



BLUE-based NO₂ data assimilation at urban scale

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

► **To cite this version:**

Anne Tilloy, Vivien Mallet, David Poulet, Céline Pesin, Fabien Brocheton. BLUE-based NO₂ data assimilation at urban scale. *Journal of Geophysical Research, American Geophysical Union*, 2013, 118 (4), pp.2031-2040. <10.1002/jgrd.50233>. <hal-00826581>

HAL Id: hal-00826581

<https://hal.inria.fr/hal-00826581>

Submitted on 2 Sep 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

BLUE-based NO₂ data assimilation at urban scale

Anne Tilloy^{1,2}, Vivien Mallet^{1,2}, David Poulet³, Céline Pesin³, Fabien

Brocheton³

©Copyright AGU

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)

¹INRIA, Paris-Rocquencourt research center, France

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

³Numtech, Aubière, France

3 **Abstract.** We aim at optimally combining air quality computations, from
4 the Gaussian model ADMS Urban, and ground observations at urban scale.
5 An ADMS simulation generated NO₂ concentration fields across Clermont-
6 Ferrand (France) down to street level, every three hours for the full year 2008.
7 A monitoring network composed of nine fixed stations provided hourly ob-
8 servations to be assimilated. Every three hours, we compute the so-called BLUE
9 (best linear unbiased estimator), which is a concentration field merging ADMS
10 outputs and ground observations. Its error variance is supposed to be min-
11 imal under given assumptions regarding the errors on observations and model
12 simulations. A key step lies in the modeling of error covariances between the
13 computed NO₂ concentrations across the city. We introduce a parameterized
14 covariance which heavily relies on the road network. The covariance between
15 two locations depends on the distance of each location to the road network
16 and on the distance between the locations along the road network. Efficient
17 parameters for the covariances are primarily chosen according to prior as-
18 sumptions, χ^2 diagnosis and leave-one-out cross-validations. According to
19 the cross-validations, the improvements due to the assimilation seem mod-
20 erate far from the observation network, but the root mean square error roughly
21 decreases by 30% to 50% in the main city where the station density is high.
22 The method is computationally tractable for the generation of improved con-
23 centration fields over a long period, or for day-to-day forecasts.

1. Introduction

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

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

39 Data assimilation has been applied in the air quality community, mostly at large scale
40 and with Eulerian chemistry-transport models [e.g., *Elbern and Schmidt*, 2001; *Segers*,
41 2002; *Chai et al.*, 2007; *Wu et al.*, 2008]. In this paper, we address the assimilation of
42 observations of an urban monitoring network in order to correct the concentrations of
43 nitrogen dioxide computed by a Gaussian urban air quality model (ADMS Urban). A key
44 step of the assimilation procedure is to model the error variance of the NO₂ concentration

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

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

2. Urban Air Quality Modeling over Clermont-Ferrand

2.1. ADMS Urban

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

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

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

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

88 determine the concentration field in the streets whose buildings are higher than 0.5 m.
89 This model is based on the Danish model Operational Street Pollution Model [*Hertel and*
90 *Berkowicz*, 1989]. For this work, the chemistry is quite simple: the NO₂ concentration is
91 determined from the NO_x concentration as described in *Derwent and Middleton* [1996].

2.2. Meteorological Data, Topography and Land Use

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

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

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

2.3. Emissions

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

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

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

2.4. Case Study

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

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

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

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

3. Assimilation Method

3.1. Problem Statement

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

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

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

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

3.2. Best Linear Unbiased Estimator (BLUE)

166 Based on c^b , B , o and R , the analysis state vector is computed as the so-called “Best
 167 Linear Unbiased Estimator” which is linearly dependent on c^b and o , has unbiased error
 168 $c^a - c^t$ and has a variance with minimum trace [see, e.g., *Bouttier and Courtier*, 1999].
 169 This estimator is uniquely defined as

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

where

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

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

3.3. Modeling of the Covariance Matrices

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

$$R = v_o I ,$$

174 where v_o is the observational error variance.

