What Can We Expect from Data Assimilation for Air Quality Forecast? Part I: Quantification with Academic Test Cases

LAURENT MENUT

Laboratoire de Météorologie Dynamique, Ecole Polytechnique, IPSL Research University, Ecole Normale Supérieure, Université Paris-Saclay, Sorbonne Universités, UPMC Université Paris 06, CNRS, Palaiseau, France

BERTRAND BESSAGNET[a]

Institut National de l'Environnement Industriel et des Risques, Verneuil en Halatte, France

(Manuscript received 5 January 2018, in final form 20 November 2018)

ABSTRACT

Data assimilation has been successfully used for meteorology for many years and is now used more and more for atmospheric composition issues (air quality analysis and forecast). The data assimilation of pollutants remains difficult and its deployment is currently in progress. It is thus difficult to have quantitative knowledge of what we can expect as the maximum benefit. In this study we propose a simple framework to make this quantification. In this first part, the gain of data assimilation is quantified using academic but realistic test cases over an urbanized polluted area and during a summertime period favorable to ozone formation. Different data assimilation configurations are tested, corresponding to different amounts of data available for assimilation. For ozone (O_3) and nitrogen dioxide (NO_2), it is shown that the benefit resulting from data assimilation lasts from a few hours to a possible maximum of 60 and 21 h, respectively. Maps of the number of hours are presented, spatializing the benefit of data assimilation.

1. Introduction

For analysis or forecast cases, one of the best ways to improve the results of a chemistry transport model is to better represent the physics and chemistry processes. Another way is to modify the trajectory by ''assimilating'' observations. Data assimilation consists of using hybridization methods of measurement data and modeling results to constraint the model prediction as close as possible to the observed data. The concept follows a simple principle whatever the studied physical problem: the closer the modeling values are to the real values of the measurements, the more we can expect better results for the places where there are no measurements. In addition, even if the model is nonlinear, the assimilation is better when observations are available, the forecast is better. But it is clear that data assimilation cannot increase our scientific knowledge: this is just a way to have better results, without any explanation as to why these results were less correct without data assimilation.

Data assimilation is widely used for meteorology (Talagrand 1997), and the applications cover three possible ways: analysis, inverse modeling, and forecast. More recently, data assimilation was also developed for atmospheric chemical composition. Several review articles were published, such as Sandu and Chai (2011) and Bocquet et al. (2015). They extensively describe the numerous data assimilation techniques, the strengths and weaknesses of the systems, the dependence on the studied chemical species (their abundance, kinetics, data availability, etc.). Nowadays, data assimilation of species is mainly used for analysis and inverse modeling.

Analysis is used to build a better data field after an event (Denby et al. 2008; Constantinescu et al. 2007; Curier et al. 2012). As an example, for a climatological study, a simulation may be performed over several years. If the goal is not to validate the model but to estimate the more realistic trend for a parameter, then the use of data assimilation gives better results

[a] Current affiliation: Hangzhou Futuris Environmental Technology Co. Ltd., Hangzhou, Zhejiang, China.

Corresponding author: Laurent Menut, menut@lmd.polytechnique.fr

DOI: 10.1175/JTECH-D-18-0002.1

than the first-guess simulation (Pierce et al. 2007). For air quality purposes, the first studies in Europe were the 4D-VAR by Elbern and Schmidt (1999) and the optimal interpolation approaches by Blond and Vautard (2004), and Zheng et al. (2018) was dedicated to building more realistic databases of surface ozone peaks or to improving the forecasts of $PM_{2.5}$.

For inverse modeling, a large number of studies were performed over the last 20 years. They have been especially applied at the global scale for the inversion of emissions of long-lived chemical species such as methane (Wang and Bentley 2002), carbon dioxide (Kaminski et al. 2001), chlorofluorocarbons (CFCs; Mahowald et al. 1997), and carbon monoxide (Bergamaschi et al. 2000; Müller and Stavrakou 2005), and at the continental scale to nitrogen oxides (NO_x; Wang et al. 2004; Konovalov et al. 2006). This methodology is completely different from the analysis approach, since it is designed to estimate input data by assimilating measurements related to model output data. At the regional scale, the problem becomes rather difficult to solve because the model considers explicitly shorter-living species, and the errors in meteorology, turbulence, and deposition dominate the system (Chang et al. 1997; Mendoza-Dominguez and Russell 2001; Enting 2002; Elbern et al. 2007; Pison et al. 2006, 2007; among many others).

The forecast is the most recent application of chemical data assimilation techniques. The scarcity of studies is not due to a lack of interest but to the numerous difficulties in conducting them. Even if surface and satellite measurements become available in near–real time, all models are not able to use them. The species of interest for air quality are not all available and the satellite sensors do not have high accuracy close to the surface (Zhang et al. 2012). For example, some measurements are available for ozone only (the photochemical reactions are thus difficult to constrain), particulate matter mass (the aerosol speciation is often missing), or aerosol optical depth (an estimate of the radiative impact but without information about the chemical composition or the altitude of the layers). Today, numerous systems exist and, as recent examples, there is the Prévisions et Observations de la Qualité de l'Air en France et en Europe (PREV'AIR) system (the first European operational air quality forecast; Honoré et al. 2008) and the European Copernicus program (Marécal et al. 2015).

Data assimilation of chemical species was found to always improve the results, including for forecasting. But the question of quantification of this benefit remains open. In this study we propose a simple approach to estimate this benefit. The starting point is that there is no need for a data assimilation system to estimate its potential gain. The key question is, If we have perfect initialization, for how many hours will the air quality forecast system be better than without this perfect initial state? To answer this question, it is not necessary to try to develop very complex systems: an academic test case can be defined to control all variables. Then, we just have to evaluate the differences between several simulations: one representing the observations and one representing the forecast with the model as is. Thus, we consider that (i) the model is state of the art and (ii) the data assimilation algorithms are perfect—that is, the forecast starts with a perfect initialization of the model. With this methodology, we can provide answers giving the maximum benefits we can expect for regional forecast applications.

