_____

IMPROVEMENT OF ATMOSPHERIC DISPERSION SIMULATIONS IN CASE OF AN ACCIDENT OR A MALEVOLENT ACTION USING DATA ASSIMILATION METHODS IN CERES® CBRN-E

Robin Locatelli[1], Vivien Mallet[2] and Patrick Armand[3]

[1]Strathom Energie, Paris, France
[2]INRIA, Paris, France
[3]CEA DAM Ile-De-France, Bruyères-le-Chatel, France

Abstract: Developed at CEA since 2008, CERES® CBRN-E is a computational tool designed for crisis management in case of accidental, malevolent or terrorist releases of hazardous radiological, chemical or biological materials. CERES® CBRN-E computes atmospheric dispersion in complex environments including buildings (industrial sites or urban areas), assesses the health consequences of the toxic releases on the population and first responders, and delivers operational results to rescue teams and decision makers. This paper aims to present a recent development, which could be implemented in CERES® with the final goal to reduce discrepancies between modeled and measured concentrations. This development is based on data assimilation techniques (combination of a state estimation algorithm with an ensemble approach) in order to improve the atmospheric simulation of a dispersion model by incorporating different sources of information (observations, simulation results, error statistics). The results obtained in a synthetic experiment show that this method is very encouraging to derive surface concentrations consistent with observations. Indeed, mean concentration levels and strong spatial and temporal variations are well represented after optimization. Before implementing this approach into CERES® CBRN-E, the next step is to analyze the behavior in real cases.

Key words: data assimilation, ensemble methods, atmospheric dispersion

INTRODUCTION

In the case of an accident or a malevolent action implying a release of gases or fines particles in the atmosphere, dispersion models are used to identify impacted areas with the final aim to protect population. However, discrepancies between modeled and measured concentrations may be large due to uncertainties in the representation of atmospheric transport, ignorance on the characteristics of the source term, errors related to measurement instruments, etc. Consequently, these different sources of uncertainties limit the trust in dispersion and impact assessment models for the decision-makers in an emergency. Thus, there is a strong interest in using all available information together (measurements, simulated concentrations, error statistics) in the course of the crisis in order to strongly reduce the uncertainties. Data assimilation methods are able to combine all these sources of information in order to improve the representation of atmospheric concentrations. Data assimilation is used operationally since the eighties in atmospheric sciences with objectives such as to improve the initial conditions of the operational meteorological forecasts, to identify uncertain parameters by inverse modelling, or to compute a field as close as possible to the ''true'' state of a dynamic system as it can be the case in air quality issues. In this study, we set up data assimilation techniques into the CERES® CBRN-E software, which is a fast-response modeling and decision-support tool, for improving simulations aiming to evaluate as accurately as possible the consequences of toxic dispersions and/or explosions.

Firstly, we detail the theory of data assimilation with a special focus on the algorithm implemented into this system for optimizing concentrations consistent both with observations and outputs from the PMSS model. The different characteristics of this method (definition of the *prior*, the observation operator, observational error covariance matrix, etc.) are detailed. Secondly, we expose the strategy set up in this study to evaluate the improvements and the drawbacks of the implementation of the data assimilation

techniques. Thirdly, we present the main results of a synthetic experiment. Finally, we give some indications for future works in order to apply this method in decision-support systems.

DATA ASSIMILATION METHOD

Data assimilation is a powerful tool which is used in different fields of geosciences (hydrology, seismology, weather forecasting…). Data assimilation refers to the process by which models and measurements are combined to produce an optimal representation of the state of the studied system. In atmospheric sciences, data assimilation techniques could be applied with different objectives: to produce an analysis, in other words to compute a field as close as possible to the "true" state; to improve the initial conditions in order to improve the forecasts; or to identify uncertain parameters, such as the emission fluxes. Here, we use data assimilation to improve atmospheric concentration simulations in order to bring valuable information for decision makers in emergency situations. There are many existing approaches to data assimilation: they include nudging methods, statistical methods, variational methods (e.g., 3-dimensional (3D-Var) and 4-dimensional variational (4D-Var)), and sequential methods (e.g., optimal interpolation and Kalman filters).