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

180 We introduce the distance d_{ij} along the road between two receptors indexed by i and
 181 j . The distance along the road is defined as the smallest distance it takes to travel on the
 182 road network between the two receptors. If the two receptors i and j are not located on
 183 a road, they are first orthogonally projected on the road network, and d_{ij} is taken as the
 184 distance along the road between the projections. We also introduce the distance P_i of the
 185 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

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

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

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

205 3.3.1. Specific examples

206 Between two receptors on the road network ($P_i = P_j = 0$), the state error covariance
 207 is equal to $\frac{1}{2}v_c$ when the distance between the receptors is $d_{ij} = 0.7L_d$. Between two
 208 receptors on the same normal to the road network ($d_{ij} = 0$) and on the same side,
 209 the state error covariance is equal to $\frac{1}{2}v_c$ when the distance between the receptors is
 210 $0.7(L_p + \alpha \min(P_i, P_j))$. By definition, this distance increases for background receptors.
 211 Between two receptors so that $P_i = P_j$, not necessarily on the road network, the covariance
 212 highly depends on the distance along the road network. Their errors correlation is equal
 213 to 1 if $d_{ij} = 0$: we assume that these two receptors are subject to the same errors.

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

218 3.3.2. Implementation

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

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

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

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

236 4.1.1. Observations and their Error Variances

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

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

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

258 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

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

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

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

4.2. Results

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

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

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

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

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

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

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

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

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

302 4.3.1. Scores

303 The cross-validation was carried out for the reference values $(v_c, L_d, L_p, \alpha) =$
304 $(220 \mu\text{g}^2 \text{m}^{-6}, 3000 \text{m}, 200 \text{m}, 1)$ from Section 4.1.2. The performance before assimilation
305 is given in Table 4. The results after assimilation are given in Table 5. The largest
306 improvements occur in urban area (at the station Jaude, the improvement is of 46%),
307 compared to peri-urban area (at the stations Gravanches and Royat, the improvements
308 are respectively of 17% and 5%). It is likely that the distance between the peri-urban
309 stations makes it difficult to obtain enough information to compute strong and reliable
310 corrections from one station to the other. Another possible explanation may be an un-
311 satisfactory modeling of the error covariances between peri-urban receptors or between
312 urban and peri-urban receptors. In some cases, the absolute bias increases but remains
313 inside the interval recommended in *Chang and Hanna* [2004] (see the first two columns
314 of Table 5).

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

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

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

324 **4.3.2. Sensibility to the Parameters of the State Error Covariance Matrix**

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

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

342 The impact of the decorrelation length transverse to the road network, L_p , is much more
343 limited. The optimal value of L_p is not clearly determined by the Figure 7. With $L_d =$
344 3000 m, the RMSE is almost identical for L_p equal to 200 m or 300 m. The RMSE at peri-

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

5. Conclusions

348 The paper demonstrates the efficiency of data assimilation at urban scale for the im-
349 provement of NO₂ concentration fields using fixed monitoring stations. Computing the
350 best linear unbiased estimator (BLUE) has proved to be efficient for the correction of prior
351 concentrations computed by the urban Gaussian model ADMS. Despite the low number
352 of stations available in the simulation domain, strong improvements (30% to 50%) were
353 found at urban monitoring stations excluded from the assimilation procedure, in a leave-
354 one-out cross-validation. This shows that, in the part of the domain where the station
355 density is high, large improvements are likely to occur at non-observed locations.

356 However, in the background, far from the monitoring network, the improvements are
357 low. It is not clear whether these low improvements at rural locations is due to lack
358 of information from the observation network or to shortcomings in the error covariance
359 modeling. In the algorithm, a key variable is indeed the error covariance matrix B that
360 determines the spatial distribution of the corrections. The proposed covariance matrix is
361 motivated by the prominent role of traffic emissions in urban NO₂ concentrations, but it
362 surely misses significant error correlations.

