Bayesian approach to data assimilation based on ensembles of forecasts and observations

E G Klimova
Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Prosp. Lavrentyeva 6, Russia, 630090, klimova@ict.nsc.ru

Abstract. Optimal assessment of geophysical fields with observational data and a mathematical model is a data assimilation problem. To solve it, a Bayesian approach is most often used. In the ensemble algorithms, ensembles of forecasts and observations are used to approximate the covariance matrices considered in the algorithm. If all probability densities are Gaussian, the problem is reduced to that of the ensemble Kalman filter. In the strongly non-Gaussian case, a particle method is used, which is based on a Bayesian approach. Ensembles are also used in this method. In the research devoted to the ensemble Kalman filter much attention is paid to deviation of ensemble elements from an average value – ensemble spread. In this paper, a comparative analysis of spread behavior over time when using different approaches to the improvement of the convergence of the algorithms is performed. The results of numerical experiments with a 1-dimensional test model are discussed. These results show that in stochastic filters the behavior over time of ensemble spread is close to the theoretical error estimate. Some of the approaches that improve convergence in the ensemble filter, such as additive inflation and multiplicative inflation, change the general formula for ensemble spread.

1. Introduction
At present, optimal assessment of geophysical fields with observational data and a mathematical model is one of the most important problems. A Bayesian approach is often used to solve it. If the considered random variables are Gaussian and the forecast and observation models are linear, the problem is equivalent to a variational data assimilation problem (4DVAR) [1].

For approximation of the covariance matrices considered in the algorithm, an ensemble of forecasts and observations can be used [2]. If the optimal estimate is sought at the end of a given time interval, the problem is reduced to the ensemble Kalman filter. If an optimal estimate is considered in a given time interval, this task is called an ensemble smoothing problem [2].

Ensemble filtering and smoothing algorithms can be divided into deterministic and stochastic. In the stochastic Kalman filter, observational data with ‘perturbations’ are considered. This allows you to make the ensemble of deviations from the mean value the same as the ensemble of estimation errors [3]. The ensemble Kalman filter and ensemble Kalman smoothing, proposed in [2], are stochastic algorithms. Along with these algorithms, there are a large number of ‘square root’ filters in which the analysis is performed for the mean over the ensemble, and the ensemble of analysis errors is obtained from the ensemble of forecast errors using the transformation matrix. Such algorithms are called deterministic. Comparative analysis of deterministic and stochastic filters is carried out in many papers [3].
Review [3] lists the main approaches to improving the convergence of the ensemble Kalman filter. One of the popular ways to control the convergence of the ensemble Kalman filter is to use the so-called ‘inflation factor’. In [3], some ways of improving of the convergence of the ensemble Kalman filter are noted, such as a multiplicative inflation and an additive inflation.

In the nonlinear case, iterative smoothing algorithms are currently used, which are a variational problem using ensembles to calculate the covariances and linearize the prediction and observation operators [1]. In the strictly nonlinear non-Gaussian case, the particle method is used, which is based on the Bayesian approach. Ensembles are also used in this method. These approaches also consider methods for controlling the convergence of the algorithms.

In papers on the ensemble Kalman filter, much attention is given to the behavior of deviations of an ensemble member from the mean (ensemble spread). It is shown in [4] that the spread of an ensemble distribution can in principle give a priori indication of forecast skill and also that the spread correlates with the forecast error. Therefore, it is important for the behavior of the spread in time to correspond to the behavior of the estimation error.

In an ensemble Kalman filter, numerous factors, such as the limited number of ensemble members (sample error), the inaccurately specified random observation errors, model noise, nonlinearity of the model, etc., affect the process of convergence and the relationship between spread and skill.

In the present paper, an approach to analyzing the behavior over time of ensemble spread and its relationship with the behavior of the estimation error based on a general formula for the deviation of an ensemble member from the ensemble mean value are considered. A comparison of the behavior over time of these quantities for deterministic and stochastic filters is made. The effects of major means for improving convergence in the ensemble Kalman filter (additive inflation and multiplicative inflation) on the behavior over time of ensemble spread are investigated.

