What can we expect from data assimilation for air quality forecast? 1: Quantification with academic test cases

Laurent Menut, Bertrand Bessagnet

JOURNAL OF ATMOSPHERIC AND OCEANIC TECHNOLOGY

What can we expect from data assimilation for air quality forecast? 1: 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 Univ Paris 06, CNRS, Route de Saclay, 91128 Palaiseau, France

BERTRAND BESSAGNET

Institut National de l'Environnement Industriel et des Risques, Verneuil en Halatte, 60550, Parc Technologique ALATA, France

Now at: Hangzhou Futuris Environmental Technology Co. Ltd, Zhejiang Overseas High-Level Talent Innovation Park, No. 998 WenYi Road, 311121, Hangzhou, Zhejiang, China

ABSTRACT

Data assimilation is successfully used for meteorology since many years and is now more and more used 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 a quantitative knowledge of what we can expect as maximum of 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 favourable to ozone formation. Different data assimilation configurations are tested, corresponding to different amount of data available for assimilation. For ozone (O_3) and nitrogen dioxide (NO_2), it is shown that the benefit due to data assimilation lasts from a few hours to a maximum possible of 60 and 21 hours, 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 way to improve results of 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 in using hybridation methods of measurement data and modelling results to constraint the model prediction the closest as possible to the observed data. The concept follows a simple principle whatever the studied physical problem: more the modelling values are close to the real values where are the measurements, more we can expect better results for the places where there is no measurements. In addition, even if the model is non-linear, better is the assimilation when observations are available, better should be the forecast. But this is clear that data assimilation cannot increase our scientific knowledge: this is just a way to have better results, without any explanation on 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 modelling 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 technics, 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 modelling.

The **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, the use of data assimilation gives better results 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; Zheng et al. 2018) dedicated

**Corresponding author address:* Laboratoire de Météorologie Dynamique, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau, France

E-mail: menut@lmd.polytechnique.fr

to build more realistic databases of surface ozone peaks or to improve $PM_{2.5}$ forecast.

For **inverse modelling**, a large number of studies were performed these last 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 and Heimann 2001), CFCs (Mahowald et al. 1997) and carbon monoxide (Bergamaschi et al. 2000; Müller and Stavrakou 2005) and, at a continental scale, to nitrogen oxides (NO_x) (Wang et al. 2004; Konovalov et al. 2005). This methodology is completely different from the analysis since it is designed to estimate an 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 explicitely shorter-living species and the errors on 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 technics. The scarcity of studies is not due to a lack of interest, but to the numerous difficulties to conduct them. Even if surface and satellite measurements become to be 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 have not a large 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) or Particulate Matter in mass (the aerosols speciation is often missing), or Aerosol Optical Depth (a budget of a radiative impact but without information of the chemical composition or the altitude layers). Today, numerous systems exist and, as recent examples, there is the PREVAIR 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 forecast. 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 of a data assimilation system to estimate its potential gain. The key question is: if we have a perfect initialization, during how many hours the air quality forecast system will 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 it is. Thus, we consider that (i) the model is in the state of the art, (ii) the data assimilation algorithms are perfect, i.e. the forecast starts with a perfect initialisation of the model. With this methodology, we can provide answers giving the maxima of 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 observations 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 in using the same model with three different configurations, as described in Figure 1. By comparing the simulations 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 academic or realistic simulation. More realistic simulations, focussed on a Particulate Matter (PM) pollution episode, are presented in the companion paper (Bessagnet and Menut 2018).

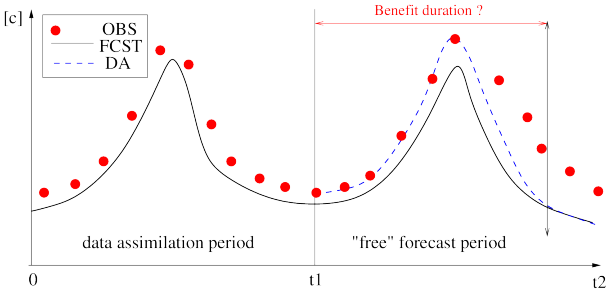

FIG. 1. *Main principle of the data assimilation procedure. A forecast without data assimilation, FCST, runs during a period [0:t_2]. The observations, OBS are available during the period [0:t_1] in forecast conditions and during the period [0:t_2] after the predicted event. Using OBS as initial conditions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

The three simulations are designed as:

- FCST: The model runs for the period [0:t_2]. This corresponds to the usual way to use it in forecast simulations, containing the current state-of-the-art. This includes errors in the meteorology and emissions, the simplifications due to the 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). In the example presented in Figure 1, the 'first-guess' simulation, FCST, underestimates the observations, OBS. Using data assimilation, the forecast restarts with values closer to the '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 acadamic 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 or temporally or spatially, to study very different configurations.