Section 2 presents the methodology. Section 3 presents the academic test case (meteorology, emissions, boundary conditions). Section 4 presents the observation dataset. Section 5 presents the principle of the pseudo data assimilation system and the results. Conclusions are presented in section 6.

2. Methodology

The methodology consists of using the same model with three different configurations, as described in Fig. 1. By comparing the simulation results, it is possible to quantify the benefit obtained with data assimilation. First, the main principle of the methodology is presented. Note that this methodology could be used with either academic or realistic simulations. More realistic simulations, focused on a PM pollution episode, are presented in Bessagnet and Menut 2018, manuscript submitted to *J. Atmos. Oceanic Technol.*, hereafter Part II).

The three simulations are designed as follows:

FCST: The model runs for the period $[0, t_2]$. This corresponds to the usual methods of state-of-the-art forecast simulations. This includes errors in the meteorology and emissions, and the simplifications resulting from parameterizations, among other possible model errors.

OBS: The model is used with perturbed meteorological variables and surface emissions fluxes during the period $[0, t_2]$. This simulation represents the pseudo-observations dataset.

DA: This simulation represents the results after data assimilation. The simulation during the period $[t_1, t_2]$ uses the FCST meteorology and emissions but is initialized at t_1 with "assimilated data" obtained from OBS.

The benefit of data assimilation is quantified by comparing these three simulations (FCST, OBS, and DA).

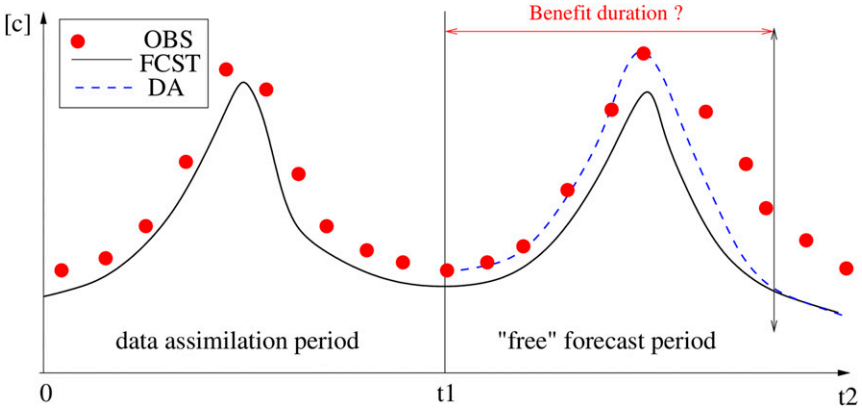

FIG. 1. Main principle of the data assimilation procedure. A forecast without data assimilation, FCST, runs during a period $[0, t_2]$. The observations, OBS, are available during the period $[0, t_1]$ in forecast conditions and during the period $[0, t_2]$ after the predicted event. Using OBS as the initial conditions at time t_1, the forecast with data assimilation, OPT, can run between t_1 and t_2.

In the example presented in Fig. 1, the "first guess" simulation—FCST— underestimates the observations, OBS. Using data assimilation, the forecast restarts with values closer to "reality" at time t_1, constituting the DA simulation. After some hours or days, and even if DA restarts with the initial state of OBS, DA will get closer to FCST, having the same forcings (meteorology, emissions, boundary conditions). For all model configurations, there is no doubt that the DA simulation will always become as uncertain as FCST. So, the question is not whether this will happen but after how long it will happen.

3. Definition of the academic test case

In this study we use academic test cases: the meteorology is realistic but completely constrained. It is the best way to really understand and interpret the results. The meteorology and emissions are chosen to be representative of a summertime pollution event. Another advantage of this configuration is that we can design several types of available measurements: only surface measurements, measurements close to satellite retrievals, etc. In addition, we can subset the different information per species, either temporally or spatially, to study very different configurations.

A specific preprocessing program was created and dedicated to the chemistry transport model used in this study: CHIMERE. This model is dedicated to the analysis and forecast of the atmospheric composition in the troposphere (Menut et al. 2013; Mailler et al. 2017). This model requires meteorological fields, chemical boundary conditions, and surface emission fluxes as forcings. All physical and chemical processes related to the spatial and temporal evolution of chemical species, gases, and aerosols are considered: transport, turbulence, chemistry, emissions, and wet and dry deposition.

The preparation of all input data is done for (i) the simulation domain, including the 3D mesh and land use; (ii) the meteorological fields; (iii) the chemical boundary conditions; and (iv) the surface anthropogenic and biogenic emissions. For simplicity, we consider no mineral dust, biomass burning, or sea salt emissions.

The simulation is performed for two periods, each lasting 5 days. The period is chosen as a summertime period, and we will focus our analysis on ozone concentrations. The first period $[0, t_1]$ ranges from 0000 UTC 1 June 2017 to 2400 UTC 5 June 2017. The second period $[t_1, t_2]$ ranges from 0000 UTC 6 June 2017 to 2400 UTC 11 June 2017.

a. *The model domain*

The domain consists of $41 \times 41 \times 20$ grid cells in the (x, y, z) dimensions, respectively, with a horizontal resolution of $0.2° \times 0.2°$. Vertically, the domain extends from the surface to 500 hPa to cover the boundary layer and a large part of the free troposphere. The first vertical level has a thickness of 20 m and then the thickness of the upper cells increases with altitude. The land cover is grassland for the whole domain, except at the center, where the land use is urban. The city is defined as a square of $0.5° \times 0.5°$ at the center of the domain. There is no orography and no sea or lake in the domain. This domain is similar to the Paris area and can thus be considered as realistic.