2. Bayesian approach to data assimilation

Assume that the time change of the estimated quantity $x^k$ is described by the model

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

where $k$ is the time step number. In addition, observational data $y^k$ are known:

$$y^k = h_k(x^k) + \epsilon^k.$$

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

The Bayesian approach consists of applying the Bayes theorem to obtain an optimal estimate from observational data and a forecast:

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}.$$

There are various options for assessing the state of data and forecast: $p(x_k \mid y_{k,1})$, $k > l$ - forecast, $p(x_l \mid y_{l,1})$ - filtering, $p(x_{k,0} \mid y_{k,1})$ - smoothing, where $x_{k,0} = \{x_k, x_{k-1}, \ldots, x_0\}$, $y_{k,1} = \{y_k, \ldots, y_l\}$. The notation and definitions are taken from review [1].

In the linear Gaussian case, the solution of the filtering problem is the Kalman filter, and the solution of the smoothing problem is the Kalman smoothing [5]. In [2], 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 covariance matrix is predicted using a nonlinear model, and the Gaussian condition is violated. Also in this case, the estimate at the analysis step is an approximation of the minimum variance estimate [2].

To solve the smoothing problem in the nonlinear case, iterative smoothing algorithms using ensembles are currently being considered [1]. In the nonlinear non-Gaussian case, the particle method is used, which is also based on the Bayesian approach. To implement a particle filter, algorithms using
the representation of the probability distribution density as a sum with the coefficients of Gaussian probability distribution densities may be considered. In this case, the algorithm is a generalization of the ensemble Kalman filter [5].

3. The ensemble Kalman filter

Consider a nonlinear dynamic system as a process equation

\[ \mathbf{x}_k = f(\mathbf{x}_{k-1}) + \mathbf{n}_{k-1} \]

and an observation equation

\[ \mathbf{y}_k = h(\mathbf{x}_k) + \mathbf{e}_k, \]

where \( \mathbf{n}_{k-1} \) is a ‘model noise’ vector, \( \mathbf{e}_k \) is an observation error vector, \( \mathbf{x}_k \) is a vector of the variables being estimated at time \( t_k \), and \( \mathbf{e}_k \) and \( \mathbf{n}_{k-1} \) are Gaussian random variables: \( E(\mathbf{e}_k) = 0 \), \( E(\mathbf{n}_{k-1}) = 0 \), \( \mathbb{E}[\mathbf{e}_k(\mathbf{e}_k)^T] = \mathbf{R}_k \), \( \mathbb{E}[\mathbf{n}_{k-1}(\mathbf{n}_{k-1})^T] = \mathbf{Q}_{k-1} \). Here \( \mathbf{x}_k \) is assumed to be the ‘true’ value.

Consider an ensemble approach in which probability mean is replaced by sample mean. In this case, it is typically assumed that the ‘true state’ can be initially represented by a sample \( \{ \mathbf{x}_0 + \delta \mathbf{x}_n, n = 1, \ldots, N \} \), where \( \mathbf{x}_0 \) is a known estimate of the sought-for quantity, and the sample \( \{ \delta \mathbf{x}_n, n = 1, \ldots, N \} \) simulates random errors of this estimate. Assume that the sample mean of this error is zero. The time variation of the ‘true state’ is simulated by the process equation. Then the sample mean will be the optimal estimate, and the deviation from the mean will be the estimation error.

The ensemble Kalman filter consists of an ensemble of forecasts \( \{ \mathbf{x}^{f,n}_k, n = 1, \ldots, N \} \)

\[ \mathbf{x}^{f,n}_k = f(\mathbf{x}^{a,n}_{k-1}) + \mathbf{n}^n_{k-1} \]

and an ensemble of analyses \( \{ \mathbf{x}^{a,n}_k, n = 1, \ldots, N \} \)

\[ \mathbf{x}^{a,n}_k = \mathbf{x}^{f,n}_k + \mathbf{K}_k (\mathbf{y}^n_k + \mathbf{e}^n_k - h(\mathbf{x}^{f,n}_k)). \]