A specific pre-processing 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 atmospheric composition in the troposphere, (Menut et al. 2013), (Mailler et al. 2017). This model requires meteorological fields, chemical boundary conditions and surface emissions fluxes as forcings. All physical and chemical processes related to the spatial and temporal evolution of chemical species, gas and aerosol, are considered: transport, turbulence, chemistry, emissions, wet and dry deposition.

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

The simulation is performed for two periods of five 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 the 1st June 2017 at 00:00 UTC to the 5 June 2017 at 24:00 UTC. The second period [t_1:t_2] ranges from the 6 June 2017 at 00:00 UTC to the 11 June 2017 at 24:00 UTC.

a. The model domain

The domain is constituted of 41 $\times$ 41 $\times$ 20 grid cells in the (x,y,z) dimensions, with an horizontal resolution of 0.2 $\times$ 0.2 degrees. 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 20m, then the thicknesses of upper cells increase with altitude. The land cover is grassland for the whole domain, except at the center where the landuse is urban. The city is defined as a square of 0.5 $\times$ 0.5 degrees 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.

Stations	Longitude	Latitude
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 1. *Names and coordinates of the pseudo-surface stations. The stations correspond to the locations where surface measurements data are available and are assimilated in the first experiment called DA1 (see next section).*

For the data assimilation management and the analysis of the results, we define several pseudo-stations, displayed in Figure 2 and with coordinates displayed in Table 1. These "measurements 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 winds conditions, no clouds, no precipitation and is representative of a stagnation period, favourable to a pollution event (in particular high ozone concentrations).

The pressure is constant in time and 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 thickness to be around 20m. The top of the model domain is 500 hPa. Using these values, the pressure profile is estimated using an 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 vertically increases using a logarithmic factor until the boundary layer height. The wind direction changes in the domain to reproduce a large scale circulation.

Variable	Min	Max	Unit	Urban
2m temperature	15	30	^{o}C	+3^{o}C
10m wind speed	2	2	m s^{-1}	-80%
2m relative humidity	0.6	1	%/100	-80%
Soil moisture	0.4	0.4	m^3/m^3	0
Boundary Layer	50	2000	m	0
Surface sensible heat flux	-30	200	W m^{-2}	0
Surface latent heat flux	-30	200	W m^{-2}	0
Short-wave radiation	50	800	W m^{-2}	0

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.*

The other meteorological variables are spatially and time varying. We consider the influence of the city located in the center of the domain: the 2m temperature, the 10m wind speed and the 2m 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 short-wave radiation between minimum and maximum values, also presented in Table 2. Between the minimum and maximum value for each meteorological variable, M, a simple sinusoidal expression is used to reproduce diurnal cycles, as:

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

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

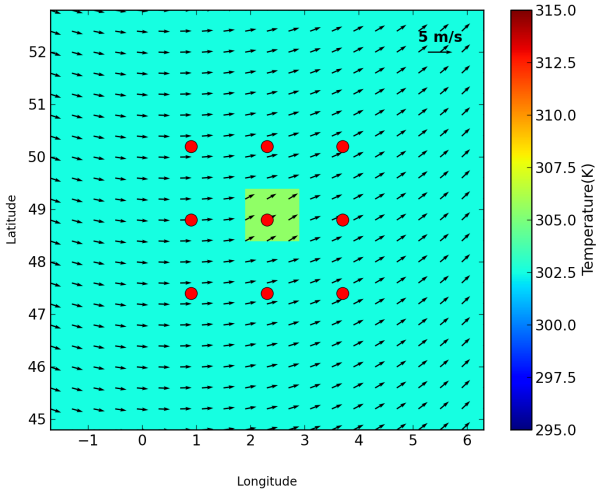

FIG. 2. *Temperature (K) for a summertime period, at the first model level (20m) 12:00 UTC, with a city in the center of the domain. Vectors represent the wind speed (m s^{-1}) in this first model level.*

An example of meteorological fields is presented in Figure 2 with the 2m temperature (colors) and the 10m wind speed (vectors), for a typical summer day at 12:00 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 fully described in (Menut et al. 2013).