Here, we have chosen to compute the so-called best linear unbiased estimator (BLUE) in regard of the static nature of the model and the constraints related to the context of emergency situations. Indeed, the CERES® CBRN-E decision support tool has to deliver operational results (e.g., danger zones, intervention zones...) in less than 15 minutes to rescue teams and decision makers. Computing BLUE has several advantages: it does not require high computational resources and is yet efficient. It is also quite easy to implement. Consequently, this method presents several advantages in agreement with the constraints imposed by the management of crisis situations.

BLUE allows to find some best compromise between a set of observations and *prior* (or background) information on the system state. The observational data set is stored in the $\mathbf{y_o}$ vector, and the state vector is represented by $\mathbf{x}$. The observation operator $\mathbf{H}$ links the observation vector with the state vector:

$$\mathbf{y_o} = \mathbf{Hx^t} + \boldsymbol{\varepsilon^o}. \tag{1}$$

$\boldsymbol{\varepsilon^o}$ is called the "observational error" vector gathering all sources contributing to the gap between the observations and the representation of the "true" state, $\mathbf{x^t}$, in the observation space. In our case, the main contributor to observational errors is the instrumental error. A background (or *prior*) information is also taken into account in this method. It is represented by the vector $\mathbf{x^b}$,

$$\mathbf{x^b} = \mathbf{x^t} + \boldsymbol{\varepsilon^b}. \tag{2}$$

We give a confidence level to this information through the $\boldsymbol{\varepsilon^b}$ vector, which gathers the different errors contributing to a gap between the *prior* state and the "true" state of the system. Then, we define the *prior* errors and the observational errors covariance matrices, $\mathbf{B}$ and $\mathbf{R}$, according to the equations (3). E[] is the expectation operator.

$$\mathbf{B} = \mathrm{E}\left[(\boldsymbol{\varepsilon^b} - \mathrm{E}[\boldsymbol{\varepsilon^b}])(\boldsymbol{\varepsilon^b} - \mathrm{E}[\boldsymbol{\varepsilon^b}])^t\right]; \ \mathbf{R} = \mathrm{E}\left[(\boldsymbol{\varepsilon^o} - \mathrm{E}[\boldsymbol{\varepsilon^o}])(\boldsymbol{\varepsilon^o} - \mathrm{E}[\boldsymbol{\varepsilon^o}])^t\right] \tag{3}$$

We are looking for the best linear unbiased estimator, $\mathbf{x^a}$. It is defined by: $\mathbf{x^a} = \mathbf{x^t} + \boldsymbol{\varepsilon^a}$, where $\boldsymbol{\varepsilon^a}$ is the optimized state error. To find the best linear unbiased estimator, we make few assumptions. We assume that background and observational errors have zero mean: $\mathrm{E}[\boldsymbol{\varepsilon^o}] = 0$ and $\mathrm{E}[\boldsymbol{\varepsilon^b}] = 0$. We also assume they are uncorrelated: $\mathrm{E}[\boldsymbol{\varepsilon^b}(\boldsymbol{\varepsilon^o})^t] = 0$. Then, we are looking for the expression of $\mathbf{x^a}$, whose the error $\boldsymbol{\varepsilon^a}$ has zero mean and has a minimum total variance (i.e., the trace of the matrix $\mathrm{E}[\boldsymbol{\varepsilon^a}(\boldsymbol{\varepsilon^a})^t]$ is minimal). Then, $\mathbf{x^a}$ is given by:

$$\mathbf{x^a = x^b + BH^t(HBH^t + R)^{-1}(y_o - Hx^b)} \qquad (4)$$

MODELING OF THE ERROR COVARIANCE MATRICES

The observational error covariance matrix $\mathbf{R}$ and the *prior* error covariance matrix $\mathbf{B}$ are essential in data assimilation. Indeed, these matrices determine the respective weights given to each piece of information in the analysis. However, the correct specification of those statistics remains a major challenge in data assimilation systems. The structure of background error correlations is particularly important, as it determines how the observed information is filtered and propagated spatially.

The observational error covariance matrix is taken diagonal, hence assuming no correlation between the observational errors at two different monitoring stations. The expression of the observational error covariance matrix is therefore $\mathbf{R} = \nu_o \mathbf{I}$, where ν_o is the observational error variance and $\mathbf{I}$ is the identity matrix. ν_o is mainly determined thanks to our knowledge on instrumental errors.