For the data assimilation management and the analysis of the results, we define several pseudostations, displayed in Fig. 2 and with coordinates displayed in

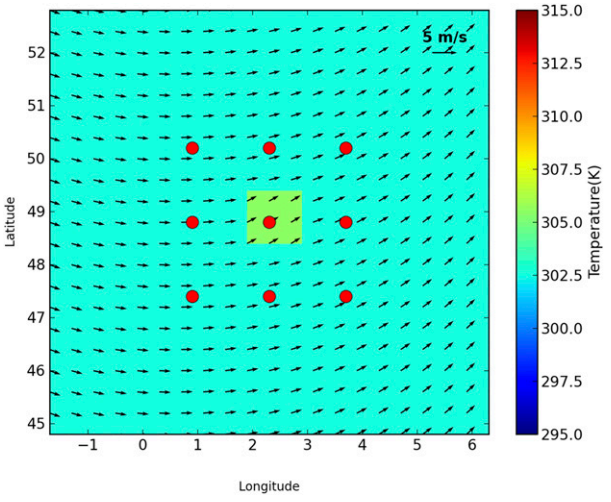

FIG. 2. Temperature (K) for a summertime period, at the first model level (20 m), at 1200 UTC, with a city in the center of the domain. Vectors represent the wind speed ($m\,s^{-1}$) in this first model level.

Table 1. These ''measurement stations'' correspond to locations where the data are considered as available and are then assimilated before the restart of the DA simulation.

b. *The meteorological fields*

The meteorological fields are calculated using very simplified parameterizations but in a realistic manner. The meteorology reproduces a summertime period with weak wind conditions, no clouds, and no precipitation, and is representative of a stagnation period, favorable to a pollution event (in particular, high ozone concentrations).

The pressure is constant in time and is horizontally over the domain but varies vertically. The two first levels are imposed: the surface pressure is 1000 hPa, and the top of the first model level is 997 hPa in order to constrain the first layer's thickness to be around 20 m. The top of the model domain is 500 hPa. Using these values, the pressure profile is estimated using exponential interpolation between the first and top levels. The altitude and thickness of each vertical cell are deduced from the pressure profile. The wind speed is temporally and horizontally constant, but it increases vertically using a logarithmic factor until the boundary layer height. The wind direction changes in the domain to reproduce a large-scale circulation.

TABLE 1. Names and coordinates of the pseudo surface stations. The stations correspond to the locations where surface measurement data are available and are assimilated in DA1.

Stations	Lon (°E)	Lat (°N)
measURB	2.3	48.8
measSW	0.9	47.4
measNW	0.9	50.2
measNE	3.7	50.2
measSE	3.7	47.4
measN	2.3	50.2
measS	2.3	47.4
measW	0.9	48.8
measE	3.7	48.8

TABLE 2. Minimum and maximum values of the time-varying meteorological parameters. For the urban cell in the center of the domain, a constant urban increment is added.

Variable	Min	Max	Urban
2-m temperature (°C)	15	30	+3
10-m wind speed ($m\,s^{-1}$)	2	2	−80%
2-m relative humidity (%/100)	0.6	1	−80%
Soil moisture ($m^3\,m^3$)	0.4	0.4	0
Boundary layer (m)	50	2000	0
Surface sensible heat flux ($W\,m^{-2}$)	−30	200	0
Surface latent heat flux ($W\,m^{-2}$)	−30	200	0
Shortwave radiation ($W\,m^{-2}$)	50	800	0

The other meteorological variables vary in space and time. We consider the influence of the city, which is located in the center of the domain. The 2-m temperature, the 10-m wind speed, and the 2-m relative humidity are modified following the values presented in Table 2. Diurnal cycles are considered for temperature, humidity, boundary layer height, surface heat fluxes, and shortwave radiation between the minimum and maximum values, which are also presented in Table 2. Between the minimum and maximum valued for each meteorological variable M, a simple sinusoidal expression is used to reproduce the diurnal cycle:

$$M = \frac{M_{\min} + M_{\max}}{2} + (M_{\max} - M_{\min}) \times \pi \frac{\sin(h-8)}{24}, \quad (1)$$

where h is the local hour (between 0 and 24).

An example of meteorological fields is presented in Fig. 2 with the 2-m temperature (colors) and the 10-m wind speed (vectors), for a typical summer day at 1200 UTC.

c. *The chemical boundary conditions*

The chemical boundary conditions are present only to preserve realistic orders of magnitude for the main studied chemical species, as explained in Table 3. The model species correspond to the Melchior mechanism used in CHIMERE and are fully described in Menut et al. (2013).

In this study only boundary conditions for gases are considered. The surface emissions of primary particles are considered, but there is no presence of biomass

TABLE 3. Constant boundary conditions chemical concentrations for each model species of the Melchior chemical mechanism.

Species	[c] (ppb)	Species	[c] (ppb)
O_3	30.0	CH4	1700.
NO	0.05	HCHO	0.7
NO_2	0.3	C2H6	0.5
HNO_3	1.0	NC4H10	0.08
PAN	0.1	C2H4	0.06
H_2O_2	1.0	C3H6	0.02
CO	80.0	OXYL	0.02

burning and mineral dust aerosol concentrations in this regional domain.

d. Anthropogenic emissions

Anthropogenic emissions are estimated for chemical species of the Melchior mechanism. The methodology follows the surface emission fluxes calculation described in Menut et al. (2012) and Mailler et al. (2017). For this model domain, composed of one city at the center and grassland and agricultural land in the surroundings, the emissions are uniform over each of these land uses. The values were extracted from the Hemispheric Transport of Air Pollution (HTAP) emissions inventory (Janssens-Maenhout et al. 2015) for the Paris, France, area in June.

4. The preparation of pseudo-observation dataset

The OBS are built using the simulation FCST but with additional perturbations for some parameters. These perturbations are chosen to be in a realistic range of known uncertainty for each parameter (Table 4). The perturbation values come from the usual known uncertainties, as already used in CHIMERE in Menut (2003), among others. OBS is performed for the same period as FCST, [0, t_2]. The perturbation is calculated with a constant bias (systematic error) and a scatter (random error). The systematic error is different for each variable. It also changes every day. To avoid stiff day-to-day differences, the perturbation is smoothed using a binomial filter. Considering that a forecast error may be spatially persistent over a regional domain, the perturbation is the same for all domain cells.

TABLE 4. Uncertainties for the perturbed meteorological parameters, after Menut (2003).

