Methods of Spatial Data Processing Based on Bayesian Approach for Environmental Monitoring

Ekaterina G. Klimova

Institute of Computational Technologies SB RAS, Novosibirsk, Russia, klimova@ict.nsc.ru

Abstract. One of the important tasks of monitoring of environment is the problem of obtaining the values of environmental parameters on a regular grid. At present, such problems are solved using all available observational data as well as the mathematical model of the process of interest to us. The mathematical formulation of the problem is included in the set of tasks of the so-called inverse modelling. If the probabilistic formulation of the problem is considered, the Bayesian approach is applied. This approach is used in popular algorithms, such as the ensemble Kalman filter, the ensemble Kalman smoothing, the particle method. The report provides a brief overview of modern methods, as well as approaches to their practical implementation.

Keywords: data assimilation, ensemble Kalman filter, satellite data

1. Introduction

One of the important problems of environment monitoring is the problem of obtaining values of environmental parameters on a regular grid. At present, problems are solved using all available observational data as well as the mathematical model of the process of interest to us. The mathematical formulation of the problem is included in the set of problems of the so-called inverse modelling. The solution of the inverse modeling problem for a given process model and a set of observational data cyclically in time is the task of data assimilation. The problem of inverse modeling also includes the problem of estimation of the model parameters.

If the probabilistic formulation of the problem is considered, the Bayesian approach is applied. This approach is based on popular algorithms, such as the ensemble Kalman filter, the ensemble Kalman smoothing, the particle method. If the considered random variables are Gaussian, and the forecast and observation models are linear, this problem statement is equivalent to the variational statement of the data assimilation problem (4DVAR) [1].

The report provides a brief overview of modern data assimilation methods used in the environment modeling, as well as approaches to their practical implementation.

2. The inverse modeling problem

The inverse modeling problem is the problem of estimation of the unknown vector of parameters $\mathbf{x} \in R^M$ using the vector of observational data $\mathbf{y} \in R^P$ [2].
Let's suppose that

$$\mathbf{y} = M(\mathbf{x}) + \boldsymbol{\delta},$$

where M - the well-known prediction-observation operator, $\boldsymbol{\delta}$ is a random noise with a given distribution function. If the time series of observational data is considered, the change in time of the estimated variable is described using a mathematical model. The task of data assimilation is usually understood as the time-sequential estimation of an unknown quantity from observational data [2].

3. Bayesian approach to the data assimilation problem

Suppose that the time change of the estimated quantity $\mathbf{x}^k$ is described by the model

$$\mathbf{x}^{k+1} = f_{k+1,k}(\mathbf{x}^k) + \boldsymbol{\eta}^k,$$

where *k* is the time step number. In addition, observational data $\mathbf{y}^k$ are known:

$$\mathbf{y}^k = h_k(\mathbf{x}^k) + \boldsymbol{\varepsilon}^k.$$

In these formulas $\boldsymbol{\eta}^k$, $\boldsymbol{\varepsilon}^k$ are random errors of forecast and observations, respectively, $f_{k+1,k}$ - model operator, h_k - transformation operator of the predicted variable into the observable one.

The Bayesian approach consists in applying the Bayes theorem to obtain the optimal estimate from observational data and the prediction:

$$p(\mathbf{x}\,|\,\mathbf{y}) = \frac{p(\mathbf{y}\,|\,\mathbf{x})\,p(\mathbf{x})}{p(\mathbf{y})}.$$

Depending on which time period is needed for state estimation, it is possible to define three estimation problems: $p(\mathbf{x}_l\,|\,\mathbf{y}_{k,1}), k > l$ - forecast, $p(\mathbf{x}_k\,|\,\mathbf{y}_{k,1})$ - filtration, $p(\mathbf{x}_{k,0}\,|\,\mathbf{y}_{k,1})$ - smoothing, where $\mathbf{x}_{k,0} = \{\mathbf{x}_k, \mathbf{x}_{k-1}, \cdots, \mathbf{x}_0\}$, $\mathbf{y}_{k,1} = \{\mathbf{y}_k, \cdots, \mathbf{y}_1\}$. Notation and definitions are taken from the review [1].