The definition of the $\mathbf{B}$ matrix is more complex and more crucial for the optimization. Indeed, a good specification of the $\mathbf{B}$ matrix is necessary to take correctly into account the spatial information brought by the model. For data assimilation at large scale, the *prior* error covariance is sometimes parameterized as a function of the geographical distance, e.g., with a decreasing exponential. At regional or local scales, the *prior* error covariances do not only depend on the distance, but also on different other parameters (distance to the source, wind speed, wind direction, turbulence…). Consequently, at regional or urban scales, it can be more complex to diagnose a satisfactory formulation for the $\mathbf{B}$ matrix. We have tested three different analytical formulations based on the works of Balgovind et al. (1983), Frydendall et al. (2009) and some variants of these. The results found with these approaches are not detailed in this paper because they suffer from limitations. Moreover, the analytical formulations of the $\mathbf{B}$ matrix are largely dependent on the configuration of the domain (size of the domain, localization of the source, wind direction…). Consequently, it does not seem reasonable in an emergency situation to take time for finding the best formulations for $\mathbf{B}$. We need to be able to quickly propose a correct configuration for the simulations of the gas dispersion. Consequently, it has been decided to use another way to define $\mathbf{B}$. Indeed, we can define $\mathbf{B}$ by using a Monte Carlo approach: with an ensemble of N "perturbed" simulations, we are able to sample the *prior* error. From N different simulations based on different perturbations in the inputs, the elements of $\mathbf{B}$ as

$$B(i,j) = \frac{1}{N-1} \sum_{k=1}^{N} (x_i^k - \overline{x_i})(x_j^k - \overline{x_j}) \quad ; \quad \overline{x} = \frac{1}{N} \sum_{k=1}^{N} x^k \qquad (6)$$

The N different simulations are created using different perturbations on the meteorological conditions, the characteristics of the gas source and the choice of numerical schemes for dispersion computations.

SYNTHETIC EXPERIMENT

We analyze the capabilities of the system in a study case using synthetic observations. The domain size of the study is 17x35 kilometers with a regular grid-space of 100 meters (171x351 points). A total tracer amount of 1.2×10^5 Becquerel is emitted during 1 hour. Outputs are available every 300 seconds (5 minutes). We describe on the Figure 1 the method used here to evaluate the assimilation results on the simulation of atmospheric concentrations in this specific case.
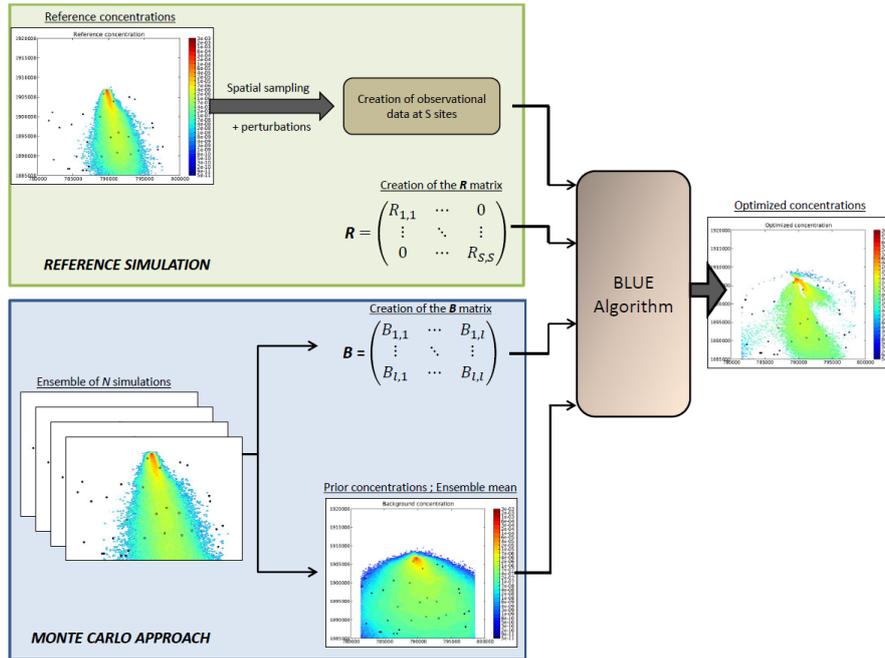
Figure 1. Scheme of our synthetic experiment with the aim to evaluate the data assimilation system.

