GEOSTATISTICS AND SEQUENTIAL DATA ASSIMILATION

HANS WACKERNAGEL[1] and LAURENT BERTINO[2]
[1] *Centre de Géostatistique, Ecole des Mines de Paris, France*
[2] *Nansen Environmental and Remote Sensing Center, Norway*

Abstract. We review possibilities of introducing geostatistical concepts into the sequential assimilation of data into numerical models. The reduced rank square root filter and the ensemble Kalman filter are presented from this perspective. Contributions of geostatistics are discussed showing that sequential data assimilation is a promising area for the application of geostatistical techniques.

1 Introduction

Traditional geostatistical space-time geostatistics (Kyriakidis and Journel, 1999) is not able to take account of the generally strongly non-linear dynamics of multivariate space-time processes. To this effect physico-chemical transport models are in general more suitable. However, as the latter do not fully master the complexity of the processes they attempt to describe, either because of simplifying hypotheses or because the information serving to set up initial and boundary conditions is imperfect, it is appropriate to introduce statistical techniques in order to assimilate a flow of measurements emanating from automatic stations.

Recent projects at Centre de Géostatistique have permitted to explore these techniques in oceanography and air pollution. Soon it became evident that geostatistics could offer concepts and approaches to enhance Sequential Data Assimilation techniques. The thesis of Laurent Bertino (Bertino, 2001) and subsequent publications (Bertino et al., 2002; Bertino et al., 2003) have permitted to develop this theme.

More precisely, when dealing with Sequential Data Assimilation (as opposed to variational techniques) two viewpoints can be adopted. On one hand, from the point of view of the designer of deterministic numerical models, data assimilation is seen as an algorithm permitting to correct the state of the mechanistic model as new data comes in. On the other hand, from the point of view of the statistician the numerical model can help improve the operational prediction taking advantage of the knowledge of the non-linear relations between the various data sources.

It is this second viewpoint that we will privilege and seek to develop by positioning Geostatistics at the center of all data flows coming from a network of stations, a transport model coupled with other data sources or remote sensing data. Geostatis-

O. Leuangthong and C. V. Deutsch (eds.), Geostatistics Banff 2004, 893–898.

tics, with its multivariate models, its anamorphosis and change-of-support models as well as its conditional simulation methods, offers a unique integrative framework for connecting these different informations, for understanding and modeling their statistical structure, for setting up prediction algorithms.

Data assimilation algorithms are needed for setting up operational forecasting systems as they are used in meteorology, oceanography, hydrology, ecology, epidemiology. An *operational forecasting system* consists of:

- a network of automatic stations,
- a dynamic forecasting model,
- a *data assimilation* algorithm.

As the station data generally provide only bad spatial coverage, the numerical model can compensate for this by a forecast based on known physical, chemical or biological relations. The data assimilation algorithm is then essential in combining these two sources of information sequentially in time, taking into account observational and model error.

This paper is divided into three sections. Section 2 describes a particular version of the extended Kalman filter, insisting on the geostatistical aspects. Section 3 presents another suboptimal Kalman filter which is close in spirit to the geostatistical simulation of Gaussian processes. Section 4 reviews a few possibilities of introducing geostatistical ideas into sequential data assimilation.

2 Kalman filter

We present the Kalman filter in its so-called *reduced rank square-root* (RRSQRT) version (Verlaan and Heemink, 1997) using notation that is close to the one used in geostatistics. Let $\mathbf{z}_t^o$ be the vector of the n observations at time t, $\mathbf{y}_t$ be the state of the system, we denote the state forecast with a f and the corrected state with a star, i.e.: $\mathbf{y}_t^f, \mathbf{y}_t^\star$. The forecast is performed by a numerical model $\mathcal{M}$, with boundary conditions $\mathbf{u}_t$, which describes the usually non-linear time dynamics. For computing error covariances we need to derive from the numerical model the tangent linear operator $\mathbf{M}$. We also need an observation linear operator $\mathbf{H}$ which serves both to transfer information from grid points to station locations and to generate from the forecast state "observations" as anticipated by the numerical model.