In the linear Gaussian case, the solution of the filtering problem is the Kalman filter, the solution of the smoothing problem is the Kalman smoothing. In [3], it was proposed to use the Monte Carlo method for solving filtering and smoothing problems, the so-called ensemble filtering and smoothing algorithms. In the ensemble Kalman filter in the nonlinear case, the prediction of the covariance matrix is performed using a nonlinear model, and the condition of Gaussianity is violated. Also in this case, the estimate at the analysis step is an approximation of the estimate of the minimum variance. To solve the smoothing problem in the nonlinear case, iterative smoothing algorithms using ensembles are currently being considered [4]. As noted in [4], iterative methods provide a good approximation for weakly non-linear models. In the nonlinear non-Gaussian case, the particle method is used, which is also based on the Bayesian approach [1, 2].

4. The ensemble Kalman filter

The ensemble Kalman filter was proposed by Evensen G. [3]. Consider a nonlinear dynamic system as a process equation

$$\mathbf{x}_k^t = f(\mathbf{x}_{k-1}^t) + \boldsymbol{\eta}_{k-1}^t$$

and an observation equation

$$\mathbf{y}_k = h(\mathbf{x}_k^t) + \boldsymbol{\varepsilon}_k^t,$$

where h is the observation operator, generally speaking, non-linear, transforming the forecast values into the observable variable, $\boldsymbol{\eta}_{k-1}^t$ is the 'model noise' vector, $\boldsymbol{\varepsilon}_k^t$ is the vector of observation errors, $\mathbf{x}_k^t$ is the vector of estimated variables at the moment of time t_k, $\boldsymbol{\varepsilon}_k^t$ and $\boldsymbol{\eta}_{k-1}^t$ are Gaussian random variables: $E[\boldsymbol{\varepsilon}_k^t (\boldsymbol{\varepsilon}_k^t)^T] = \mathbf{R}_k^t$, $E[\boldsymbol{\eta}_{k-1}^t (\boldsymbol{\eta}_{k-1}^t)^T] = \mathbf{Q}_{k-1}^t$. We will assume $\mathbf{x}_k^t$ to be a 'true' value.

The stochastic ensemble Kalman filter consists of an ensemble of forecasts $\{\mathbf{x}_k^{f,n}, n = 1, \cdots, N\}$

$$\mathbf{x}_k^{f,n} = f(\mathbf{x}_{k-1}^{a,n}) + \boldsymbol{\eta}_{k-1}^n \tag{1}$$

and an ensemble of analysis $\{\mathbf{x}_k^{a,n}, n = 1, \cdots, N\}$

$$\mathbf{x}_k^{a,n} = \mathbf{x}_k^{f,n} + \mathbf{K}_k(\mathbf{y}_k^n + \boldsymbol{\varepsilon}_k^n - h(\mathbf{x}_k^{f,n})). \tag{2}$$

The ensembles (1) and (2) provide a sample of 'true' values. Here the sample mean will be the optimal estimate, and the deviations from the mean will be the ensembles of the analysis and forecast errors, respectively. To implement the ensemble version of the Kalman filter algorithm, an ensemble of observation errors $\{\boldsymbol{\varepsilon}_k^n, n = 1, \cdots, N\}$ and ensemble of forecast errors $\{\mathbf{dx}_k^{f,n} = \mathbf{x}_k^{f,n} - \overline{\mathbf{x}_k^{f,n}}, n = 1, \cdots, N\}$, where $\overline{\mathbf{x}_k^{f,n}} \cong \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_k^{f,n}$, and an ensemble of model noise $\{\boldsymbol{\eta}_{k-1}^n, n = 1, \cdots, N\}$, $E[\boldsymbol{\eta}_{k-1}^n (\boldsymbol{\eta}_{k-1}^n)^T] = \mathbf{Q}_k$ are specified. $\mathbf{K}_k$ is a matrix of the form