Firstly (top box on the left part of the Figure 1), a reference simulation is created with the PMSS model (Tinarelli et al., 2013) using base meteorological conditions and base source characteristics. From that simulation, a data set of synthetic observations is created by spatial sampling and perturbing with Gaussian noise with zero mean and variance $\mathbf{R}$. The diagonal $\mathbf{R}$ matrix is also computed knowing the number, the localization of the sites and the observational errors. Secondly (bottom box on the left part of the Figure 1), a Monte Carlo approach is set up to produce N "perturbed" simulations from the reference one. The perturbations are Gaussian and are applied to meteorological parameters (temperature, wind speed and direction) and source characteristics. Perturbations on numerical choices in PMSS are also applied. The *prior* simulation is computed as the mean of the simulation ensemble. The $\mathbf{B}$ matrix is computed according to the equation (6). Finally (right part of the Figure 1), the optimized concentrations are computed as BLUE using the synthetic observations data set, the $\mathbf{R}$ and $\mathbf{B}$ matrices and the *prior* knowledge of the concentrations. Then, the quality of the optimization is evaluated from a direct comparison between the optimized and the reference simulations.

RESULTS

After several tests, it has been decided to use 60 members in the ensemble and synthetic observations coming from 40 different sites to constrain the assimilation system. It has been found that this configuration gave the best tradeoff between the quality of the optimization state and the computing time. Here, we evaluate the spatial distribution of the optimized concentrations by comparing the map of the *prior*, reference and optimized concentrations (see Figure 2). We focus on the final step (the 12th time step) of the simulation.
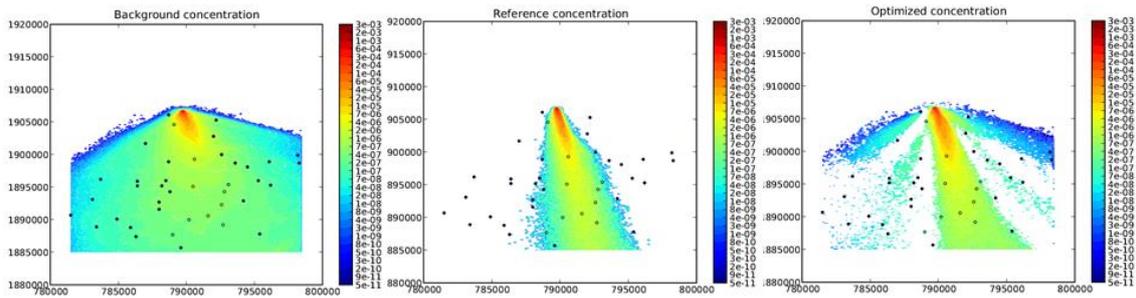
Figure 2. Spatial distribution of the *prior* (left), reference (middle) and optimized (right) concentrations (at the final time step) based on a configuration of the assimilation system using 60 members and 40 observation sites. The points represent the localization of the observation sites.

The *prior* concentration map (left plot of the Figure 2) is computed as the mean of the 60 concentration maps simulated by the 60 members of the ensemble. The aim is to reproduce the reference concentration map (center plot of the Figure 2) after the assimilation process. Here, we show that we are able to reproduce the main characteristics of the plume (shape, direction, size, etc.). Moreover, the magnitude of the optimized concentrations looks very similar to the reference concentrations. However, some low concentrations are located at the edges of the *prior* plume coming from an inadequate specification of the **B** matrix in these areas.
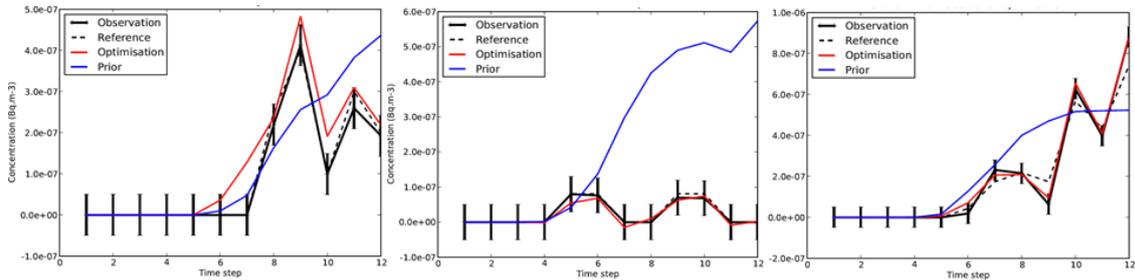

Figure 3. Time series of the *prior*, reference and optimized concentrations at three sites. The observations "measured" at these sites are used in the assimilation process, which explains the close proximity of the optimized concentrations. The observations have been created as a deviation from the reference simulation using a Gaussian distribution. The errors bars represent the observational errors.