The ensembles (2) and (3) provide a sample of ‘true’ values. Here the sample mean will be the optimal estimate, and the deviations from the mean will be ensembles of analysis and forecast errors, respectively. To implement the ensemble version of the Kalman filter algorithm, an ensemble of observation errors \( \{ \mathbf{e}^n_k, n = 1, \ldots, N \} \), an ensemble of forecast errors \( \{ \mathbf{d}^{f,n}_k = \mathbf{x}^{f,n}_k - \mathbf{d}^{f,n}_k, n = 1, \ldots, N \} \), where \( \overline{\mathbf{d}^{f,n}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}^{f,n}_k \), and an ensemble of model noise \( \{ \mathbf{n}^n_{k-1}, n = 1, \ldots, N \} : E(\mathbf{n}^n_{k-1}(\mathbf{n}^n_{k-1})^T) = \mathbf{Q}_{k-1} \) are specified. \( \mathbf{K}_k \) is a matrix of the form

\[ \mathbf{K}_k = \mathbf{P}^f_k \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}^f_k \mathbf{H}_k^T + \mathbf{R}_k)^{-1}, \]

where \( \mathbf{P}^f_k \) and \( \mathbf{R}_k \) are matrices that are estimated over the ensemble as

\[ \mathbf{P}^f_k = \frac{1}{N-1} \sum_{n=1}^{N} \mathbf{d}^{f,n}_k (\mathbf{d}^{f,n}_k)^T, \]
\[ \mathbf{R}_k = \frac{1}{N-1} \sum_{n=1}^{N} \mathbf{n}^n_k (\mathbf{n}^n_k)^T, \]

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

\[ h(\mathbf{x}_k) \approx h(\overline{\mathbf{x}^{f,n}}) + \mathbf{H}_k \mathbf{e}_k. \]

With a given analysis error \( \{ \mathbf{d}^{a,n}_k = \mathbf{x}^{a,n}_k - \overline{\mathbf{x}^{a,n}}_k, n = 1, \ldots, N \} \), where \( \overline{\mathbf{x}^{a,n}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}^{a,n}_k \), the analysis error covariance matrix can be calculated by the formula

\[ \mathbf{P}^a_k = \frac{1}{N-1} \sum_{n=1}^{N} \mathbf{d}^{a,n}_k (\mathbf{d}^{a,n}_k)^T. \]
Formulas (2) and (3) represent a stochastic ensemble Kalman filter [2]. These formulas can be written in the following form:

$$x^k_i = (I - K_i H_i) \left[ f(x^i_{i-1}) + \eta^i_{i-1} \right] + K_i (y^i_i + \epsilon^i_i),$$

where $x^i_k$ is the ensemble of variables being estimated. As noted above, the optimal estimate in the ensemble Kalman filter is the ensemble mean value $\overline{x^i_k}$. Deviation from the mean (spread) simulates the estimate error with the equation

$$dx^k_i = (I - K_i H_i)(f(x^i_{i-1}) - K_i^\top \eta^j_{i-1}) + K_i \epsilon^i_i,$$ (4)

A 'theoretical' estimation error, $dx^j_i = x^j_i - \overline{x^j_i}$, satisfies the equation

$$dx^j_i = (I - K_j H_j)(f(x^j_{i-1}) - K_j^\top \eta^j_{i-1}) + K_j \epsilon^j_i.$$ (5)

The deterministic ensemble Kalman filter (analysis step) consists of the following equation for the mean:

$$\overline{x^i_k} = x^{f,i}_k + K_i (y^i_i - h(x^{f,i}_k))$$

and an estimate of the ensemble of analysis errors such that the corresponding covariance matrix satisfies the Kalman filter equation $P^k = (I - K_k H_k)P^f_k$.

As noted in some papers, the transformation of an ensemble of forecast errors into analysis errors in a deterministic filter can be represented in the form of left multiplication [6]:

$$dx^{f,i}_k = A_i dx^{f,i}_k,$$

where $A_i$ is an operator that satisfies the equation $A_i P^i_f A_i^\top = P^i_f$, and the operator $A_i$ may be not unique. The ensemble of analyses of the deterministic filter has the form

$$x^{a,i}_k = \overline{x^{a,i}_k} + A_i (x^{f,i}_k - \overline{x^{f,i}_k}).$$

There are many analysis ensembles corresponding to a covariance matrix $P^f_k$ and a mean value $\overline{x^{a,i}_k}$. Specifically, deterministic filters calculate an ensemble of fields with a corresponding mean value and covariance matrix [3]. However, the analysis ensemble specified by formula (3) has the following important property: the equation for deviation from the mean ensemble value (4) will have the same form as the equation for the theoretical estimation error (5).

It is shown in [6] that the ensemble of analyses being estimated has different distribution functions in deterministic and stochastic filters. It is evident that the same is true for the ensemble spread.