$$\mathbf{K}_k = \mathbf{P}_k^f \mathbf{H}_k{}^T (\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k{}^T + \mathbf{R}_k)^{-1},$$

where $\mathbf{P}_k^f$ and $\mathbf{R}_k$ are matrices that are estimated by the ensemble averaging as

$$\mathbf{P}_k^f \approx \frac{1}{N-1}\sum_{n=1}^{N} \mathbf{dx}_k^{f,n} (\mathbf{dx}_k^{f,n})^T, \quad \mathbf{R}_k \approx \frac{1}{N-1}\sum_{n=1}^{N} \boldsymbol{\varepsilon}_k^n (\boldsymbol{\varepsilon}_k^n)^T,$$

And $\mathbf{H}_k$ is the linearized operator of $h(\mathbf{x}_k^{f,n})$ with respect to $\overline{\mathbf{x}_k^{f,n}}$:

$$h(\mathbf{x}_k) \cong h(\mathbf{x}_k^{f,n}) + \mathbf{H}_k \boldsymbol{\varepsilon}_k^f \quad .$$

Formulas (1) - (2) are a stochastic ensemble Kalman filter. The deterministic ensemble Kalman filter (analysis step) consists of an equation for the mean

$$\overline{\mathbf{x}_k^{a,n}} = \overline{\mathbf{x}_k^{f,n}} + \mathbf{K}_k(\mathbf{y}_k^n - \overline{h(\mathbf{x}_k^{f,n})})$$

and an estimate of the ensemble of analysis errors such that the corresponding covariance matrix satisfies the Kalman filter equation $\mathbf{P}_k^a = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)\mathbf{P}_k^f$ [5].

5. The problem of optimal smoothing. Ensemble smoothing (EnKS - ensemble Kaman smoother)

In the classical optimal estimation theory, the problem of optimal filtering is the problem of obtaining an optimal estimate at the end of the considered time interval. Also, there is an optimal smoothing problem, which is the problem of obtaining the optimal estimate for the given time interval. We suppouse, that an optimal estimate minimizes the trace of the covariance matrix of estimation errors [6].

Ensemble algorithms are also applied for the optimal smoothing problem. In [3] was shown that the problem of ensemble smoothing can be solved sequentially in time by using the available data at each time step to estimate values over the given time interval. In this case, the formulas for the optimal estimate are similar to those of the ensemble Kalman filter. With this variant of ensemble smoothing (EnKS), calculations computing in the opposite direction of time are not required. It should be noted that the EnKS algorithm is equivalent to the variational data assimilation algorithm with the use of the corresponding weight matrices (4DVAR) [1, 2].

6. The parameters estimation in the data assimilation algorithm

Consider the equation of process as

$$\mathbf{x}_k^t = f(\mathbf{x}_{k-1}^t, \boldsymbol{\alpha}_{k-1}^t) + \boldsymbol{\eta}_{k-1}^t,$$

observations are

$$\mathbf{y}_k = h(\mathbf{x}_k^t, \boldsymbol{\alpha}_k^t) + \boldsymbol{\varepsilon}_k^t,$$

where $\boldsymbol{\alpha}_k^t$ is the vector of parameters. We assume that the parameter does not change with time: $\boldsymbol{\alpha}_{k+1}^t = \boldsymbol{\alpha}_k^t$. Consider the generalized estimation problem for vector $\mathbf{z} = [\mathbf{x}, \boldsymbol{\alpha}]^T$. Omitting the intermediate calculations, we write down the result of the estimation procedure in general:

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{P}_{xx} h_x^T (h_x \mathbf{P}_{xx} h_x^T + \mathbf{R})^{-1} [\mathbf{y} - h(\mathbf{x}^f, \boldsymbol{\alpha}^f)] + \mathbf{P}_{x\alpha} h_\alpha^T (h_x \mathbf{P}_{xx} h_x^T + \mathbf{R})^{-1} [\mathbf{y} - h(\mathbf{x}^f, \boldsymbol{\alpha}^f)],$$

$$\boldsymbol{\alpha}^a = \boldsymbol{\alpha}^f + \mathbf{P}_{\alpha x} h_x^T (h_x \mathbf{P}_{xx} h_x^T + \mathbf{R})^{-1} [\mathbf{y} - h(\mathbf{x}^f, \boldsymbol{\alpha}^f)] + \mathbf{P}_{\alpha\alpha} h_\alpha^T (h_x \mathbf{P}_{xx} h_x^T + \mathbf{R})^{-1} [\mathbf{y} - h(\mathbf{x}^f, \boldsymbol{\alpha}^f)].$$

In these formulas, the index 'k' is omitted. The analysis step is considered, the index 'a' means analysis, the index 'f' means the prediction, $\mathbf{P}_{\alpha x}$ - the cross-covariance of errors $\mathbf{x}$ and $\boldsymbol{\alpha}$, $\mathbf{P}_{\alpha\alpha}$ - the covariance matrix of errors $\boldsymbol{\alpha}$, h_x and h_α - linearized operators with respect to $\mathbf{x}$ and $\boldsymbol{\alpha}$. If h does not depend on $\boldsymbol{\alpha}$, the estimation $\mathbf{x}^a$ is carried out using the same formula as in the classical Kalman filter [6]. In modern works on data assimilation, such an approach is applied to the estimation of greenhouse gas fluxes [7].

7. Variational approach to the data assimilation problem

The variational approach to the data assimilation problem is very popular and is used in the operational data assimilation systems in the leading prognostic centers [1, 2]. The data analysis algorithm called 3DVAR (3-Dimensional VARiational) consists in finding a vector $\mathbf{x}_\mathbf{a}$ that minimizes the functional

$$2\mathrm{J}(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_\mathrm{b})^\mathrm{T} \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_\mathrm{b}) + [\mathbf{y}_0 - \mathrm{h}(\mathbf{x})]^\mathrm{T} \mathbf{R}^{-1} [\mathbf{y}_0 - \mathrm{h}(\mathbf{x})].$$

In this formula, **B** and **R** are the covariance matrices of forecast and observation errors, respectively. The operator *h* is, generally speaking, nonlinear, transforms prognostic value to the observations (observed variables).

Minimum of functional is the solution of the equation

$$\nabla \mathrm{J}(\mathbf{x}_\mathrm{a}) = 0.$$

If the gradient of the functional is written in the following form:

$$\nabla \mathrm{J}(\mathbf{x}) = \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_\mathrm{b}) + \mathbf{H}^\mathrm{T} \mathbf{R}^{-1} \mathbf{H}(\mathbf{x} - \mathbf{x}_\mathrm{b}) - \mathbf{H}^\mathrm{T} \mathbf{R}^{-1} \{\mathbf{y}_0 - \mathrm{h}(\mathbf{x}_\mathrm{b})\},$$

then the solution of the problem is

$$\mathbf{x}_\mathrm{a} = \mathbf{x}_\mathrm{b} + (\mathbf{B}^{-1} + \mathbf{H}^\mathrm{T} \mathbf{P}^{-1} \mathbf{H}) \{\mathbf{y}_0 - \mathrm{h}(\mathbf{x}_\mathrm{b})\}.$$

Where **H** is the linearized operator *h* with respect to $\mathbf{x}_\mathrm{b}$. This formula is equivalent to the analysis step of the Kalman filter algorithm [1, 2, 6].