Parameter	Error type: Bias (systematic)	Error type: Scatter (random)
Anthropogenic emissions		
NO_x (%)	−40	±40
VOCs (%)	−40	±40
Boundary conditions (gas) (%)	−30	±20
Meteorology		
Wind components u and v ($m\,s^{-1}$)	0	±1
Temperature (K)	0	±3
Boundary layer height (%)	0	±20

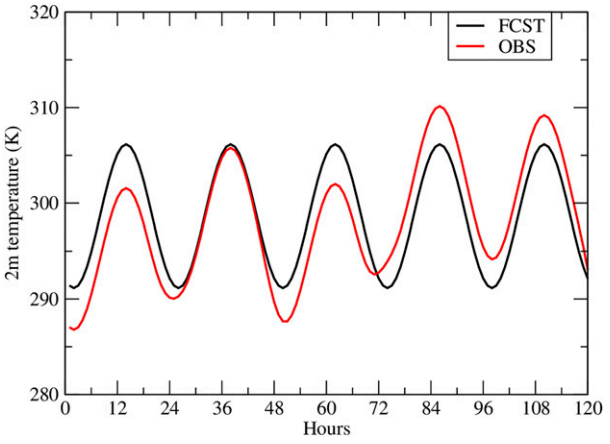

FIG. 3. Time series of 2-m temperature for the urban cell. The two model configurations are represented as FCST (current version of the model) and OBS (perturbed version of the model and pseudo observations).

For the meteorology, three variables are perturbed: the wind speed (zonal and meridional components), the temperature, and the boundary layer height. Figure 3 presents the time series of 2-m temperature for the "urban" cell at the center of the domain. The FSCT simulation has the same diurnal cycle every day. The OBS simulation corresponds to FCST but after multiplying the variables by the perturbation. Note that the wind components u and v are perturbed with the same factor. In the CHIMERE model, the vertical transport w is always diagnosed from the u and v values known at each model cell interface. It is thus possible to randomly change the zonal and meridional wind components, and to ensure mass conservation for the transport calculation.

For the anthropogenic emissions, NO_x and volatile organic compound (VOC) fluxes are perturbed. In addition to the meteorological variables, we consider for emissions a bias and an uncertainty (random error). The bias is constant and represents an example of poor knowledge of what is really emitted in a city. The variability is applied after the bias: for example and for the NO_x fluxes, a random perturbation of ±40% is applied after the bias effect of −40%. This bias is realistic knowing the current available regional inventories.

Representative of large-scale chemical concentrations fields, the boundary conditions are also perturbed. All chemical species are changed with a constant negative bias of 30%. This bias represents the fact that if anthropogenic emissions are underestimated at the regional

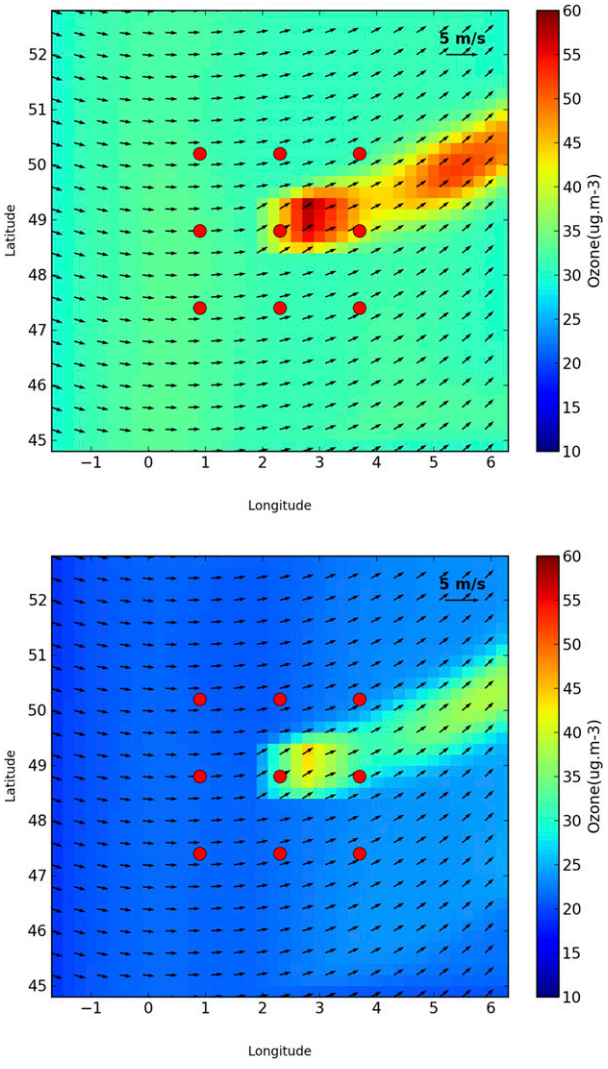

FIG. 4. Surface ozone concentrations ($\mu g\,m^{-3}$) for the sixth day of the simulation at 1200 UTC. (top) The FSCT simulation and (bottom) the OBS simulation (corresponding to FCST with perturbations).

scale, there is a chance to have the same effect at a larger scale, and then on the chemical concentrations transported in the regional domain, and then on these boundary conditions.

Results with surface concentration maps

Surface concentration maps of ozone for the FCST and OBS simulations are presented in Fig. 4. The surface ozone concentration values are presented for 1200 UTC 6 June 2017, corresponding to the first day of the second simulation period. For the FCST simulation, the maximum values are modeled downwind of the city and follow the mean wind flow. Two local maxima are identified: close to the city for ozone just formed in the preceding hours and in the northeast part of the domain for the ozone produced the day before. For the OBS simulation, the ozone plume follows the same trajectory. The wind speed and direction are perturbed but with a maximum of 10% only. The surface ozone concentrations are lower and represent a lower photochemical production as a result of lower surface emissions of NO_x and VOCs, and lower boundary conditions of ozone.

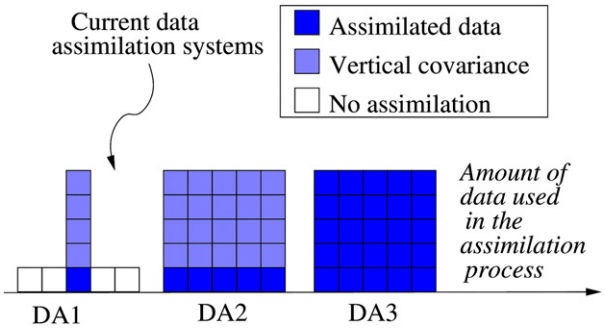

FIG. 5. Synthetic presentation of the "data assimilation" cases. All cases correspond to the idealized setup and are not designed to be realistic but only to represent minimum and maximum possible benefits of data assimilation. The "real" current systems are between DA1 and DA2.