4. Analysis of ensemble spread behavior over time in the ensemble Kalman filter

The equation for ensemble spread in the stochastic Kalman filter may be written in the following form:

$$dx^k_i = (I - K_i H_i) (F_i dx^i_{i-1} + \eta^i_{i-1}) + K_i \epsilon^i_i.$$ (6)

Here, instead of the nonlinear model operator $f$, we take the linearized operator $F_i$. This equation can be rewritten as

$$dx^k_i = \Psi(k,0) dx^0_i + \sum_{j=1}^{i-1} \Psi(k,i) K_i \epsilon^j_i + \sum_{j=1}^{i-1} \Psi(k,i) (I - K_j H_j) \eta^j_{i-1},$$ (7)

where $\Psi(k,i) = \prod_{j=i+1}^{i} (I - K_j H_j) F_j$, if $k > i$, $\Psi(k,k) \equiv I$. Note that a similar equation for the classical Kalman filter was obtained in [7, 8].

The estimation error (i.e., the deviation of the mean from the ‘true’ value) satisfies the following equation (which is similar to (6)):

$$dx^j_i = (I - K_j H_j) (F_i dx^i_{i-1} + \eta^i_{i-1}) + K_j \epsilon^j_i.$$ (8)
It follows from equations (6) for $dx^n_k$ and (8) for $dx^n_k$ that the simulated estimation error tends to the theoretical error if the random vectors of observational errors and model noise being simulated have the same covariance matrices as the true ones. The random vectors of observational errors and model noise being simulated are considered here in the context of the theory of Kalman filtering, that is, they are assumed to be Gaussian white noise with a given covariance matrix [7].

In deterministic filters the analysis is made for the mean. Then some perturbations $an_k$ corresponding to the covariance matrix of analysis errors are generated. Writing the formula for the analysis perturbations in terms of 'left multiplication', we obtain the following equation for ensemble spread in the deterministic Kalman filter:

$$dx^n_k = A_k(F_k dx^n_{k-1} + \eta^n_{k-1})$$

or

$$dx^n_k = \Psi_{det}(k,0)dx^n_0 + \sum_{i=1}^k \Psi_{det}(k,i)A_i \eta^n_{i-1} ,$$

where $\Psi_{det}(k,i) = \prod_{j=i+1}^{k} A_j F_j$. It is evident that the behavior over time of ensemble spread in the deterministic Kalman filter differs from that in the stochastic filter, as well as from the behavior over time of the estimation error. Specifically, the formula for the deterministic filter lacks the term with $K_i e^n_i$, which simulates ensemble spread as a function of observational data distribution and observational error covariances. Note that in the linear case the covariance matrices of the stochastic and deterministic filters differ only by 'sampling error'.

5. Methods of improving convergence in the ensemble Kalman filter

Let us consider the ensemble spread modification in general form. In the first case when the modification is used in the analysis step it has the form

$$dx^n_k = \alpha_k dx^n_k + \beta^n_k ,$$

where $\beta^n_k$ is a random vector with a specified covariance matrix, and in the second case when the modification is used in the forecast step, the form

$$dx^n_k = (I - K_k H_k)(\alpha_k F_k dx^n_{k-1} + \alpha_k \eta^n_{k-1} + \beta^n_k) + K_k e^n_k .$$

Then in the first case the formula for a stochastic filter is

$$dx^n_k = \tilde{\Psi}(k,0)dx^n_0 + \sum_{i=1}^k \tilde{\Psi}(k,i)\alpha_i K_i e^n_i + \sum_{i=1}^k \tilde{\Psi}(k,i)(I - \tilde{K}_i H_i)\eta^n_{i-1} + \sum_{i=1}^k \tilde{\Psi}(k,i)\beta^n_i ,$$

and in the second case,

$$dx^n_k = \tilde{\Psi}(k,0)dx^n_0 + \sum_{i=1}^k \tilde{\Psi}(k,i)\tilde{K}_i e^n_i + \sum_{i=1}^k \tilde{\Psi}(k,i)(I - \tilde{K}_i H_i)\eta^n_{i-1} + \sum_{i=1}^k \tilde{\Psi}(k,i)(I - \tilde{K}_i H_i)\beta^n_i ,$$

where $\tilde{\Psi}(k,i) = \prod_{j=i+1}^{k} (\alpha_j - \tilde{K}_j H_j) F_j$, $\tilde{K}_j$ is calculated using the modified covariance matrices.