The algorithm 4DVAR is a generalization of 3DVAR to the space-time case. 4DVAR allows you to use observations from a certain time interval $(\mathrm{t_n} - \mathrm{t_0})$. It is assumed that observations $\{\mathbf{y}_\mathrm{i}^0, \mathrm{i} = 0, \cdots, \mathrm{N}\}$ are specified at this interval. It is required to find a minimum of functional

$$\mathrm{J}[\mathbf{x}(\mathrm{t}_0)] = \frac{1}{2} [\mathbf{x}(\mathrm{t}_0) - \mathbf{x}^\mathrm{b}(\mathrm{t}_0)]^\mathrm{T} \mathbf{B}_0^{-1} [\mathbf{x}(\mathrm{t}_0) - \mathbf{x}^\mathrm{b}(\mathrm{t}_0)] + \frac{1}{2} \sum_{\mathrm{i}=0}^{\mathrm{N}} [\mathrm{h}(\mathbf{x}_\mathrm{i}) - \mathbf{y}_\mathrm{i}^0]^\mathrm{T} \mathbf{R}^{-1} [\mathrm{h}(\mathbf{x}_\mathrm{i}) - \mathbf{y}_\mathrm{i}^0]$$

provided that

$$\mathbf{x}(\mathrm{t_n}) = \mathrm{M_n}[\mathbf{x}(\mathrm{t}_0)].$$

In this optimization problem, the control variable $\mathbf{x}(\mathrm{t}_0)$ is the value at the initial time. That is, the algorithm 4DVAR finds an initial condition such that a forecast with this initial value best approximates observations in a given time interval. The minimum functional in 4DVAR is searched by iterative methods (for example, the quasi-Newton method). To implement the iterative process, we need to estimate the gradient of the functional J(x). The gradient of functional is calculated using the conjugate linearized model. At present, there are data assimilation algorithms based on 4DVAR, with use the ensemble approach to estimate changes in covariances over time [1, 2].

8. Problems of implementation of ensemble algorithms

Tasks of data assimilation, the estimation of the parameters in the environment modelling has a very large dimension, requires huge computing resources. The use of ensemble algorithms partially solves this problem, but still the problem remains extremely time-consuming. To carry out the analysis step of ensemble Kalman filter, algorithms of transformation of the ensemble of forecasts are used to obtain the ensemble of the analysis. In this case, the analysis is performed only for the ensemble mean value, and then the ensemble of analyses is calculated. There are deterministic and stochastic variants of such algorithms for realization of both ensemble Kalman filter and ensemble Kalman smoother [5].

A variant of a stochastic Kalman filter with ensemble transformations is the ensemble π-algorithm [8-10]. This algorithm performs operations with matrices of the order of the ensemble dimension. In addition, this algorithm can be used to implement ensemble smoothing (EnKS). An important property of the algorithm is its locality. The estimate

the desired value can be implemented independently for the given subdomains. This property can be used for estimation of parameters locally in the given region.

When using ensembles in data assimilation, there are a number of problems associated with a small sample size (ensemble size). In particular, it is a problem of high correlation values of forecast errors at long distances. In this case, the so-called localization may be used, namely, the element-by-element multiplication of covariance matrices by the function decreasing with the distance [1,5].

The small sample size is also one of the reasons for the divergence of the algorithm. The algorithms of the ensemble Kalman filter (EnKF) and ensemble Kalman smoothing (EnKS) use the 'inflation factor'. In this case the elements of the covariance matrix are multiplied by a certain multiplier (multiplicative inflation), or an additional perturbation is added to the ensemble of perturbations (additive inflation). It should be noted that these techniques are also used in the iterative smoothing algorithms using ensembles, as well as in the particle method using ensembles [1, 2, 5].

Consider the features of the assimilation of satellite data [11]. Satellite data have a number of distinctive features compared to ground-based and upper-air observations. These features include the following:

(a) The satellite measures radiation information, so it is necessary to convert model variables into observable variables using the radiation transfer equation. Thus, the observation operator h is nonlinear.

b) Observations are received continuously over time.

C) Observation errors correlate, the matrix of observational errors has large dimension and it is non-diagonal.

d) There is a of systematic error in observations (bias). This makes it difficult to use the formulas of the above algorithms.

Satellite data can be used in two ways: directly, in the form of radiation data, or it is required to obtain meteorological values from satellite data in advance and then use them in the data assimilation procedure. The first option is more preferable for a number of reasons:

(a) The transformation of radiation data uses forecast information, so the errors of the recovered data and forecast will be correlated.

b) There is the problem of estimating the covariance matrix of observation errors.