Species	[c] (ppb)	Species	[c] (ppb)
O3	30.0	CH4	1700.
NO	0.05	HCHO	0.7
NO2	0.3	C2H6	0.5
HNO3	1.0	NC4H10	0.08
PAN	0.1	C2H4	0.06
H2O2	1.0	C3H6	0.02
CO	80.0	OXYL	0.02

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

In this study, only boundary conditions for gases are considered. Surface emissions of primary particles are considered, but there is no arrival of biomass 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 emissions fluxes calculation as 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 landuses. The values were extracted from the HTAP emissions inventory (Janssens-Maenhout et al. 2015) for the Paris area in June.

4. The preparation of pseudo-observation dataset

The pseudo-observations (called 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 parameters, Table 4. The perturbations values come from usual known uncertainties as already used in CHIMERE in (Menut 2003), among others. OBS is performed for the same period than 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.

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

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

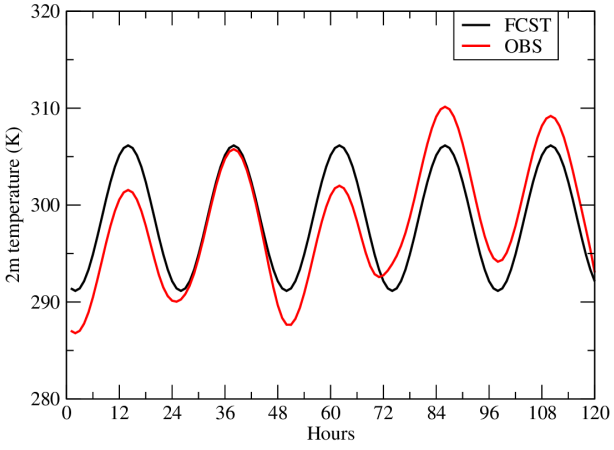

FIG. 3. *Time series of 2m temperature for the "urban" cell. The two model configurations are represented as FCST (current version of the model) and OBS (perturbed version of the model and pseudo-observations).*

For the meteorology, three variables are perturbed: the wind speed (zonal and meridional components), the temperature and the boundary layer height. Figure 3 presents time series of 2m 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 finally ensure mass conservation for the transport calculation.

For the anthropogenic emissions, NO_x and VOCs 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 represent an example of poor knowledge of what are 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 scale, there is a chance to have the same effect at larger scale, then on the chemical concentrations transported in the regional domain, then on these boundary conditions.

a. Results with surface concentrations maps

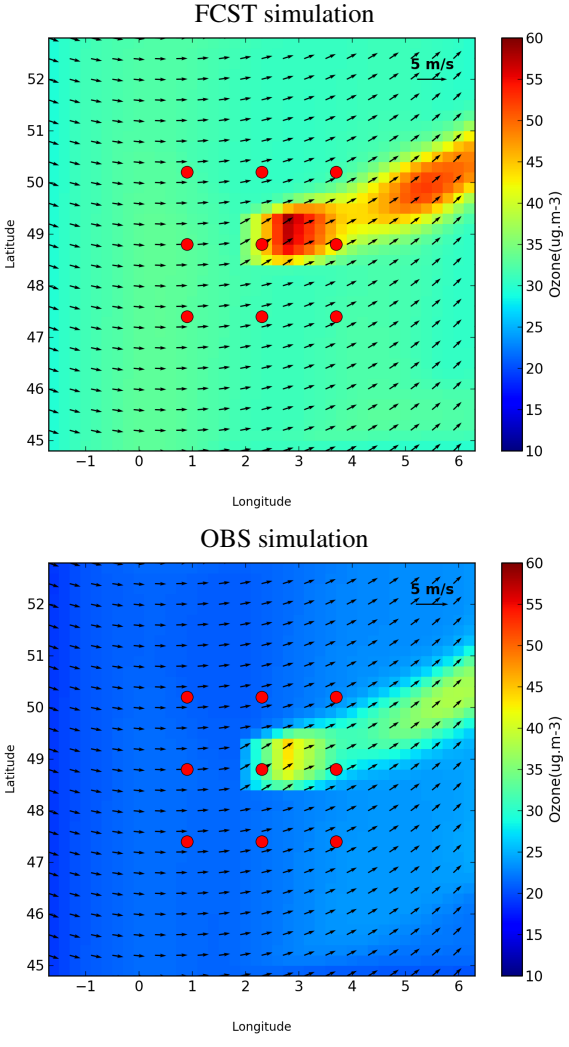

FIG. 4. *Surface ozone concentrations ($\mu g\ m^{-3}$) for the 6th day of the simulation at 12:00 UTC. (top) The FSCT simulation, (bottom) the OBS simulation (corresponding to FCST with perturbations).*

