GEOSTATISTICAL ANALYSIS OF VALIDATION DATA OF AN AIR POLLUTION SIMULATOR

JEAN-PAUL CHILES [1], SERGE SEGURET [1] and PIERRE-MARC RIBOUD [2]

[1] MINES ParisTech, Fontainebleau, France
[2] EDF R&D, Chatou, France

ABSTRACT

Chemistry-transport models for air quality forecasting are affected by the uncertainty on the input data (emissions of pollutants and meteorological conditions), the approximations in the modelling of the physicochemical reactions, and numerical approximations (space and time discretization). The validation of the accuracy of these simulators can be done by comparing predictions with actual measurements. This exercise has been carried out for a model for daily forecasting at the scale of Europe, with reference to daily measurements at about one hundred stations over one year. A thorough variographic analysis shows that the error field cannot be characterized independently of the predicted and observed fields. Indeed the forecasts usually display space and time variations similar to those of the measurement data, up to a multiplicative factor, but are often poorly correlated with the reality. These results can be used to define priorities in the improvement of the chemistry-transport model. The presentation is focused on sulphates and nitrogen dioxide.

INTRODUCTION

An increasing number of phenomena of Earth sciences can now be modelled by process simulators. The numerical models they provide are an approximation to the reality, due to simplification assumptions made by the simulator, numerical approximations, and imperfect knowledge of the boundary and initial conditions. Several methods have been proposed to improve the quality of the model simulations once some observations are available. In the 60's, meteorologists developed the guess-field approach, which consists in using the numerical weather forecast model for time *t* as a first guess of reality, analysing the forecast errors deduced from observations at time *t*, and adding an interpolation of that error to the forecast grid (Cressman, 1959; Rutherford, 1972). Later on, the external-drift approach fulfilled a similar objective when the field produced by the physical simulator, or some secondary variable, is only linearly linked with the studied variable (Delfiner et al., 1983). More recently, data assimilation provided a significant improvement as it enables the correction of both the output of the simulator and the statistical model parameters; see, e.g., Talagrand and Courtier (1987) and Talagrand (1997) for the variational approach, and Burgers et al. (1998), Bertino et al. (2003), and Evensen (2007) for sequential data assimilation.

Even if corrections are valuable, they are effective under the conditions that the chemistry-transport model has a certain degree of quality and the discrepancy between the simulator output and the reality has been clearly understood, at least from a statistical perspective. There is thus a need for a validation of the approach as well as for improvements in the simulator. It could be tempting, for simplicity, to only study the forecast error (difference between measurement and simulator output). But to understand the forecast error, it is usually necessary to analyse it in relation to the real input or output field. Pebesma et al. (2005) give an example of such an analysis, in the very different context of rainfall-runoff event time series. In our application—air quality forecasting at the scale of Western Europe—the input fields are complex (meteorology) or imprecise (emissions). Therefore, we jointly analyse the spatial variations of the real output field (pollution), the predicted field, and the error field.

THE SIMULATOR AND THE VALIDATION DATA

The Chemistry-Transport Model

The CHIMERE chemistry-transport model is a multi-scale model for air quality forecasting and simulation. It is primarily designed to produce daily forecasts of ozone, aerosols and other pollutants and make long-term simulations (entire seasons or years) for emission control scenarios. CHIMERE runs over a range of spatial scales from the regional scale (several thousand kilometres) to the urban scale (100-200 km) with resolutions from 1-2 km to 100 km. CHIMERE proposes many different options for simulations which make it also a powerful research tool for testing parameterizations and hypotheses. For further information, see the official site: http://euler.lmd.polytechnique.fr/chimere/.

The present study concerns the validation of a model based on CHIMERE and whose chemical part has been developed by the LISA (Laboratory of atmospheric systems of Paris 7 and Paris 12 Universities and CNRS, the French national centre for scientific research). This simulator models the transport and the chemical evolution (up to several hundred reactions) of a series of several tens of constituents at the scale of Europe, depending on the emissions and the meteorological conditions. The main sources of uncertainty are the imperfect knowledge about the emissions, the uncertainty of the meteorological model, and possibly approximations in the modelling of the physicochemical reactions as well as numerical approximations (geographical, vertical, and time discretization).