Consider the case of a deterministic filter. Let us denote $\tilde{\Psi}_{det}(k,i) = \prod_{j=i+1}^{k} \alpha_j A_j F_j$. We have for a deterministic filter in the first case (analysis step)

$$dx^n_k = \tilde{\Psi}_{det}(k,0)dx^n_0 + \sum_{i=1}^k \tilde{\Psi}_{det}(k,i)A_i \eta^n_{i-1} + \sum_{i=1}^k \tilde{\Psi}_{det}(k,i)\beta^n_i ,$$

and in the second case (forecast step)

$$dx^n_k = \tilde{\Psi}_{det}(k,0)dx^n_0 + \sum_{i=1}^k \tilde{\Psi}_{det}(k,i)A_i \eta^n_{i-1} + \sum_{i=1}^k \tilde{\Psi}_{det}(k,i)A_i \beta^n_i .$$
The equation for the error when modifying the ensemble spread takes the form
\[ dx'_k = \Psi_f (k,0) \Delta \xi_0 + \sum_{i=1}^k \Psi_f (k,i) \Delta \xi_i + \sum_{i=1}^k \Psi_f (k,i) (I - \bar{K} \bar{H}) \eta_i, \]
where \( \Psi_f (k,i) = \prod_{j=i+1}^k (I - \bar{K} \bar{H}) F_j \).

One can see from the above formulas that the perturbation ensembles of deterministic and stochastic filters with the thus modified ensemble spread do not correspond to the error ensemble. For \( \beta^a = 0 \) we obtain a version of multiplicative inflation. For \( \alpha_a = 1 \) we obtain a version of additive inflation. In this case, additive inflation can be specified so that the covariance matrix after correction coincides with the matrix obtained when using multiplicative inflation: \( \beta^a = \delta Q_1 \xi^a \), where
\[ \delta Q_1 = (\alpha_a^2 - 1) P_1. \]

Here \( P_1 \) is the covariance matrix of analysis or forecast errors corresponding to the above versions of modification of ensemble spread, \( \xi^a \) is a Gaussian random vector \( N(0,1) \), and \( I \) is a unit matrix. Here \( N(a,B) \) is a random vector distributed normally with mean \( a \) and covariance \( B \). In this case we obtain a version of additive inflation in which the covariance matrix is the same as the matrix obtained when using multiplicative inflation, but the ensemble spreads are different.

It follows from the formulas presented in this section that in the case of additive inflation the behavior of stochastic filter spread over time is closer to that of the error than in the case of multiplicative inflation (the matrices \( \Psi \) and \( \Psi_f \) are closer). In the case of additive inflation the formulas after the forecast step and after the analysis step differ only in the multiplier of \( \beta^a \). However, when using additive inflation after the analysis step the 'theoretical' relation between the covariance matrices, \( P_1 = (I - \bar{K} \bar{H}) \eta F_1 \) [7], is violated.

It should be noted that in the practical implementation of the particle filter, similar methods are used to improve the convergence of the algorithm over time.

6. Numerical experiments with the Lorenz-96 model

In the present paper, numerical experiments on data assimilation are performed for an ensemble Kalman filter with perturbed observations (a stochastic ensemble Kalman filter) and a version of a deterministic filter (LETKF algorithm [9]).

6.1. Ensemble \( \pi \)-algorithm

A version of a stochastic ensemble Kalman filter in square root form (ensemble \( \pi \)-algorithm) proposed in [10, 11] is used. This algorithm has a property of locality: the analysis error ensemble is calculated at each grid node independently of the others. The algorithm is efficient: the arithmetic operations are performed with matrices whose dimensions are equal to those of the ensemble.

The ensemble \( \pi \)-algorithm is a stochastic filter in which the analysis step is performed only for the ensemble mean. In the ensemble of analysis errors, \( D \) is an \( (L \times N) \) matrix whose columns are vectors \( \{dx'_k, n = 1, \ldots, N\} \), \( \{k = 1, \ldots, L\} \) is a grid node number obtained by transforming the ensemble of forecast errors, \( B \) is a matrix with columns \( \{b'_k, n = 1, \ldots, N\} \): \( b'_k = x^{f,a}_k - x^{e,a}_k \). \( D^T = (I + \Pi^T)^{-1} B^T \), and
\[ \Pi^T = (C + 0.25 I)^2 - 0.5 I, \]
\[ C = \frac{1}{N-1} F^T H^T R^{-1} (H B + E) = C_1 + C_2. \]

\( E \) is a matrix whose columns are equal to a vector \( \xi^a_k \), the ensemble of observation errors, and \( I \) is a unit matrix. All details are presented in [10] and [11].
The elements of the matrix $\Pi$ are calculated for the matrices $H$ and $R$ from the ensemble of values \{dx^i_n, n=1,\ldots,N\} independently of the grid nodes. Therefore, the algorithm can be used locally, and the operations can be performed with matrices the size of the ensemble. Notice that the above properties are valid for the highly popular LETKF algorithm [9], which is a deterministic filter.