Leaving aside a detailed state space presentation of the Kalman filter, we merely present the algorithm which is composed of two steps. The first step is a propagation of the state from time $t-1$ to time t, using the numerical model to do the forecast:

$$\mathbf{y}_t^f \;=\; \mathcal{M}_t\left(\mathbf{y}_{t-1}^\star, \mathbf{u}_t\right) \tag{1}$$

and using the tangent linear operator to compute the corresponding error covariances,

$$\mathbf{C}_t^f \;=\; \mathbf{M}_t \mathbf{C}_{t-1}^\star \mathbf{M}_t^\top + \mathbf{Q}_t \tag{2}$$

The model noise covariance matrix $\mathbf{Q}_t$ needs to be carefully calibrated as it will condition the behavior of the filter. In a study of the hydrodynamics of the Baltic sea the sensitivity of the system stemmed mainly from the errors in the boundary conditions. Under the assumption that the water level field at the open boundary can be described by the wave equation a geostatistical model in the form of a space-time covariance model could be proposed (Wolf et al., 2001).

The second step is a correction of the state by kriging, perfomed at time t as soon as new data comes in. Kriging weights are computed from the forecast error covariances as well as the observation error covariances,

$$\mathbf{W}_t \;=\; \mathbf{C}_t^f \mathbf{H}^\top \left(\mathbf{H} \mathbf{C}_t^f \mathbf{H}^\top + \mathbf{C}^o \right)^{-1} \tag{3}$$

The corrected state is obtained by simple kriging

$$\mathbf{y}_t^\star \;=\; \mathbf{y}_t^f + \mathbf{W}_t \left(\mathbf{z}_t^o - \mathbf{H} \mathbf{y}_t^f \right) \tag{4}$$

and corresponding error covariances are computed,

$$\mathbf{C}_t^\star \;=\; \left(\mathbf{I} - \mathbf{W}_t \mathbf{H} \right) \mathbf{C}_t^f \tag{5}$$

In the RRSQRT algorithm the most important eigenvectors ("square roots") of the error covariance matrices $\mathbf{C}^f$, $\mathbf{C}^o$ are propagated ensuring both the positive definiteness of the matrices and a drastic reduction of dimensionality.

3 Ensemble Kalman filter

The *Ensemble Kalman* filter (EnKF) due to Geir Evensen (Evensen, 1994; Burgers et al., 1998) is based on a Monte-Carlo framework and has the avantage of not requiring a linearization of the numerical forecasting model $\mathcal{M}$. At each time step an ensemble of N forecasts

$$\left\{ \mathbf{y}_t^{f,i} = \mathcal{M}_{t-1}(\mathbf{y}_{t-1}^{\star,i}, \mathbf{q}_t^i);\; i = 1 \ldots N \right\}, \tag{6}$$

are propagated using simulated model errors $\{\mathbf{q}_0^i\}$. In geostatistical terms this first step can be seen as a *non-conditional simulation* generating N realisations of a non-stationary random function. The average forecast $\mathbf{y}_t^f$ and the covariance matrix $\mathbf{C}_t^f$ are computed directly on this ensemble of realizations.

The second step is the *conditioning* of the realizations by kriging on the basis of n observations collected at time t,

$$\left\{ \mathbf{y}_t^{\star,i} = \mathbf{y}_t^{f,i} + \mathbf{W}_t(\mathbf{z}_t - \mathcal{H}\mathbf{y}_t^{f,i} + \mathbf{u}_t^{o,i});\; i = 1 \ldots N \right\}, \tag{7}$$

where the observation errors are simulated according to a normal distribution $\mathcal{N}(0, \mathbf{C}^o)$ and the observation operator $\mathcal{H}$ is allowed to be non-linear. The first two moments of this ensemble of realizations approximate $\mathbf{y}_t^\star$ and $\mathbf{C}_t^\star$ in the same way as the mean of a number of conditional geostatistical simulations is equivalent to the

solution of the kriging of the data. The details of the algorithmic implementation of the EnKF are discussed in (Evensen, 2003).