The Validation Data

The simulator models the air pollution at a regional scale, the grid has 68×48 nodes, its spatial discretization is about 50 km (Fig. 1). The validation exercise was carried out with the daily average concentration data of the EMEP regional monitoring sites across Europe (EMEP is a program under the Convention on long-range transboundary air pollution for international co-operation to solve transboundary air pollution problems; see http://www.emep.int/). One complete year has been considered, namely the 366 situations of year 2000. A total of 91 stations spread over 24 countries were available, but the set of measured constituents differed from one country to the other. Ozone data originate from another network with 73 stations, and are averages over one hour. Five species were considered: nitrogen dioxide (NO_2), nitrates, sulphates, ozone, and the particulate matter of less than 10μm. We focus here on sulphates and NO_2.

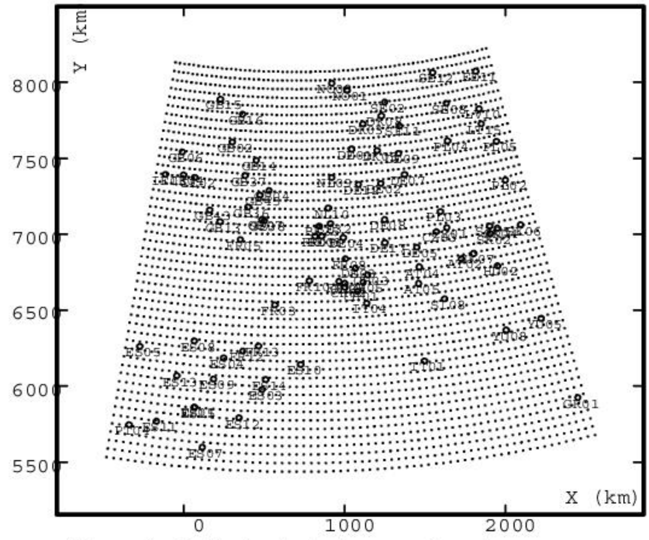


Figure 1: Simulated 50×50 km^2 grid and the 91 point stations giving daily measurements.

Gilles Forêt from LISA collected the EMEP data. He also ran the CHIMERE simulator for these 366 situations and stored the results relative to the lowest atmospheric layer, which is about 100-m thick. The support of these data is thus 50×50 km^2 horizontally and 100 m vertically. The simulated values at the EMEP stations were obtained by interpolation of the simulated grids.

A SIMPLE REFERENCE MODEL

Standard statistical tools (mean, variance, correlation coefficient, histogram, scatter diagram) give useful global information about the relationship between the three variables. We will see that the simulation gives a biased view of the pollution at the stations, sometimes overestimating it (NO_2), sometimes underestimating it (nitrates). This bias does not amount to a simple shift of the values but produces a deformation of the histogram. It comes with a bias of the standard deviation similar to that of the mean, though less pronounced.

The variograms allow us to examine the space and time structure of the measured and simulated data, as well as of their difference, and particularly to examine if the morphology of the simulated fields is similar to what the measurement data suggest.

Since our objective is not to correct the numerical model, the variograms will be used in a descriptive way and do not need to be fitted to a model. To facilitate their interpretation it is nevertheless useful to have in mind the relations that exist between the various variograms in the framework of a simple model.

We could imagine a priori that the simulation captures one part of the reality and that the other part, the residual, is not correlated to it. In such a model, the variogram of the reality is the sum of the variograms of the simulation and of the residual. Even by including measurement errors, that model did not hold to the first sample variogram calculations. The main reason is that the simulation usually displays less or more (depending on the constituent) variability than the reality. This lead us to consider a more general, and yet very simple, model.

The Model

Pollution $Y(\mathbf{x})$, considered as a function of point $\mathbf{x}$, behaves homogeneously in the study domain: it is represented by a stationary or intrinsic random function with variogram $\gamma(\mathbf{h})$.

The measurement $M(\mathbf{x})$ at point $\mathbf{x}$ is the sum of pollution $Y(\mathbf{x})$ and a measurement error $\varepsilon(\mathbf{x})$

$$M(\mathbf{x}) = Y(\mathbf{x}) + \varepsilon(\mathbf{x})$$

The measurement errors are supposed to be non-systematic, uncorrelated with each other and with $Y(.)$, and with the same variance σ_ε^2 for all measurements.