5. The pseudo data assimilation

a. *The data assimilation cases*

The studied cases—DA1, DA2, and DA3—are presented in Fig. 5 and described in this section.

1) DA1

Of all the defined studied cases, DA1 is the closest to the current systems. The available surface data correspond to current regional air quality networks, thus at the surface and for some locations only. These locations may be in and around urbanized areas, defining the urban and suburban sites, respectively. For this configuration, the first model level of FCST is replaced by OBS for the initialization step. In the boundary layer, above these stations, the FCST concentration is corrected following a "pseudo vertical covariance" relationship. This pseudo vertical covariance is applied to the "assimilated concentrations." For this academic study, the simulation corresponds to a low-wind-speed case with a grid cell with 0.2° width: considering numerical diffusion, there is no need to add horizontal error covariance. It is different in the vertical dimension, where the vertical mixing acts quickly and efficiently in the convective boundary layer. A correction is thus applied in the boundary layer as follows:

$$c^{\mathrm{FSCT}}_{z=2,z_{\mathrm{ABL}}} = \max\left[0, c^{\mathrm{OBS}}_{z=2,z_{\mathrm{ABL}}} + (c^{\mathrm{OBS}}_{z=1} - c^{\mathrm{FSCT}}_{z=1})\right], \quad (2)$$

where the model vertical levels extend from 1 to z_{ABL} (the altitude of the boundary layer). This simple relation is defined to report the error correction diagnosed from the surface to upper levels. It is not a real ''vertical error covariance'' correction, but it is able to reproduce the benefit we can calculate when assimilating surface data. Note that negative values of concentrations are not allowed.

2) DA2

In this case we consider we have enough information close to the surface to have the complete first model level identical to the observations. This configuration does not exist yet. It is based on the same principle as DA1, but it is applied to all model cells. This case is more complete than the current existing systems.

3) DA3

We consider we have enough information (boundary layer and free troposphere) to have the whole model domain identical to the observations. This is the ultimate ''data assimilation'' system, since in this case we consider the available data to be numerous and the data assimilation system to be ''perfect.'' This could correspond to future combined in situ and satellite observation systems, where all chemical species of interest are measured with high spatial and temporal resolution. The complete domain used the concentrations calculated with OBS as a restart for the FCST simulation.

Note that there is no case really similar to existing data assimilation systems, even if DA1 is relatively close to the state of the art. These systems are very complex, and the goal of this paper is not to reproduce one of these configurations. The cases are thus defined to be clearly less or more efficient than the existing systems.

To quantify the time when the DA simulation, starting with OBS chemical values, reaches the values of the FCST simulation, a simple criterion is used:

$$B = \frac{|c_{\mathrm{DA}} - c_{\mathrm{FCST}}|}{|c_{\mathrm{DA}} - c_{\mathrm{OBS}}|}. \quad (3)$$

This calculation represents the difference between the DA simulation and FCST and OBS: the closer the DA simulation is to FCST, the smaller B is. We define a threshold value of $B_t = 0.1$ (i.e., 10%), corresponding to a 90% loss in the benefit of restarting with OBS. It is arbitrary, but the results show that another value would not have changed the conclusions. To avoid the division of small values by other small values, another threshold is fixed: the number of hours is calculated only if the concentrations c_{FCST} or the difference $|c_{\mathrm{DA}} - c_{\mathrm{FCST}}|$ is larger than $0.1\,\mu\mathrm{g\,m^{-3}}$. For lower values, we consider that the ratio is not significant and that there is no gain, since the FCST, OBS, and DA are already very close.

b. Results with time series

The results are presented as a time series in Fig. 6. The abscissa axis represents the number of hours after t_1. Surface concentrations of O_3 and NO_2 are compared in the simulations FCST and OBS and the scenarios DA1–DA3. Two stations are selected: the station located in the city center (measURB) and the station located downwind of the city (measNE). Results were also studied for some other stations, but they provided no valuable additional information.

Because of the emissions perturbations, the FCST surface ozone and NO_2 concentrations are larger than the OBS concentrations. The benefit is studied here by using the B criterion. For the two sites and the two species, it is noteworthy that only the DA3 configuration is able to propose a benefit more important than a few hours. For the configurations DA1 and DA2, the benefit vanishes after 3–7 h.

During these 3–7 h, there is no convection and photochemistry. For DA1, the concentrations are assimilated at station locations only. The benefit is low because the horizontal transport dominates the potential vertical motions: even if the concentrations are updated for the first hour at the stations, the benefit is annihilated after a few hours as a result of advection. For DA2, the complete surface level is assimilated. Even with this configuration, the benefit remains low and does not exceed a few hours.

The only configuration with a remarkable benefit is DA3, when the whole atmosphere receives data assimilation. In this case, the benefit exceeds 24 and 36 h for ozone and for measURB and measNE, respectively. For NO_2, the benefit is lower and is mainly significant for the urbanized site, with 16 h. The way the concentrations changed from OBS to FSCT is different for the two species. For ozone, the shift is sudden and with a duration of a few hours only: for example, in measURB, the DA3 case shows ozone concentrations close to OBS during the first 26 h and then reaching the FCST values in only 1 h. The same tendency is observed for measNE, when the ozone concentrations of DA3 change from OBS to FCST calculated concentrations in only 4 h. For NO_2, since the beginning of the simulation, the DA3 concentration is close to FCST. These differences are due to the lifetime of these species. Ozone is a secondary pollutant, with a lifetime of several days. Thus, the boundary conditions and the vertical mixing play a more important role than the local production. On the other