Surface concentrations maps of ozone for the two simulations, FCST and OBS, are presented in Figure 4. The surface ozone concentrations values are presented for the 6 June 2017 at 12:00 UTC, corresponding to the first day of the second simulation period. For the FCST simulation, the maximum values are modelled downwind the city and follow the mean wind flow. Two local maximum are identified: close to the city for ozone just formed in the preecedent hours and in the north-east 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 due to lower surface emissions of NO_x and VOCs and lower boundary conditions of ozone.

5. The pseudo data assimilation

a. The data assimilation cases

The studied cases, DA1, DA2 and DA3, are presented in Figure 5.

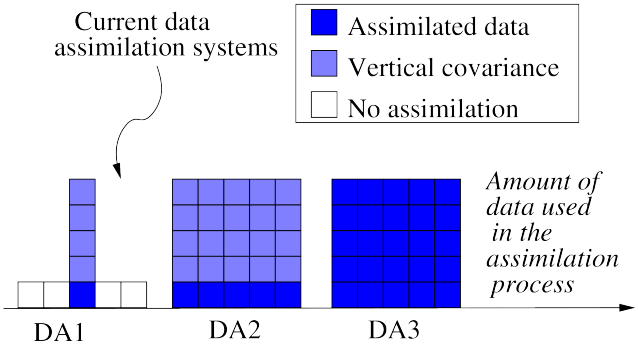

FIG. 5. *Synthetic presentation of the "Data Assimilation" cases. All cases corresponds to idealized set-up and are not designed to be realistics but only to represent minimum and maximum possible benefits of data assimilation. The "real" current systems are between DA1 and DA2.*

DA1: From all defined studied cases, it is the closest to 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. 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 correspond to a low wind speed case with grid cell of 0.2^o width: considering numerical diffusion, there is no need to add horizontal error covariance. It is different in the vertical dimension when the vertical mixing acts quickly and efficiently in the convective boundary layer. A correction is thus applied in the boundary layer, as:

$$c^{FSCT}_{z=2,zABL} = max\left(0., c^{OBS}_{z=2,zABL} + (c^{OBS}_{z=1} - c^{FSCT}_{z=1})\right) \quad (2)$$

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

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 is not existing yet. It is the same principle as DA1 but applied to all model cells. This case is more complete than the current existing systems.

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 that the available data are numerous and the data assimilation system is 'perfect'. This could correspond to future combined in-situ and satellite observations 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 current 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 simulation DA, starting with OBS chemical values, reaches the values of the FCST simulation, a simple criterion s defined as:

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

This calculation represents the difference between the DA simulation and the two cases: FCST and OBS. More the DA simulation is close to FCST, more B is small. We define a threshold value of B_t=0.1 (i.e 10%), corresponding to the fact that the benefit of the restart with OBS was lost with a percentage of 90%. 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 $|c_{DA} - c_{FCST}|$ are larger than 0.1 μg m^{-3}. For

lower values, we consider that the ratio is not significant and there is no gain since the FCST, OBS and DA are already very close.

b. Results with time series

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

Due to 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 criterium. 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 to 7 hours.

During these 3 to 7 hours, there is no convection and photochemistry. For DA1, the concentrations are assimilated at stations 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 due to 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 hours 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 hours. The way how the concentrations changed from OBS to FSCT is different for the two species. For ozone, the shift is sudden and with a few hours duration only: for example, in measURB, the DA3 case shows ozone concentrations close to OBS during the first 26 hours, then reaching the FCST values in one hour only. The same tendency is observed for measNE, when the ozone concentrations of DA3 turns from OBS to FCST calculated concentrations in 4 hours only. 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 hand, NO_2 is directly emitted by local emissions and is rapidly converted during summertime period in the presence of oxidants like ozone. In this case, the local emissions are the key factor to explain the modelled 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 Figure 7 for ozone. For DA1, the benefit is 6 hours at the stations. The benefit is larger for DA2 and may reach 10 hours in the ozone plume. Finally, with DA3, the benefit increases again and may reach 60 hours downwind 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 in 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 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 is mainly due to the fact there is no city (no ozone fast titration by NO_x), but mainly biogenic VOCs, favourable to ozone production. The whole column of ozone being 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 Figure 8. The benefit is more local and mainly downwind the city. In the city, the benefit is due to the replacement of ozone, NO and NO_2 where the main anthropogenic emissions occurs. The effect is thus to counterbalance the bias in emissions by more realistic concentrations. Here, the assimilation has a positive (but short) effect by compensating discrepancies on emissions.