The simulated field $S(\mathbf{x})$ has the same structure as the reality, up to a multiplicative factor. It is in intrinsic correlation with $Y(.)$, which means that its variogram as well as the cross-variogram of $Y(.)$ and $S(.)$ are proportional to $\gamma(\mathbf{h})$.

Let r^2 denote the ratio between the variogram of $S(.)$ and that of $Y(.)$ (which is equal to the ratio of the variances of Y and S), and ρ denote the correlation coefficient between these variables. This model then corresponds to the case where the simulation is of the form

$$S(\mathbf{x}) = r\,\rho\,Y(\mathbf{x}) + r\sqrt{1-\rho^2}\,X(\mathbf{x}) \qquad (1)$$

where $X(\mathbf{x})$ is a random function with the same spatial structure as $Y(.)$ but independent of it.

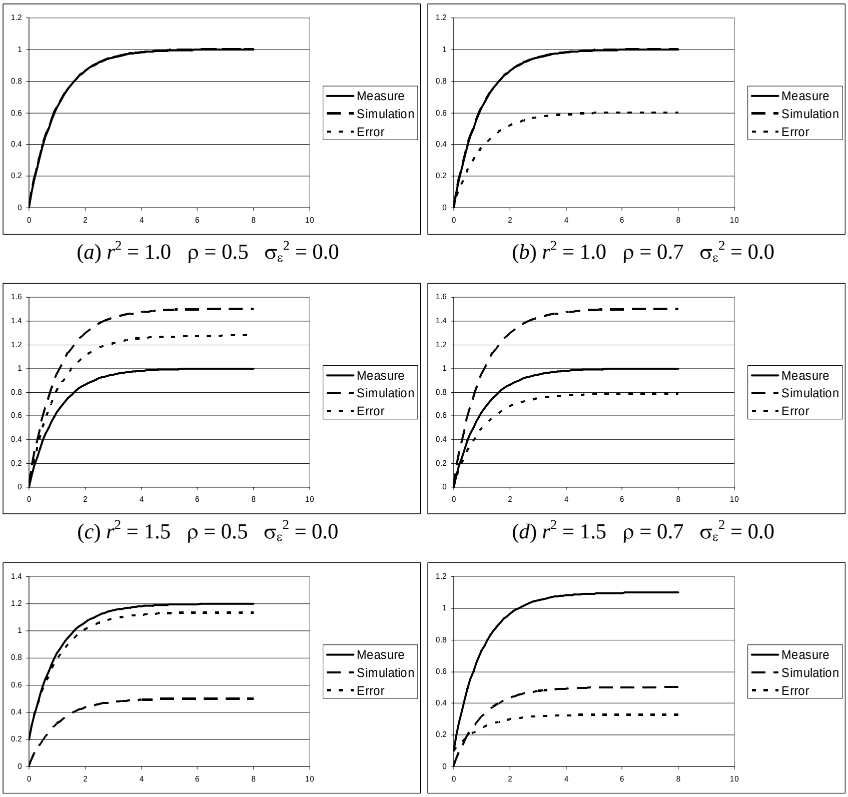
Variograms

The variograms of the measured pollution $Z(\mathbf{x})$, of the simulated pollution $S(\mathbf{x})$, and of the error $E(\mathbf{x}) = Z(\mathbf{x}) - S(\mathbf{x})$ are then:

$$\begin{aligned}
\gamma_M(\mathbf{h}) &= \sigma_\varepsilon^2 + \gamma(\mathbf{h}) \\
\gamma_S(\mathbf{h}) &= r^2\,\gamma(\mathbf{h}) \\
\gamma_E(\mathbf{h}) &= \sigma_\varepsilon^2 + \left((1 - r\rho)^2 + r^2(1-\rho^2)\right)\gamma(\mathbf{h})
\end{aligned}$$