6.2. Lorenz-96 model
The Lorenz-96 model [12] is used for the numerical experiments. The equations of the model are
\[
\frac{dx}{dt} = (x_{j+1} - x_j)x_{j-1} - x_j + F_0, \quad j = 1, \ldots, J
\]
where $x_1, \ldots, x_J$ ($J=40$) are the variables being forecasted. This is one of the simplest systems that simulate the properties of many atmospheric models. To solve the equations, a fourth-order finite-difference Runge-Kutta scheme is used. The time step $\Delta t = 0.05$, which corresponds to 6 hours ($t=1$ is taken for five days), $F_0 = 8$. To simulate ‘true values’ in the numerical data assimilation experiments, $x^0_i \equiv N(F_0/4; F_0/2)$ is taken to be an initial value, as in [12], and $N_i$ time steps are made. In the numerical experiments, the results of a forecast for $N_i=1000$ time steps are taken as initial data for the ‘truth’. To obtain initial data for forecasting $\hat{x}_n$ by using the model, a perturbation is added to the ‘true’ initial data $x_d(0) = x_{i}(0) + \delta$, $\delta \equiv N(0, s_0)$.

6.3. Numerical experiments
The following parameters are specified for the numerical experiments: an ensemble of initial fields $x^n(0) = x_{i}(0) + \delta^n$, $\delta^n \equiv N(0, s_0)$, $n=1, \ldots, N$; observations $y_o = x_{i}(0) + \delta_o$, $\delta_o \equiv N(0, e_0)$; and an ensemble of observations with perturbations $y^n_o = y_o + \delta^n_o$, $\delta^n_o \equiv N(0, e_0)$, $n=1, \ldots, N$. In the experiments it is assumed that the model noise $\eta^i = 0$, and in simulating the ‘truth’ $\eta = 0.01$. The observations are available at each of the $J=40$ model grid points. The experimental period has length $N_i=3000$ time steps, with assimilation being done at each time step or at every four time steps. The following values were used in the experiments: $s_0 = e_0 = 1$, $N = 20$.

In all of the numerical experiments $R = e_0 I$. The observational data from the interval $(l-id,l+id)$ are taken for the analysis at a grid node $l$. In this case, the matrix $\hat{R} = R \cdot e^{0.5(s_0/bc)^2}$ is used instead of $R$, where $\rho_n$ is the distance between the grid node and the observation, and "\cdot" denotes element-by-element multiplication. In the experiments we took $id = 5$, $bc = 5\Delta x$ ($\Delta x$ is the grid spacing).

The numerical experiments were performed for ten versions of the ‘truth’, and all estimates were calculated as average values over these ten versions. The following estimates were considered:
\[
\text{rms} = \frac{1}{K} \sum_{m=1}^{K} \left\{ \frac{1}{L} \sum_{i=1}^{L} (x_{m,i}^{n} - x_{m,i}^{0})^2 \right\}^{1/2},
\]
the root-mean-square error averaged over $K=10$ versions of calculations ($n$ is the number of a version and $i$ is the number of a grid node), with random values of the initial data errors, observational data errors, and the ensemble of perturbations specified at the initial time in each version. If the ‘truth’ is considered to be a random quantity, the $\text{rms}$ corresponds to the trace of the error covariance matrix (the covariance is estimated from a sample of ten values).
\[
\text{sp} = \frac{1}{K} \sum_{m=1}^{K} \left\{ \frac{1}{L(N-1)} \sum_{i=1}^{L} \sum_{n=1}^{N} (x_{m,i}^{n} - x_{m,i}^{0})^2 \right\}^{1/2}.
\]
This is the mean value of the covariance matrix trace calculated over $K=10$ versions of calculations ($n$ is the number of an ensemble member).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Root-mean-square ($\text{rms}$) and covariance matrix trace ($\text{sp}$) calculated in experiments 1 – 2 in the first series of experiments for the ensemble $\pi$-algorithm.}
\end{figure}

The numerical experiments with the ensemble $\pi$-algorithm described above and the LETKF algorithm were performed under the same conditions. The ensemble spread was modified after the forecast step. Two series of experiments were performed. In the first series observational data were simulated in one time step interval, and in the second series, in four time step intervals. The following experiments were performed in each of the series:

\textit{Experiment 1}. In this experiment multiplicative inflation $\alpha$ was used ($\alpha = 1.1$ in the first series and $\alpha = 1.2$ in the second series).