For DA2, the benefit reaches 10 hours. During these hours, from midnight to 10:00 UTC, there is no chemistry and the impact is mainly due to the unperturbed anthropogenic emissions, slowly transported to the northeast. The increase of the number of hours is thus just the reflect of this transport of NO_2. When the photochemistry starts, as well as the vertical mixing, the benefit fades in time. The addition of the vertical covariance slightly increases the number of hours. For DA3, the benefit has a maximum of 15 hours even close to the city. The shape of the maxima over the city is 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 the city where the surface concentrations of NO_2 are low.

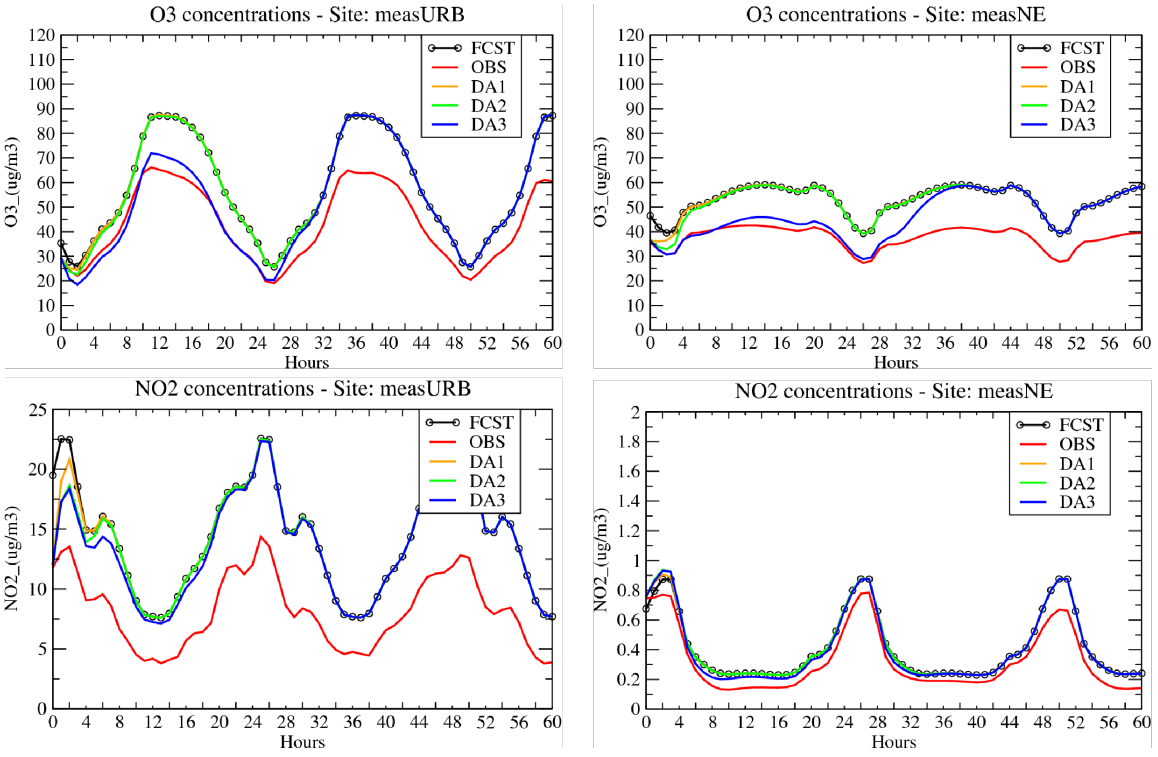

FIG. 6. *Time-series of surface concentrations (in $\mu g\ m^{-3}$) for O_3 (top) and NO_2 (bottom). The simulations for FCST, OBS and the five DA cases are presented on each Figure and for the two sites measURB and measNE.*