On the Figure 3, it is found that the optimized concentrations (in red) fit very well the observations (in black with the errors bars). The system is able to reproduce the fast variations in the concentrations, while the *prior* was only able to reproduce the magnitude of the concentrations (left and right plots). Moreover, when the *prior* concentrations are very far from the observations, the system is also able to strongly move from the *prior* to derive optimized concentrations fitting the observations (center plot). These results are of course explained by the fact that these observations are assimilated.
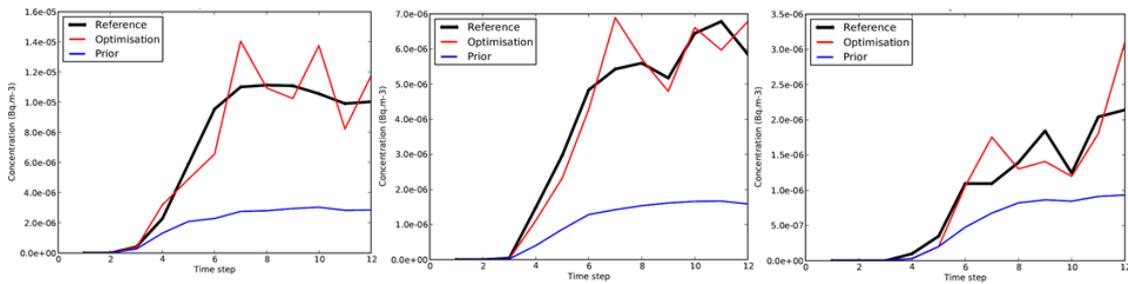
Figure 4. Time series of *prior*, reference and optimized concentrations at 3 different independent stations. *Independent* means here that observations used in the assimilation process are not coming from these independent stations.

Hence, these results only mean that the assimilation system works properly, but they do not give much information on the quality of the assimilation outside areas surrounding the observations. One way to evaluate the quality of the optimization is to compare the time series of optimized and reference concentrations at independent stations (which are not included in the assimilation process). The Figure 4 presents these time series at 3 different independent stations. We show here that the optimized concentrations perform much better than the *prior* concentrations at these stations, even if some variations are too strong in the optimized concentrations. Statistically, the scores confirm that the optimization works well in the areas far from the observations assimilated: the normalized bias, correlation and normalized root mean square error shift before and after optimization from -0.62 to 0.1, from 0.81 to 0.92 and from 1.9 to 0.3 respectively.

CONCLUSION

In this paper, we present a data assimilation system based on the BLUE in order to improve the representation of atmospheric dispersion at regional or local scales in the CERES® CBRN-E system. The definition of the variance/covariance matrix B is a fundamental point of this method: a good specification of this matrix is needed to spread the information given by the observations. Here, we compute the B matrix by using an ensemble of simulations based on different perturbation of the inputs, what gives indications on the covariances of simulation errors.

The analysis of the system in the context of a synthetic experiment shows that the system developed here is able to correctly assimilate the synthetic observations in order to produce an optimized state close to the reference state. Indeed, the statistical scores of the optimized concentrations are much better that those obtained with the *prior* concentrations. Moreover, the temporal variations of the concentrations show that the system is able to reproduce the main characteristics contained in the observations. On-going and future work will apply this method to real cases and experimental cases in wind tunnels with more or less complex situations (considerations of buildings, complex topography, etc.).

REFERENCES

Balgovind, R., A. Dalcher, M. Ghil and E. Kalnay, 1983: A Stochastic-Dynamic Model for the Spatial Structure of Forecast Error Statistics, *Monthly Weather Review*, 111, 701-722.

Frydendall, J., J. Brandt and J.H. Christensen, 2009: Implementation and testing of a simple data assimilation algorithm in the regional air pollution forecast model, DEOM. *Atmospheric Chemistry and Physics*, 9, 5475-5488.

Tinarelli, B., S. Trini-Castelli, G. Carlino, J. Moussafir, C. Olry, P. Armand and D. Anfossi, 2013: Review and validation of Micro-SPRAY, a Lagrangian particle model of turbulent dispersion. *In Lagrangian Modeling of the Atmosphere*, Volume 200, AGU, doi: 10.1029/2012GM001242.