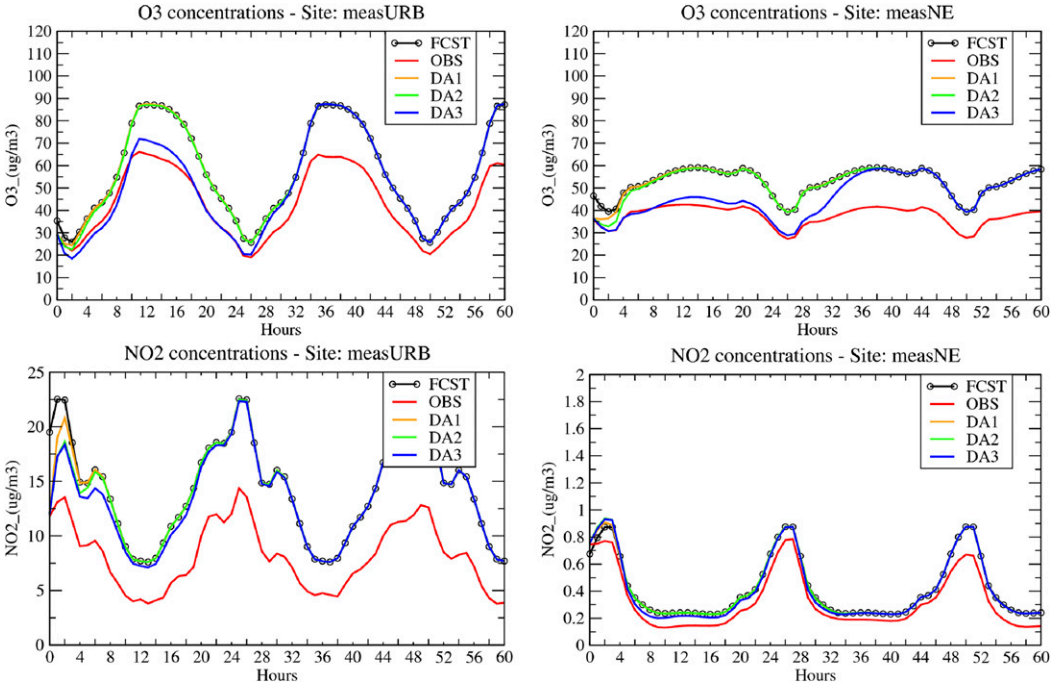

FIG. 6. Time series of surface concentrations ($\mu g\,m^{-3}$) for (top) O_3 and (bottom) NO_2. The simulations for FCST, OBS, and the five DA cases are presented in each panel and for the two sites, (left) measURB and (right) measNE.

hand, NO_2 is directly emitted by local emissions and is rapidly converted during the summertime period in the presence of oxidants like ozone. In this case, the local emissions are the key factor in explaining the modeled time evolution. This is why, even if DA3 initializes the whole domain, the emissions injected in the following hours will quickly suppress the local benefit of data assimilation.

c. *Results with benefit maps*

To evaluate the benefit on maps, values of B are calculated for each hour and all model surface cells. When $B < B_t$, the corresponding hour is stored. Results are then presented as maps of hours showing the end of the benefit period. The calculation is performed for O_3 and NO_2 surface concentrations.

Results are presented in Fig. 7 for ozone. For DA1, the benefit is 6 h at the stations. The benefit is larger for DA2 and may reach 10 h in the ozone plume. Finally, with DA3, the benefit increases again and may reach 60 h downwind of the city. On the western part of the domain, the benefit is close to zero, showing that the advection of boundary conditions instantaneously suppresses a potential benefit. The wind being from west to east, injecting assimilated data into the domain, has an impact increasing with time and following the mean flow. It is also interesting to note that the benefit tends to zero in on the southeastern part of the domain, where there is also a wind entering the domain through the boundary conditions. The high benefit values on the northwestern part of the domain are mainly due to the fact there are no city (no ozone fast titration by NO_x)—mainly biogenic—VOCs, which are favorable to ozone production. The whole column of ozone is updated during the initialization; these high concentrations values are the reason for this longer benefit at the surface.

For NO_2, results are presented in Fig. 8. The benefit is more local and mainly downwind of the city. In the city the benefit is due to the replacement of ozone, NO, and NO_2, where the main anthropogenic emissions occur. The effect is thus to counterbalance the bias in emissions with more realistic concentrations. Here, the assimilation has a positive (but short) effect by compensating for the discrepancies in emissions.

For DA2, the benefit reaches 10 h. During these hours—from 0000 to 1000 UTC—there is no chemistry and the impact is mainly due to the unperturbed anthropogenic emissions, slowly transported to the northeast. The increase in the number of hours is thus just the reflection of this transport of NO_2. When the photochemistry starts, as well as the vertical mixing, the benefit fades with time. The addition of the vertical covariance slightly increases the number of hours. For DA3, the benefit has a maximum of 15 h even close to the city. The shape of the maxima over the city is

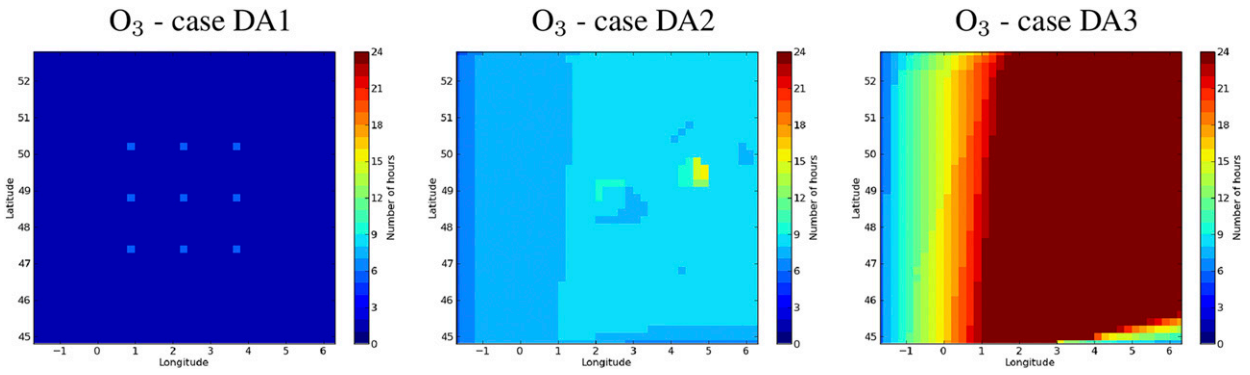

FIG. 7. Number of hours of data assimilation benefit for ozone. The three data assimilation test cases with different initializations of the simulation are presented.

different because in this case the complete vertical column is assimilated for all species: the gain, mixed with the advection, enhanced the number of hours of benefit, especially upwind of the city, where the surface concentrations of NO_2 are low.