Apart from the measurement error terms, they are all proportional. Figure 2 shows some possible behaviours. The exact shape of the variogram is of little importance at that level, so that we have assumed an exponential variogram with unit scale parameter and unit sill, namely

$$\gamma(h) = \exp(-h) \qquad (h \geq 0)$$



(a) $r^2 = 1.0 \quad \rho = 0.5 \quad \sigma_\varepsilon^2 = 0.0$

(b) $r^2 = 1.0 \quad \rho = 0.7 \quad \sigma_\varepsilon^2 = 0.0$

(c) $r^2 = 1.5 \quad \rho = 0.5 \quad \sigma_\varepsilon^2 = 0.0$

(d) $r^2 = 1.5 \quad \rho = 0.7 \quad \sigma_\varepsilon^2 = 0.0$

(e) $r^2 = 0.5 \quad \rho = 0.4 \quad \sigma_\varepsilon^2 = 0.2$

(f) $r^2 = 0.5 \quad \rho = 0.9 \quad \sigma_\varepsilon^2 = 0.1$

Figure 2: Some configurations of variograms, depending on the ratio r, the correlation ρ, and the measurement error variance σ_ε^2.

In Figure 2*a* and *b* the simulation displays the same variability as reality and the measurement errors are supposed to be negligible. In *a*, where the correlation coefficient between the simulation and the reality equals 0.5, the error has the same variogram as the simulation and the reality. If the correlation is lower (resp. higher), the variogram of the error comes above (resp. below) the others (case *b* corresponds to $\rho = 0.7$).

In Figure 2*c* and *d*, the simulation presents more variability than the reality. The variogram of the error lies above the variogram of the simulation if the correlation between simulation and reality is low (case not represented), between the variogram of the reality and that of the simulation in the case of a medium correlation (*c*), and below the variogram of the reality in the case of a high correlation (*d*).

In Figures 2*e* and *f* the simulation displays less variability than the reality, and measurement errors are introduced. Notice that the variogram of the error can cross that of the simulation in case of a high correlation.

Cross-variograms are usually an essential complement to the direct variograms. They give much more information than a simple correlation coefficient, because they show how correlation evolves with distance (in space or time). This is less true here: since the error is the difference between measurement and simulation, cross-variograms are linear functions of the direct variograms. They will not be shown here, even if they give a more immediate view of the joint structures of the three variables.

In the presence of measurement errors, let us also notice that the variographic analysis provides the true correlation ρ between simulation and reality. Indeed, the measurement errors give a degraded image of that correlation. In situation *f*, for example, where ρ has been fixed to 0.9, the correlation between simulation and measurements, which is the only one to be directly accessible in an actual application, is only 0.85.

RESULTS

We focus on nitrogen dioxide and sulphates. We recall that all the data considered are located at the EMEP stations to make the variograms comparable (the simulation data at these stations were obtained by interpolation of the simulation grid). Variograms concern daily measurements or seasonal and annual ones. They are calculated in the time domain and in space. We present some typical examples and link them to the above reference model.

Sulphates

The 64 stations form a relatively good coverage of Europe with the exception of its south-east part (a single station). The measurements are given in $\mu g/m^3$. The measurement data as well as the simulation data display a lognormal-type histogram, with a lower mean (-25%) and standard deviation (-15%) for the simulation data in comparison to the measurement data.

The average of the spatial variograms calculated daily (Fig. 3) show the presence of a nugget effect in the measurement and error variograms but not in the simulation (this is consistent with intuition—the simulation includes no measurement error—but the simulated data of some constituents display an apparent nugget effect). The general behaviour of the three curves is of the same exponential type and if we shift the simulation variogram by the nugget value, it is in the range of the other variograms. The error and measurement variograms are quite similar, and we can refer to the situation of Figure 2*e* with here a measurement error variance σ_ε^2 equal to about 1.5, a ratio *r* close to 1, and a correlation coefficient ρ of 0.5.

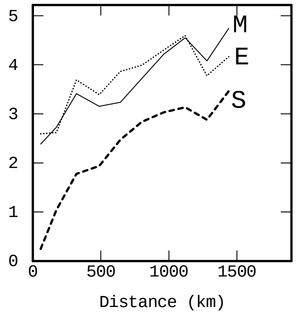


Figure 3: Sulphates. Spatial variograms calculated with the daily measurements and simulated values. S: simulated values, M: measurements, E: error.