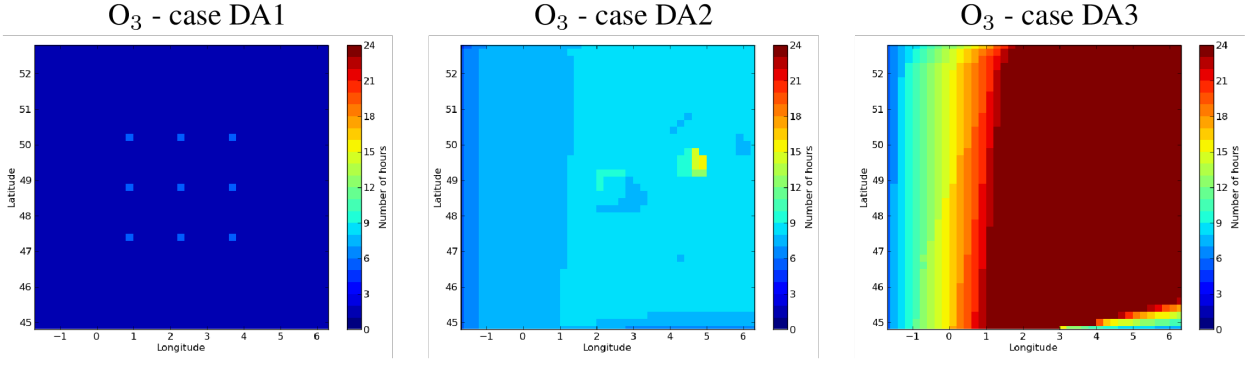

FIG. 7. *Number of hours of data assimilation benefit for ozone. The five data assimilation test cases with different initializations of the simulation are presented.*

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

6. Conclusions

If the current data assimilation systems are able to improve the analysis and the forecast of regional air quality, 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 situa-

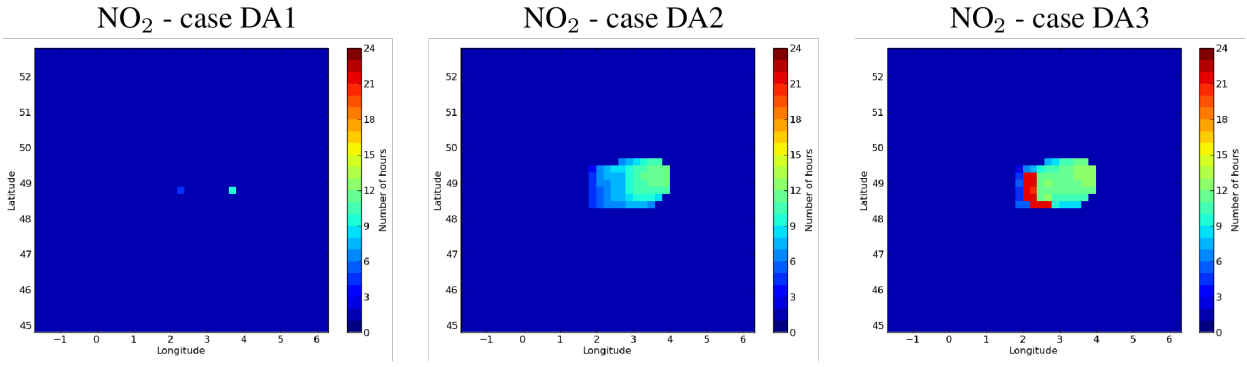

FIG. 8. *Hours of data assimilation benefit for NO_2. The five data assimilation test cases with different initializations of the simulation are presented.*

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

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 modelled domain.*

tion, favourable 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 hours for ozone and NO_2. For DA2, the maxima are 15 hours and 11 hours for ozone and NO_2, respectively. In this case, ozone, a secondary species, is able to be longer transported, then the benefit is higher than the primary species NO_2. This effect increases with the case DA3. This case showed that at the maximum, and considering that we have a lot of data to assimilate, the maximum of benefit would be 60 hours for ozone and 21 hours 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 temperature up to 25^{o}C during the afternoon. A faster wind will dampen the effects of data assimilation. This episode of stagnation thus represents a maximum of possible gain. To strengthen these results, made using academic test cases, the second part of this work, presented in (Bessagnet and Menut 2018), 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., Hein, R., Heinmann, M., and Crutzen, P.: Inverse modeling of the global CO cycle: 1. Inversion of CO mixing ratio, Journal of Geophysical Research, 105, 1909–1927, 2000.