Satellite data is a huge amount of information. However, as noted in a number of papers, a large number of observations do not always give the best result, since these data are an array of correlating random variables and therefore they are not too informative [11].

9. The data assimilation in the environment modelling

The modern environmental monitoring is based on the mathematical modeling. In particular, models of the propagation of greenhouse gases, active gas constituents, and aerosols in the atmosphere are considered.

Information on all these substances is measured using both ground-based measuring instruments and satellites. A large number of investigations are currently devoted to the environmental data assimilation problem, using the basic mathematical apparatus developed in the meteorological data assimilation [7].

An important task in monitoring of the environment is to assess the fluxes from the earth's surface of greenhouse gases with the help of the data assimilation system [7].

The problems of the environment modeling are currently using models of the atmosphere with a "chemical" block. As an example, the popular model WRF-Chem [12] may be considered. Modern databases of meteorological fields (ERA-interim reanalysis [13], for example) are used for numerical experiments. ECMWF has created a MACC database (Monitoring Atmospheric Composition and Climate), which includes active gas components, aerosols and greenhouse gases [14]. In the regional modelling, an important problem is the determination of boundary values. Reanalysis data may be used for this purpose. Also, in the literature, methods of data assimilation for determination values at the boundary of the region are considered [7].

10. Conclusion

The report provides a brief overview of current approaches to the problem of data assimilation in the environment modelling. The main aspects of practical realization of this task are considered.

References

[1] Carrassi A., Bocquet M., Bertino L., Evensen G. Data assimilation in the geosciences: An overview of methods, issuers and perspectives // Wiley interdisciplinary reviews: Climate Change. 2018. Vol. 131. Issue5, e535.

[2] Nakamura G., Potthast R. Inverse Modeling. IOP Publishing. 2015.

[3] Evensen, G. Data assimilation. The ensemble Kalman filter.Berlin Heideberg: Spriger-Verlag, 2009.

[4] Evensen G. Analysis of iterative ensemble smoother for solving inverse problems //Computational Geosciences. 2018. Vol. 22. P.885-908.

[5] Houtekamer, H.L. Zhang, F. Review of the ensemble Kalman filter for atmospheric data assimilation // Monthly Weather Review. 2016. Vol. 144. P. 4489-4532.

[6] Jazwinski, A.H. Stochastic processes and filtering theory. New York: Academic Press, 1970.

[7] Bocquet M.et al. Data assimilation in atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models // Atmospheric Chemistry and Physics. 2015. Vol. 15. P. 5325-5358.

[8] Klimova, E. A suboptimal data assimilation algorithm based on the ensemble Kalman filter // Quarterly Journal of the Royal Meteorological Society. 2012. Vol. 138. P. 2079-2085.

[9] Klimova E.G. Application of ensemble Kalman filter in environment data assimilation // IOP Conference Series: Earth and Environmental Science. 2018. Vol. 211. P. 012049.

[10] Klimova E G 2019 The Kalman stochastic ensemble filter with transformation of perturbation ensemble // Siberian. J. Num. Math. Vol. 22. No.1. P. 27-40.

[11] Thepaut J.-N. Satellite data assimilation in numerical weather prediction: an overview // Proceedings of the ECMWF Seminar on Recent development in data assimilation for atmosphere and ocean, 8-12 September 2003. 2003. P.75-96.

[12] Grell G. et al. Fully coupled “online” chemistry within the WRF model // Atmos. Environment. 2005. Vol. 39. P. 6957- 6975.

[13] Dee D.P. et al. The ERA-interim reanalysis: Configuration and performance of the data assimilation system // Q. J. Roy. Meteorol. Soc. 2011. Vol. 137. P. 553-579.

[14] Inness A. et al. The MACC reanalysis: an 8 yr data set of atmospheric composition // Atmos. Cgem. Phys. 2013. Vol. 13. P. 4073-4109.