4 Contributions of geostatistics

We have seen that the correction step of the Kalman filter implies a kriging and that the EnKF is similar in spirit to the conditional simulations used in geostatistics. We also mentioned that geostatistics can be used to model the spatial correlation of the model error (Cañizares, 1999; Sénégas et al., 2001; Wolf et al., 2001).

UNIVERSALITY CONDITIONS

The correction step of the Kalman filter implies a simple kriging of the differences between the observations and their forecast according to the numerical model. It is possible to add universality conditions to this kriging in order to remove multiplicative or additive bias (Bertino, 2001). The approach is then equivalent to the one solved by external drift using numerical model output as external drift (Wackernagel et al., 2004). However, the difference is that geostatistics fits a covariance model to the forecast error at time t using some form of stationarity assumption, while in sequential data assimilation the covariances are propagated from the past and are not necessarily stationary.

In the RRSQRT filter the error covariance of the corrected state $\mathbf{C}_t^\star$ depends exclusively on the initial covariances $\mathbf{C}_0$, the model error covariances, the model operator, the location of observations through the matrix $\mathbf{H}$ and the observation error $\mathbf{C}_t^o$ (generally composed of white noise covariances). So the RRSQRT filter does not actually learn from the data but depends exclusively on how the error matrices were calibrated. The EnKF depends on the way how the errors $\mathbf{q}_t^i$ and $\mathbf{u}_t^{o,i}$ are generated, yet this affects only the mean and not the error covariances $\mathbf{C}_t^\star$.

The bias filtering through universality conditions in applications requires more stations than the five that were available in our study of the Odra lagoon. With a few stations only it turns out that the results may deteriorate when including universality conditions.

ANAMORPHOSIS OF NON-GAUSSIAN VARIABLES

The data assimilation methods presented above imply Gaussian assumptions. In applications the distributions may be skew and an anamorphosis, i.e. a transformation of the distribution, as used in non-linear geostatistics, may be of advantage. This idea was tested performing a lognormal transform to reduce skewness when implementing an EnKF for three variables (nutrients, phytoplancton, herbivores) in the context of a simplified ecological model of a water column in the ocean (Bertino et al., 2003). It turned out that data assimilation with anamorphosis generated smaller errors than with the standard EnKF. In particular, the spring bloom, which is the principal cause of non-linearity in the dynamics, less perturbs

the filter with anamorphosis and the number of "false starts" of the phytoplancton bloom in springtime was significantly reduced.

MODELING THE SUPPORT EFFECT

It appears to be important in data assimilation problems to take account of the difference in the support of numerical model forecast and of observations, the support of the latter bein pointwise as compared to that of the state variables, which is of the size of the numerical model cells. Classical results in geostatistics have been adapted to the data assimilation context (Lajaunie and Wackernagel, 2000; Bertino, 2001).

For Gaussian state variables and observations the support correction resumes to an affine correction of the variances, because in the absence of bias the first moment of the observations is identical to that of the state variable by Cartier's relation.

In the framework of a lognormal model with assumption of permanence of lognormality for the different supports, merely the change of support coefficient needs to be inferred. The Gaussian anamorphosis generalizes the lognormal approach in the sense that it permits to transform an unspecified distribution towards a Gaussian distribution. The discrete Gaussian change of support model governed by a change of support coefficient can be applied in this context. Other change of support models like e.g. the gamma model studied by (Hu, 1988) could be used in the context of sequential data assimilation.

Finally it is also possible to work without an explicit change of support model in an approach based on geostatistical simulation on point support, where values on larger support are obtained by spatial averaging. By considering an ensemble of realizations empirical conditional distributions can then be easily computed.

The modelling of change of support has not yet been studied in detail and experimented in operational forecasting systems. This is due to the fact that there is a lack of awareness to implications of the support effect and this awareness is confined to domains in which geostatistics is already well known. Furthermore nonlinear geostatistical techniques need to be carefully adapted to applications in data assimilation in order to be able to add performance. A discrete Gaussian change of support model has been used for the purpose of downscaling air pollution forecasts at a resolution below that of the model cells by uniform conditioning (Wackernagel et al., 2004).