Results presented on the maps are summarized in Table 5: for each pollutant and for each test case (DA1, DA2 and DA3), the lower and higher numbers of hours are extracted.

6. Conclusions

If the current data assimilation systems are able to improve the analysis and the forecast of regional air quality, then it is also important to quantify the number of hours of this benefit. In this study, we propose a simple framework based on academic test cases. The goal is to reproduce the equivalent of a forecast having initial conditions improved using data assimilation. By comparing two simulations (one with and one without some assimilated data), we are able to estimate the number of hours of benefit of the data assimilation.

An academic test case of meteorology and pollution was defined, corresponding to a summertime period, over a region similar to the Paris area and for an anticyclonic situation, favorable to a pollution event. We simulated the assimilation of a few surface stations (DA1), the whole surface (DA2), and the whole troposphere (DA3). The current existing systems are between DA1 and DA2. For the case DA1, it was shown that the benefit is less than 10 h for ozone and NO_2. For DA2, the maxima are 15 and 11 h for ozone and NO_2, respectively. In this case ozone, a secondary species, is able to be transported longer; thus, the benefit is higher than the primary species, NO_2. This effect increases with the case DA3. This case showed that, considering that we have a lot of data to assimilate, the maximum benefit would be 60 h for ozone and 21 h for NO_2.

The configuration presented in this study is a specific case and is not representative of all possible cases. But this selected case corresponds to a summertime pollution with low wind speed and a temperature up to 25°C during the afternoon. A faster wind will dampen the effects of data assimilation. This episode of stagnation

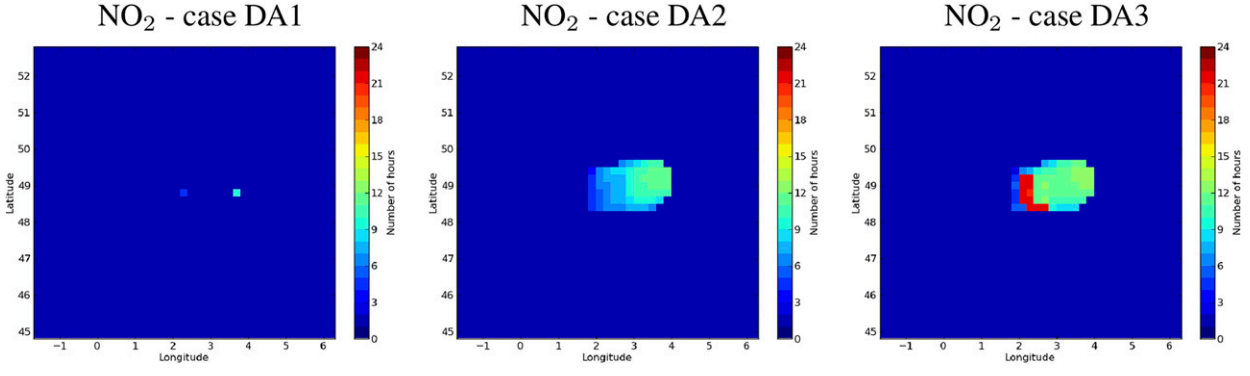

FIG. 8. Hours of data assimilation benefit for NO_2. The three data assimilation test cases with different initializations of the simulation are presented.

TABLE 5. Summary of the number of hours of benefit over the domains. The "Min" and "Max" values are the extrema observed whatever their location on the modeled domain.

	DA1		DA2		DA3	
Parameter	Min	Max	Min	Max	Min	Max
Ozone	0	5	4	15	5	60
NO_2	0	8	0	11	0	21

thus represents a maximum of possible gain. To strengthen these results, made using academic test cases, Part II is dedicated to the same kind of quantification but for a more real test case over Europe, including an analysis for gaseous and aerosol chemical species.

Acknowledgments. The authors acknowledge their colleagues Sylvain Mailler and Guillaume Siour of the CHIMERE development team for their fruitful comments and discussions. This work was partly funded by the French Ministry in Charge of Ecology (MTES).

REFERENCES

Bergamaschi, P., R. Hein, M. Heinmann, and P. Crutzen, 2000: Inverse modeling of the global CO cycle: 1. Inversion of CO mixing ratio. *J. Geophys. Res.*, **105**, 1909–1927, https://doi.org/10.1029/1999JD900818.

Blond, N., and R. Vautard, 2004: Three-dimensional ozone analyses and their use for short-term ozone forecasts. *J. Geophys. Res.*, **109**, D17303, https://doi.org/10.1029/2004JD004515.

Bocquet, M., and Coauthors, 2015: Data assimilation in atmospheric chemistry models: Current status and future prospects for coupled chemistry meteorology models. *Atmos. Chem. Phys.*, **15**, 5325–5358, https://doi.org/10.5194/acp-15-5325-2015.

Chang, M., D. Hartley, C. Cardelino, D. Haas-Laursen, and W. Chang, 1997: On using inverse methods for resolving emissions with large spatial inhomogeneities. *J. Geophys. Res.*, **102**, 16 023–16 036, https://doi.org/10.1029/97JD00964.

Constantinescu, E. M., A. Sandu, T. Chai, and G. R. Carmichael, 2007: Assessment of ensemble-based chemical data assimilation in an idealized setting. *Atmos. Environ.*, **41**, 18–36, https://doi.org/10.1016/j.atmosenv.2006.08.006.

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

Denby, B., M. Schaap, A. Segers, P. Builtjes, and J. Horálek, 2008: Comparison of two data assimilation methods for assessing PM_{10} exceedances on the European scale. *Atmos. Environ.*, **42**, 7122–7134, https://doi.org/10.1016/j.atmosenv.2008.05.058.