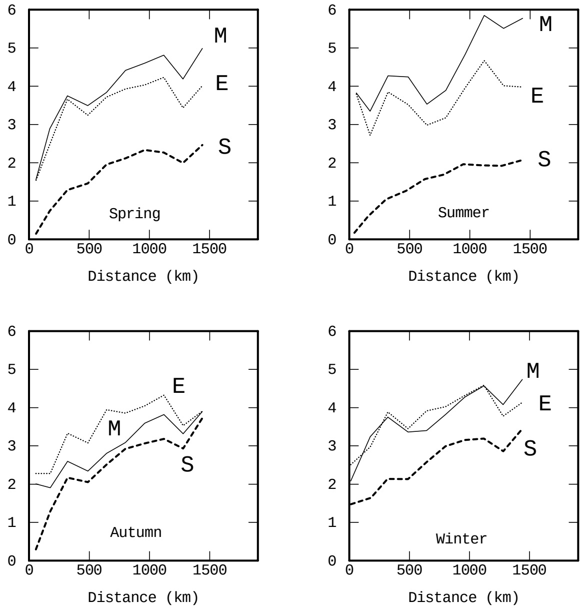


Figure 4: Sulphates. Spatial variograms calculated season by season. S: simulated values, M: measurements, E: error.

In other words, the simulator reproduces the shape of the spatial variations of the daily concentrations correctly (*r* = 1) and without measurement error component, but the simulated field is poorly correlated (ρ = 0.5) to the actual one. Therefore, the error data have the same amplitude as the measurement data.

We considered until now the average spatial behaviour along the year, but that behaviour can vary with the season. Figure 4 still presents variograms in space, but they are now seasonal averages of daily variograms. In spring the situation is similar to that of the yearly average. Summer is characterized by the largest measurement error variance. Autumn is more difficult to predict than the other seasons: the variogram of the error is higher than that of the measurement data, which shows that the simulator is not a good local predictor of reality in that case. In winter we notice that the simulation includes an apparent nugget effect. The ratio *r* seems to vary around 1, with the exception of autumn where it is larger than 1 while the correlation ρ is poor (about 0.3); this is why the error is so large in that season.

The variograms calculated along the time axis all reach a sill at a time lag of about seven days (Fig. 5; for larger time lags they continue to grow, but very slowly). The variograms show the same hierarchy as the daily variograms in the spatial domain. In particular, they confirm that the shape of the time variations is correctly simulated though they seem to be slighly underestimated, and that the average correlation between the simulated data and the actual ones is about 0.5.

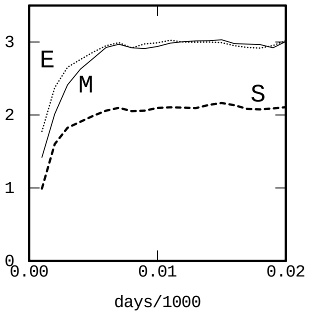


Figure 5: Sulphates. Time variograms calculated with the daily measurements and simulated values. S: simulated values, M: measurements, E: error.

We notice that the nugget effect of the measurement data is smaller in the time domain (about 0.75) than in the spatial domain (1.5). This originates probably in that some part of the so-called "measurement error" is in fact a site effect attached to each measurement station, which is constant through time in a first approximation. This is confirmed by the variograms of yearly averages. The interest of these variograms, in comparison to yearly averages of daily variograms, is that the measurement errors that are not correlated in time will vanish, whereas those corresponding to a site effect and constant through time will remain. The variogram of yearly averages of measurement data (not shown here) has a nugget effect of about 0.75. Therefore, the nugget effect of the yearly average of daily variograms comprises 50% of site effect (constant through time) and 50% of conventional measurement error (uncorrelated in space and time).

Nitrogen Dioxide

The data (47 stations; unit: µg/m^3) are mainly representative of the northern and central part of Europe. The simulation data have an average 40% larger than that of the measurement data. The time variograms of nitrogen dioxide all behave similarly (no noticeable nugget effect, a time range of 4 days) but, contrarily to sulphates, the variability of the simulation is more important than that of the measurement data (Fig. 6). The time variations are therefore very well simulated but they are overestimated and poorly correlated to the reality: referring to formula (1), we can reproduce this behaviour when we set $r = 1.35$ and $\rho = 0.55$. The spatial variograms (not shown here) confirm the larger variability of the simulation, especially in summer. The variograms are roughly linear up to 1000 km and all present an apparent nugget effect (even the simulation) which can be partly due to a structure with a range shorter than 100 km.