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

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

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

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

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

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

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

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

References

- 399 Airparif, Guide pratique d’utilisation pour l’estimation de l’incertitude de mesure des
400 concentrations en polluants dans l’air ambiant, *Tech. Rep. Version 9*, AIRPARIF, 2007.
- 401 Bouttier, F., and P. Courtier, Data assimilation concepts and methods, *Meteorological*
402 *training course lecture series*, ECMWF, 1999.
- 403 Chai, T., et al., Four-dimensional data assimilation experiments with International Con-
404 sortium for Atmospheric Research on Transport and Transformation ozone measure-
405 ments, *Journal of Geophysical Research*, 112(D12S15), doi:10.1029/2006JD007763,
406 2007.
- 407 Chang, J. C., and S. R. Hanna, Air quality model performance evaluation, *Meteorology*
408 *and Atmospheric Physics*, 87, 167–196, 2004.
- 409 Derwent, R. G., and D. R. Middleton, An empirical function for the ratio NO₂:NO_x, *Clean*
410 *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,

- 434 Greece, 2011.
- 435 Riishojgaard, L. P., A direct way of specifying flow-dependent background error correla-
436 tions for meteorological analysis systems, *Tellus*, 50A, 42–57, 1998.
- 437 Segers, A., Data assimilation in atmospheric chemistry models using Kalman filtering,
438 Ph.D. thesis, Delft University, 2002.
- 439 Tennekes, H., A model for the dynamics of the inversion above a convective boundary
440 layer, *Journal of the Atmospheric Sciences*, 30, 558–567, 1973.
- 441 Tennekes, H., and A. Driedonks, Basic entrainment equations for the atmospheric bound-
442 ary layer, *Boundary-Layer Meteorology*, 20, 515–229, 1981.
- 443 Wu, L., V. Mallet, M. Bocquet, and B. Sportisse, A comparison study of data assimilation
444 algorithms for ozone forecasts, *Journal of Geophysical Research*, 113(D20310), 2008.

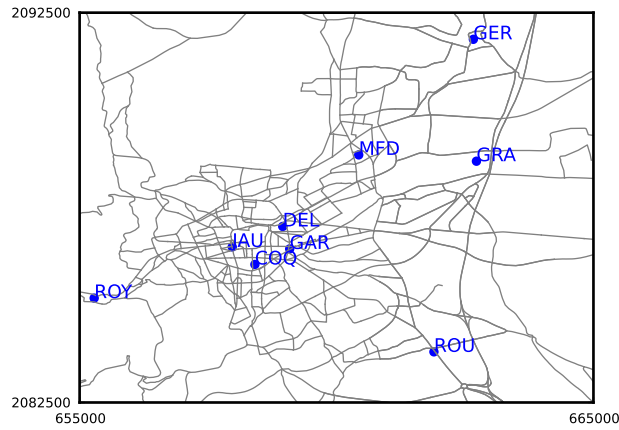


Figure 1. The modeled road network of the city of Clermont-Ferrand and the location of the observation stations: GAR stands for Gare, ROU for Roussillon, COQ for Lecoq, DEL for Delille, JAU for Jaude, MFD for Montferrand, GER for Gerzat, GAR for Gravanches and ROY for Royat. The coordinate projection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