5 Conclusion

To keep the presentation simple, we have presented here the basic version of the ensemble Kalman filter as our main aim was to show the links and cross-fertilization potential between sequential data assimilation and geostatistical theory and methods. The EnKF is presently without doubt the most popular algorithm in sequential data assimilation (Mackenzie, 2003). Most recent developments (Evensen, 2004) can be found at the web site www.nersc.no/∼geir/EnKF/. Applications are found in many areas of operational forecasting for oceanography, meteorology,

environmental and ecological monitoring. While the Kalman filter is a classical tool in hydrogeology (Eigbe et al., 1998), some new developments could occur in petroleum reservoir modelling (Naevdal et al., 2002).

References

Bertino, L. (2001). *Assimilation de Données pour la Prédiction de Paramètres Hydrodynamiques et Ecologiques: Cas de la Lagune de l'Oder*. Doctoral thesis, Ecole des Mines de Paris, Fontainebleau. http://pastel.paristech.org.

Bertino, L., Evensen, G., and Wackernagel, H. (2002). Combining geostatistics and Kalman filtering for data assimilation in an estuarine system. *Inverse Problems*, 18:1–23.

Bertino, L., Evensen, G., and Wackernagel, H. (2003). Sequential data assimilation techniques in oceanography. *International Statistical Review*, 71:223–241.

Burgers, G., van Leeuwen, P. J., and Evensen, G. (1998). On the analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126:1719–1724.

Cañizares, R. (1999). *On the Application of Data Assimilation in Regional Coastal Models*. PhD thesis, TU Delft, Rotterdam.

Eigbe, U., Beck, M. B., Wheather, H. S., and Hirano, F. (1998). Kalman filtering in groundwater flow modelling: problems and prospects. *Stochastic Hydrology and Hydraulics*, 12:15–32.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *J. Geophysical Research*, 99(C5):10143–10162.

Evensen, G. (2003). The Ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367.

Evensen, G. (2004). Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics*, 54:539–560.

Hu, L. Y. (1988). *Mise en oeuvre du modèle gamma pour l'estimation des distributions spatiales*. Doctoral thesis, Ecole des Mines de Paris, Fontainebleau.

Kyriakidis, P. C. and Journel, A. G. (1999). Geostatistical space-time models: a review. *Mathematical Geology*, 31:651–684.

Lajaunie, C. and Wackernagel, H. (2000). Geostatistical approaches to change of support problems: Theoretical framework. IMPACT Project Deliverable Nr 19, Technical Report N–30/01/G, Centre de Géostatistique, Ecole des Mines de Paris, Fontainebleau.

Mackenzie, D. (2003). Ensemble Kalman filters bring weather models up to date. *SIAM News*, 36(8). http://www.siam.org/siamnews/10-03/tococt03.htm.

Naevdal, G., Mannseth, T., and Vefring, E. H. (2002). Near-well reservoir monitoring through ensemble Kalman filtering. *Society of Petroleum Engineers*, SPE 75235.

Sénégas, J., Wackernagel, H., Rosenthal, W., and Wolf, T. (2001). Error covariance modeling in sequential data assimilation. *Stochastic Environmental Research and Risk Assessment*, 15:65–86.

Verlaan, M. and Heemink, A. W. (1997). Tidal flow forecasting using reduced rank square root filters. *Stochastic Hydrology and Hydraulics*, 11(5):349–368.

Wackernagel, H., Lajaunie, C., Blond, N., Roth, C., and Vautard, R. (2004). Geostatistical risk mapping with chemical transport model output and ozone station data. *Ecological Modelling*, 179:177–185.

Wolf, T., Sénégas, J., Bertino, L., and Wackernagel, H. (2001). Application of data assimilation to three-dimensional hydrodynamics: the case of the Odra lagoon. In Monestiez, P., Allard, D., and Froidevaux, R., editors, *GeoENV II – Geostatistics for Environmental Applications*, pages 157–168, Amsterdam. Kluwer.