Bessagnet, B. and Menut, L.: What can we expect from data assimilation for air quality forecast? 2: Quantification with a semi-ideal case for a particulate matter episode., Journal of Atmospheric and Oceanic Technology, -, submitted, 2018.

Blond, N. and Vautard, R.: Three-dimensional ozone analyses and their use for short-term ozone forecasts, Journal of Geophysical Research-Atmospheres, 109, doi:10.1029/2004JD004515, 2004.

Bocquet, M., Elbern, H., Eskes, H., Hirtl, M., Žabkar, R., Carmichael, G. R., Flemming, J., Inness, A., Pagowski, M., Pérez Camaño, J. L., Saide, P. E., San Jose, R., Sofiev, M., Vira, J., Baklanov, A., Carnevale, C., Grell, G., and Seigneur, C.: Data assimilation in atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models, Atmospheric Chemistry and Physics, 15, 5325–5358, doi:10.5194/acp-15-5325-2015, 2015.

Chang, M., Hartley, D., Cardelino, C., Haas-Laursen, D., and Chang, W.: On using inverse methods for resolving emissions with large spatial inhomogeneities, Journal of Geophysical Research, 102, 16 023–16 036, 1997.

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

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

Denby, B., Schaap, M., Segers, A., Builtjes, P., and Horlek, J.: Comparison of two data assimilation methods for assessing PM10 exceedances on the European scale, Atmospheric Environment, 42,

7122 – 7134, doi:https://doi.org/10.1016/j.atmosenv.2008.05.058, 2008.

Elbern, H. and Schmidt, H.: A four-dimensional variational chemistry data assimilation scheme for Eulerian chemistry transport modeling, Journal of Geophysical Research, 104, 18 583–18 598, 1999.

Elbern, H., Strunk, A., Schmidt, H., and Talagrand, O.: Emission rate and chemical state estimation by 4-dimensional variational inversion, Atmospheric Chemistry and Physics, 7, 3749–3769, doi:10.5194/acp-7-3749-2007, 2007.

Enting, I.: Inverse problems in atmospheric constituent transport, Cambridge University Press, 2002.

Honoré, C., Rouïl, L., Vautard, R., Beekmann, M., Bessagnet, B., Dufour, A., Elichegaray, C., Flaud, J., Malherbe, L., Meleux, F., Menut, L., Martin, D., Peuch, A., Peuch, V., and Poisson, N.: Predictability of European air quality: The assessment of three years of operational forecasts and analyses by the PREV'AIR system, Journal of Geophysical Research, 113, D04 301, doi:10.1029/2007JD008761, 2008.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Dentener, F., Muntean, M., Pouliot, G., Keating, T., Zhang, Q., Kurokawa, J., Wankmüller, R., Denier van der Gon, H., Kuenen, J. J. P., Klimont, Z., Frost, G., Darras, S., Koffi, B., and Li, M.: HTAP v2.2: a mosaic of regional and global emission grid maps for 2008 and 2010 to study hemispheric transport of air pollution, Atmospheric Chemistry and Physics, 15, 11 411–11 432, doi:10.5194/acp-15-11411-2015, 2015.

Kaminski, T. and Heimann, M.: Inverse modeling of atmospheric carbon dioxide fluxes, Science, 294, 259, 2001.

Konovalov, I., Beekmann, M., Richter, A., and Burrows, J.: Inverse modelling of the spatial distribution of NO_x emissions on a continental scale using satellite data, Atmospheric Chemistry ans Physics Discussions, 5, 12 641–12 695, 2005.

Mahowald, N., Prinn, R., and Rasch, P.: Deducing CCl3 F emissions using an inverse method and chemical transport models with assimilated winds, Journal of Geophysical Research, 102, 28 153–28 168, 1997.

Mailler, S., Menut, L., Khvorostyanov, D., Valari, M., Couvidat, F., Siour, G., Turquety, S., Briant, R., Tuccella, P., Bessagnet, B., Colette, A., Létinois, L., Markakis, K., and Meleux, F.: CHIMERE-2017: from urban to hemispheric chemistry-transport modeling, Geoscientific Model Development, 10, 2397–2423, 2017.

Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., Chéroux, F., Colette, A., Coman, A., Curier, R. L., Denier van der Gon, H. A. C., Drouin, A., Elbern, H., Emili, E., Engelen, R. J., Eskes, H. J., Foret, G., Friese, E., Gauss, M., Giannaros, C., Guth, J., Joly, M., Jaumouillé, E., Josse, B., Kadygrov, N., Kaiser, J. W., Krajsek, K., Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M., Poupkou, A., Queguiner, S., Robertson, L., Rouïl, L., Schaap, M., Segers, A., Sofiev, M., Tarasson, L., Thomas, M., Timmermans, R., Valdebenito, A., van Velthoven, P., van Versendaal, R., Vira, J., and Ung, A.: A regional air quality forecasting system over Europe: the MACC-II daily ensemble production, Geoscientific Model Development, 8, 2777–2813, doi:10.5194/gmd-8-2777-2015, URL `http://www.geosci-model-dev.net/8/2777/2015/`, 2015.

Mendoza-Dominguez, A. and Russell, A.: Estimation of emission adjustments from the application of four-dimensional data assimilation to photochemical air quality modeling, Atmospheric Environment, 35, 2879–2894, 2001.

Menut, L.: Adjoint modelling for atmospheric pollution processes sensitivity at regional scale during the ESQUIF IOP2, Journal of Geophysical Research, 108, 8562, doi:10.1029/2002JD002549, 2003.

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

Menut, L., Bessagnet, B., Khvorostyanov, D., Beekmann, M., Blond, N., Colette, A., Coll, I., Curci, G., Foret, F., Hodzic, A., Mailler, S., Meleux, F., Monge, J., Pison, I., Siour, G., Turquety, S., Valari, M., Vautard, R., and Vivanco, M.: CHIMERE 2013: a model for regional atmospheric composition modelling, Geoscientific Model Development, 6, 981–1028, doi:10.5194/gmd-6-981-2013, 2013.

Müller, J.-F. and Stavrakou, T.: Inversion of CO and NO_x emissions using the adjoint of the IMAGES model, Atmospheric Chemistry and Physics, 5, 1157–1186, 2005.

Pierce, R. B., Schaack, T., Al-Saadi, J. A., Fairlie, T. D., Kittaka, C., Lingenfelser, G., Natarajan, M., Olson, J., Soja, A., Zapotocny, T., Lenzen, A., Stobie, J., Johnson, D., Avery, M. A., Sachse, G. W., Thompson, A., Cohen, R., Dibb, J. E., Crawford, J., Rault, D., Martin, R., Szykman, J., and Fishman, J.: Chemical data assimilation estimates of continental U.S. ozone and nitrogen budgets during the Intercontinental Chemical Transport Experiment - North America, Journal of Geophysical Research: Atmospheres, 112, n/a–n/a, doi:10.1029/2006JD007722, d12S21, 2007.

Pison, I., Menut, L., and N.Blond: Inverse modeling of emissions for local photo-oxidant pollution: Testing a new methodology with kriging constraints, Annales Geophysicae, 24,, 1523–1535, 2006.

Pison, I., L.Menut, and G.Bergametti: Inverse modeling of surface NOx anthropogenic emissions fluxes in the Paris area during the ESQUIF campaign, Journal of Geophysical Research, Atmospheres, 112, D24 302, doi:10.1029/2007JD008871, 2007.

Sandu, A. and Chai, T.: Chemical Data Assimilation—An Overview, Atmosphere, 2, 426–463, doi:10.3390/atmos2030426, 2011.

Talagrand, O.: Assimilation of observations: an introduction, Journal of the Meteorological Society of Japan, 75, 191–209, 1997.

Wang, Y. and Bentley, S.: Development of a spatially explicit inventory of methane emissions from Australia and its verification using atmospheric concentration data, Atmospheric Environment, 36, 4965–4975, 2002.

Wang, Y., McElroy, M., Wang, T., and Palmer, P.: Asian emissions of CO and NO_x: Constraints from aircraft and Chinese station data, Journal of Geophysical Research, 109, doi:10.1029/2004JD005250, 2004.

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

Zheng, H., Liu, J., Tang, X., Wang, Z., Wu, H., Yan, P., and Wang, W.: Improvement of the Real-time PM2.5 Forecast over the Beijing-Tianjin-Hebei Region using an Optimal Interpolation Data Assimilation Method, Aerosol and Air Quality Research, 18, 1305–1316, doi:10.4209/aaqr.2017.11.0522, URL `https://doi.org/10.4209/aaqr.2017.11.0522`, 2018.