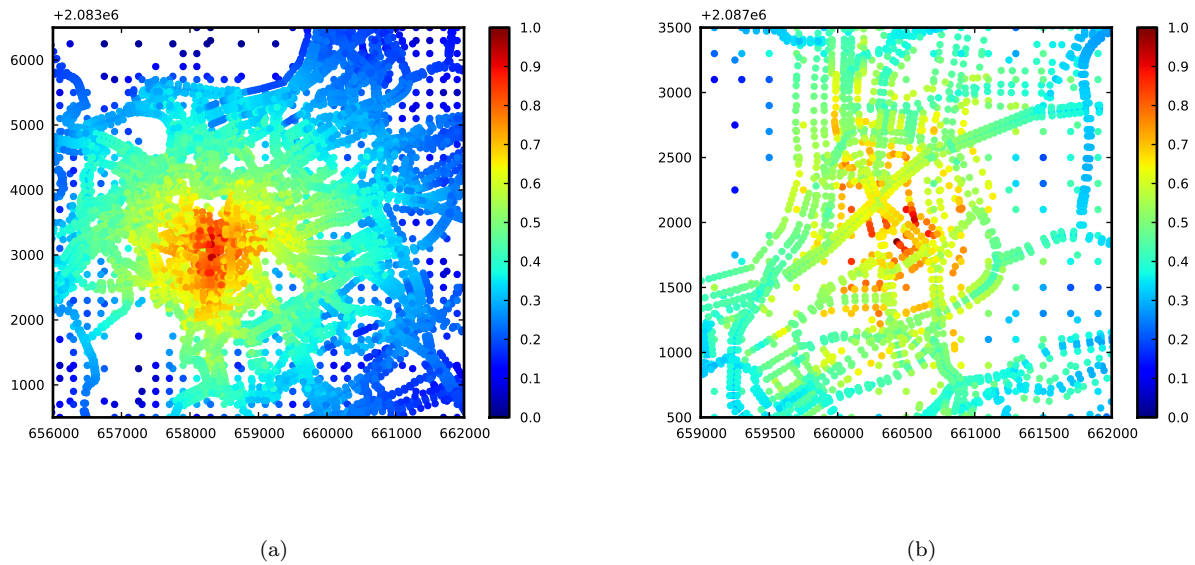
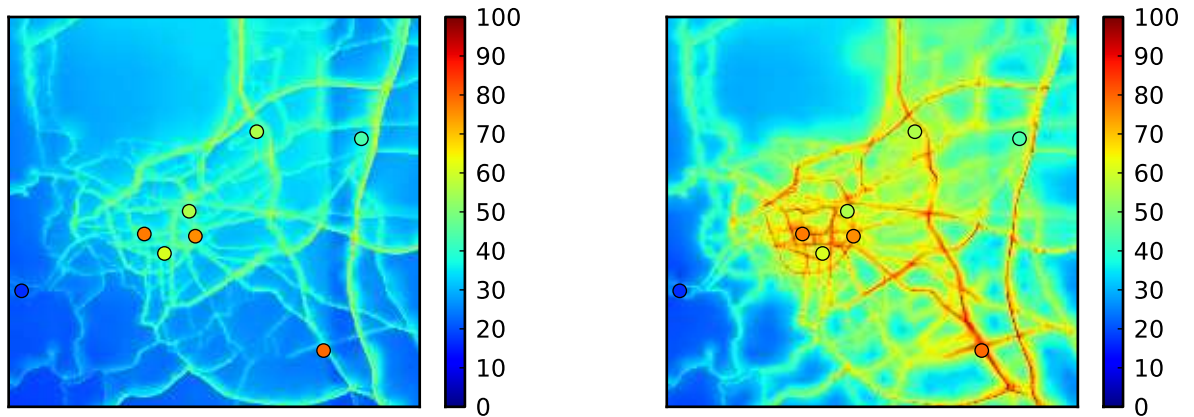


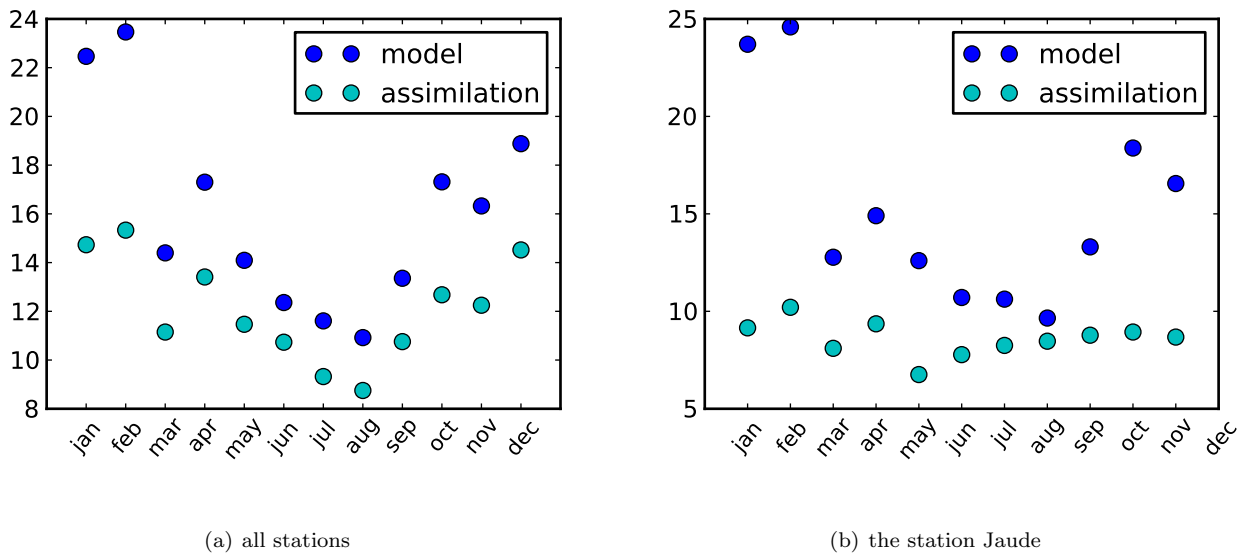
Figure 2. The state error correlations (B_{ij}/v_c) between the receptor located at (a) the station Gare or at (b) the station Montferrand and the other receptors. This corresponds to one row of the state error covariance matrix divided by v_c . Notice that the figures do not correspond to the same domain 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 $\mu\text{g m}^{-3}$. 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 $\mu\text{g m}^{-3}$ of the model in blue and after data assimilation in cyan, for (a) all stations and at (b) Jaude.