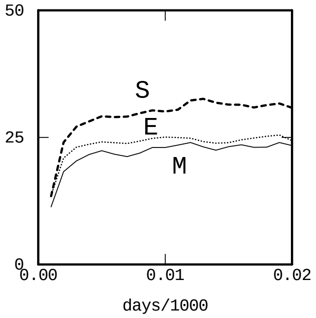


Figure 6: Nitrogen dioxide. Time variograms calculated with the daily measurements and simulated values. S: simulated values, M: measurements, E: error.

DISCUSSION AND CONCLUSION

The whole study confirms that the simulator correctly reproduces the type of variations in time and space that are seen in the measurement data. This gives credit to the simulator, even if these variations are often either overestimated (nitrogen dioxide) or underestimated (sulphates, nitrates).

Unfortunately, even after filtering of the measurement error, the simulations are poorly correlated with the reality: typical values for ρ are 0.5 for sulphates and 0.65 for nitrates. Consequently, the texture of the simulations is similar to that of reality but the details that can be seen on the simulations are not necessarily at the correct location. This can be due to the large uncertainty in the initial conditions (the emissions are known very approximately and only at the scale of one year), to errors in the meteorological model (particularly concerning the occurrence of rainfall), and to defects of the chemistry-transport simulator. The detailed analysis of the results points some ways of improvement for the simulator.

Validating a physical simulator is based on the comparison of simulator forecasts with measurement data. The present study shows a typical example where the error data can be understood statistically when they are considered with reference to the simulation and measurement data. The analysis, based on a variography in the space and time domains, at the scale of daily values as well as seasonal averages, shows a large variety of behaviours which can be explained, at least in a first approximation, by a very simple model. It would be worth to complete it by an analysis in connection with the input data, for example with the type of meteorological conditions by separately studying anticyclonic situations and situations with large perturbations. The present analysis can nevertheless already help to detect directions of research for improving the simulator. It could also be used to correct the simulations, even if this was not the aim of the study.

ACKNOWLEDGEMENTS

The authors would like to acknowledge EDF R&D for supporting this research and Gilles Bergametti, Isabelle Coll and Gilles Forêt from the LISA of Paris 7 and Paris 12 Universities for delivering the data and very fruitful discussions.

REFERENCES

Bertino, L., Evensen, G. and Wackernagel, H. (2003) *Sequential Data Assimilation Techniques in Oceanography*, International Statistical Review, vol. 38, no. 2, pp. 223-241.

Burgers, G., van Leeuwen, P.J. and Evensen, G. (1998) *Analysis Scheme in the Ensemble Kalman Filter*, Monthly Weather Review, vol. 126, no. 6, pp. 1719-1724.

Cressman, G.P. (1959) *An Operational Objective Analysis System*, Monthly Weather Review, vol. 87, no. 10, pp. 367-374.

Delfiner, P., Delhomme, J.P. and Pélissier-Combescure, J. (1983) *Application of Geostatistical Analysis to the Evaluation of Petroleum Reservoirs with Well Logs*, in Proceedings of the 24th Annual Logging Symposium of SPWLA, Calgary, June 1983.

Evensen, G. (2007) *Data Assimilation, The Ensemble Kalman Filter*, Springer, Berlin.

Pebesma, E.J., Switzer, P. and Loague, K. (2005) *Error Analysis for the Evaluation of Model Performance: Rainfall-Runoff Event Time Series Data*, Hydrological Processes, vol. 19, no 8, pp. 1529-1548.

Rutherford, I.D. (1972) *Data Assimilation by Statistical Interpolation of Forecast Error Fields*, Journal of Atmospheric Sciences, vol. 29, no. 5, pp. 809-815.

Talagrand, O. (1997) *Assimilation of Observations, an Introduction*, Journal of the Meteorological Society of Japan, vol. 75, no. 1B, pp. 191-209.

Talagrand, O. and Courtier, P. (1987) *Variational Assimilation of Meteorological Observations with the Adjoint Vorticity Equation. I: Theory*, Quarterly Journal of the Royal Meteorological Society, vol. 113, no. 478, pp. 1311-1328.