\textit{Experiment 2}. In this experiment additive inflation was used so that the change in the covariance matrix coincided with the change made in experiment 1. In this case formula (9) from Section 4 was used.

Figures 1-3 show the time behavior of $\text{rms}$ and $\text{sp}$ obtained in the ensemble $\pi$-algorithm (Figures 1 and 3) and the LETKF algorithm (Figure 2).

Figures 1-2 present the results of the first series of experiments for the ensemble $\pi$-algorithm (Figure 1) and the LETKF algorithm (Figure 2).
Figure 2. Root-mean-square error (rms) and covariance matrix trace (sp) calculated in experiments 1 - 2 in the first series of experiments for the LETKF algorithm.

Figure 3. Root-mean-square error (rms) and covariance matrix trace (sp) calculated in experiments 1 - 2 in the second series of experiments for the ensemble π-algorithm.
Figure 3 presents the results of the second series of experiments for the ensemble π-algorithm.

The goal of the experiments was to make a qualitative comparison of the rms and sp time behavior depending on the method being used (deterministic or stochastic ensemble filters) and specifications of the inflation parameter and assimilation frequency. One can see from Figures 1-2 that in the ensemble π-algorithm the rms is closer to the sp than in the LETKF algorithm. In the second series of experiments (Figure 3) the additive inflation in the stochastic filter (ensemble π-algorithm) improves rms as compared to multiplicative inflation.

7. Conclusions

A Bayesian approach to data assimilation based on ensembles was discussed. A brief overview of the existing algorithms was presented.

The major factors affecting the behavior of ensemble spread in the ensemble Kalman filter were considered. A general formula for ensemble spread in stochastic and deterministic ensemble Kalman filters has been derived. The impact of some methods that improve convergence on the behavior of the ensemble Kalman filter over time was investigated. The results of this study show that ensemble spread in stochastic filters rather than in deterministic filters is closer to the theoretical estimation error. Multiplicative inflation and additive inflation change the general formula for ensemble spread. The formula for ensemble spread in the stochastic filter with additive inflation rather than with multiplicative inflation is closer to the estimation error.

References

[1] Carrassi A, Bocquet M, Bertino L, Evensen G Data assimilation in the geosciences: An overview of methods, issuers and perspectives 2018 Wiley interdisciplinary reviews: Climate Change 131 5 e535
[2] Evensen G 2009 Data assimilation. The ensemble Kalman filter (Berlin Heidelberg: Spriger-Verlag)
[3] Houtekamer H L and Zhang F 2016 Review of the ensemble Kalman filter for atmospheric data assimilation Mon. Wea. Rev. 144 4489-532
[4] Murphy J M 1988 The impact of ensemble forecasts on predictability Quarterly Journal of the Royal Meteorological Society 114 463-93
[5] Nakamura G, Potthast R 2015 Inverse Modeling (IOP Publishing)
[6] Lei J, Bickel P, Snyder C 2010 Comparison of ensemble Kalman filters under non-Gaussianity. Monthly Weather Review 138 1293-306
[7] Jazwinski A H 1970 Stochastic Processes and Filtering Theory (New York: Academic Press)
[8] Cohn S E, Dee D P 1988 Observability of discretized partial differential equations SIAM J. Numer. Anal. 25 3 586-617
[9] Hunt B R, Kostelich E J, Szunyogh I 2007 Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter Physica D 230 112-26
[10] Klimova E 2012 A suboptimal data assimilation algorithm based on the ensemble Kalman filter Quarterly Journal of the Royal Meteorological Society 138 2079-85
[11] Klimova E G 2019 The Kalman stochastic ensemble filter with transformation of perturbation ensemble Siberian. J. Num. Math. 22 1 27-40
[12] Lorenz E N, Emanuel K A 1998 Optimal sites for supplementary weather observations: simulation with a small model Journal of the Atmospheric Sciences 55 399-414