Elbern, H., and H. Schmidt, 1999: A four-dimensional variational chemistry data assimilation scheme for Eulerian chemistry transport modeling. *J. Geophys. Res.*, **104**, 18 583–18 598, https://doi.org/10.1029/1999JD900280.

——, A. Strunk, H. Schmidt, and O. Talagrand, 2007: Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys.*, **7**, 3749–3769, https://doi.org/10.5194/acp-7-3749-2007.

Enting, I., 2002: *Inverse Problems in Atmospheric Constituent Transport*. Cambridge Atmospheric and Space Science Series, Cambridge University Press, 394 pp.

Honoré, C., and Coauthors, 2008: Predictability of European air quality: Assessment of 3 years of operational forecasts and analyses by the PREV'AIR system. *J. Geophys. Res.*, **113**, D04301, https://doi.org/10.1029/2007JD008761.

Janssens-Maenhout, G., and Coauthors, 2015: HTAP_v2.2: A mosaic of regional and global emission grid maps for 2008 and 2010 to study hemispheric transport of air pollution. *Atmos. Chem. Phys.*, **15**, 11 411–11 432, https://doi.org/10.5194/acp-15-11411-2015.

Kaminski, T., M. Heimann, P. Peylin, P. Bousquet, and P. Ciais, 2001: Inverse modeling of atmospheric carbon dioxide fluxes. *Science*, **294**, 259, https://doi.org/10.1126/science.294.5541.259a.

Konovalov, I., M. Beekmann, A. Richter, and J. Burrows, 2006: Inverse modelling of the spatial distribution of NO_x emissions on a continental scale using satellite data. *Atmos. Chem. Phys.*, **6**, 1747–1770, https://doi.org/10.5194/acp-6-1747-2006.

Mahowald, N., R. Prinn, and P. Rasch, 1997: Deducing CCl_3F emissions using an inverse method and chemical transport models with assimilated winds. *J. Geophys. Res.*, **102**, 28 153–28 168, https://doi.org/10.1029/97JD02086.

Mailler, S., and Coauthors, 2017: CHIMERE-2017: From urban to hemispheric chemistry-transport modeling. *Geosci. Model Dev.*, **10**, 2397–2423, https://doi.org/10.5194/gmd-10-2397-2017.

Marécal, V., and Coauthors, 2015: A regional air quality forecasting system over Europe: The MACC-II daily ensemble production. *Geosci. Model Dev.*, **8**, 2777–2813, https://doi.org/10.5194/gmd-8-2777-2015.

Mendoza-Dominguez, A., and A. Russell, 2001: Estimation of emission adjustments from the application of four-dimensional data assimilation to photochemical air quality modeling. *Atmos. Environ.*, **35**, 2879–2894, https://doi.org/10.1016/S1352-2310(01)00084-X.

Menut, L., 2003: Adjoint modelling for atmospheric pollution process sensitivity at regional scale. *J. Geophys. Res.*, **108**, 8562, https://doi.org/10.1029/2002JD002549.

——, A. Goussebaile, B. Bessagnet, D. Khvorostiyanov, and A. Ung, 2012: Impact of realistic hourly emissions profiles on modelled air pollutants concentrations. *Atmos. Environ.*, **49**, 233–244, https://doi.org/10.1016/j.atmosenv.2011.11.057.

——, and Coauthors, 2013: CHIMERE 2013: A model for regional atmospheric composition modelling. *Geosci. Model Dev.*, **6**, 981–1028, https://doi.org/10.5194/gmd-6-981-2013.

Müller, J.-F., and T. Stavrakou, 2005: Inversion of CO and NO_x emissions using the adjoint of the IMAGES model. *Atmos. Chem. Phys.*, **5**, 1157–1186, https://doi.org/10.5194/acp-5-1157-2005.

Pierce, R. B., and Coauthors, 2007: Chemical data assimilation estimates of continental U.S. ozone and nitrogen budgets during the Intercontinental Chemical Transport Experiment–North America. *J. Geophys. Res.*, **112**, D12S21, https://doi.org/10.1029/2006JD007722.

Pison, I., L. Menut, and N. Blond, 2006: Inverse modeling of emissions for local photo-oxidant pollution: Testing a new methodology with kriging constraints. *Ann. Geophys.*, **24**, 1523–1535, https://doi.org/10.5194/angeo-24-1523-2006.

——, ——, and G. Bergametti, 2007: Inverse modeling of surface NO_x anthropogenic emissions fluxes in the Paris area during the Air Pollution Over Paris Region (ESQUIF) campaign. *J. Geophys. Res.*, **112**, D24302, https://doi.org/10.1029/2007JD008871.

Sandu, A., and T. Chai, 2011: Chemical data assimilation—An overview. *Atmosphere*, **2**, 426–463, https://doi.org/10.3390/atmos2030426.

Talagrand, O., 1997: Assimilation of observations: An introduction. *J. Meteor. Soc. Japan*, **75**, 191–209, https://doi.org/10.2151/jmsj1965.75.1B_191.

Wang, Y., and S. Bentley, 2002: Development of a spatially explicit inventory of methane emissions from Australia and its verification using atmospheric concentration data. *Atmos. Environ.*, **36**, 4965–4975, https://doi.org/10.1016/S1352-2310(02)00589-7.

——, M. McElroy, T. Wang, and P. Palmer, 2004: Asian emissions of CO and NO_x: Constraints from aircraft and Chinese station data. *J. Geophys. Res.*, **109**, D24304, https://doi.org/10.1029/2004JD005250.

Zhang, Y., M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov, 2012: Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. *Atmos. Environ.*, **60**, 656–676, https://doi.org/10.1016/j.atmosenv.2012.02.041.

Zheng, H., J. Liu, X. Tang, Z. Wang, H. Wu, P. Yan, and W. Wang, 2018: Improvement of the real-time $PM_{2.5}$ forecast over the Beijing-Tianjin-Hebei region using an optimal interpolation data assimilation method. *Aerosol Air Qual. Res.*, **18**, 1305–1316, https://doi.org/10.4209/aaqr.2017.11.0522.