lecoq	25.7	25.9	0.01	0.74	10.6	0.41	15.1	0.34
Delille	27.7	28.0	0.01	0.73	11.1	0.40	15.0	0.29
Jaude	27.2	21.0	-0.25	0.73	11.0	0.46	16.7	0.49
Montferrand	25.9	25.1	-0.03	0.73	10.3	0.41	14.5	0.32
Peri-urban stations								
Gerzat	23.1	19.5	-0.17	0.75	8.8	0.41	12.5	0.35
Gravanches	23.6	22.3	-0.06	0.73	9.5	0.41	13.6	0.35
Royat	12.5	16.1	0.25	0.59	10.0	0.70	14.4	1.03

Table 3. The value of A for several choices of parameters (v_c, L_d, L_p, α). The state error variance is in $\mu g^2\,m^{-6}$, the characteristic lengths in m and α without units.

State error variance	L_d	L_p	α	A
200	3000	200	1	1.10
215	3000	200	1	1.03
220	3000	200	1	1.01
230	3000	200	1	0.97
220	**4000**	200	1	1.06
220	**3000**	200	1	1.01
220	**5000**	200	1	1.13
220	**6000**	200	1	1.20
220	3000	**100**	1	0.98
220	3000	**200**	1	1.01
220	3000	**300**	1	1.03
220	3000	200	**0**	1.01
220	3000	200	**1**	1.01
220	3000	200	**2**	1.01
220	3000	200	**3**	1.02

Table 4. Model performance at 1.5 m. Contrary to Table 2, the simulation values are computed at 1.5 m whereas the stations can be at higher altitude. The bias and the RMSE are in $\mu g\,m^{-3}$, the correlation and the normalized mean square error are indicators without units. All the indicators formulae are defined in Table 1.

	Observation mean	Bias	Correlation	RMSE	NMSE
Traffic stations					
Gare	49.5	-10.7	0.68	24.7	0.32
Roussillon	37.6	-7.5	0.69	19.5	0.34
Urban stations					
Lecoq	25.7	0.8	0.74	15.1	0.33
Delille	27.7	0.5	0.72	15.1	0.29
Jaude	27.2	-3.0	0.72	16.0	0.39
Montferrand	25.9	-0.5	0.73	14.5	0.32
Peri-urban stations					
Gerzat	23.1	-3.4	0.75	12.4	0.34
Gravanches	23.6	-1.2	0.74	13.6	0.35
Royat	12.5	3.9	0.59	14.5	1.02

Table 5. Scores of the cross validation for the configuration ($L_d = 3000\,\mathrm{m}, L_p = 200\,\mathrm{m}, \alpha = 1$). The simulation values are computed at 1.5 m height whereas the stations can be at higher altitude. The "improvement" is the relative change in % of the RMSE before and after assimilation of observations at the other stations.

	Bias	Correlation	RMSE	NMSE	Improvement
Traffic stations					
Gare	-9.4	0.87	17.9	0.36	28%
Roussillon	-5.5	0.74	17.6	0.47	10%
Urban stations					
Lecoq	4.3	0.95	8.4	0.33	44%
Delille	3.3	0.93	8.6	0.31	43%
Jaude	-0.2	0.92	8.6	0.32	46%
Montferrand	0.6	0.91	9.0	0.35	38%
Peri-urban stations					
Gerzat	-3.1	0.86	9.9	0.43	33%
Gravanches	-0.9	0.83	11.3	0.48	17%
Royat	3.6	0.65	13.8	1.10	5%

Table 6. The RMSE in $\mu\mathrm{g\,m^{-3}}$ over all stations against the scalar α.

α	0	0.5	1	2	3	4
RMSE	12.28	12.25	12.23	12.20	12.19	12.18

Table 7. The RMSE in $\mu\mathrm{g\,m^{-3}}$ over all stations against the decorrelation length L_d, in m. L_p and α are set respectively to 200 m and to 1. The variance v_c is determined by the χ^2 diagnosis.

L_d	2000	3000	4000	5000	6000	7000
v_c	218	220	235	250	265	280
RMSE	12.80	12.23	12.01	11.95	11.96	12.01