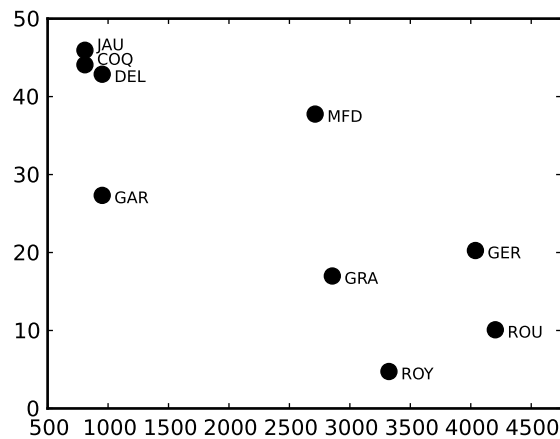


Figure 5. Improvement of stations RMSE in % (see Table 5), against the distance (m) to the rest of the network. See Figure 1 for the position of the stations.

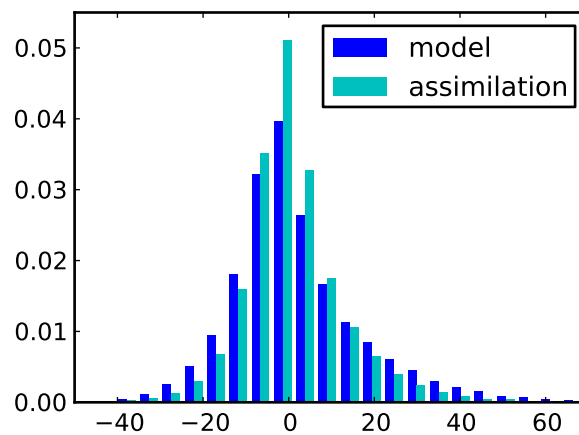


Figure 6. In blue the dispersion of the innovations, in cyan the dispersion of the discrepancy to observations after data assimilation (in leave-one-out cross-validation). The abscissa is a concentration discrepancy in $\mu\text{g m}^{-3}$ and the ordinate is the relative occurrence frequency.

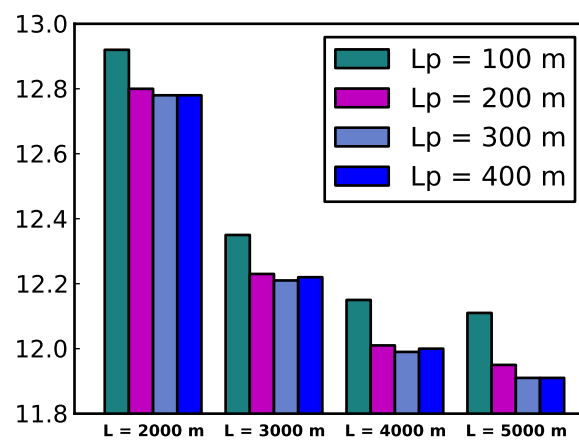


Figure 7. The RMSE in $\mu\text{g m}^{-3}$ over all stations for several pairs of parameters L_d and L_p .

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 c_i - o_i $
Normalized mean absolute error	$\frac{1}{n} \sum_{i=1}^n \frac{ c_i - o_i }{\bar{o}}$
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 $\mu\text{g m}^{-3}$, 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 $\mu\text{g}^2 \text{m}^{-6}$, 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 $\mu\text{g m}^{-3}$, 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 ($L_d = 3000$ m, $L_p = 200$ 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 $\mu\text{g m}^{-3}$ 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 $\mu\text{g m}^{-3}$ 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