Data assimilation algorithm based on an approximate solution of the optimal nonlinear filtering problem

E G Klimova¹

¹Federal Research Center for Information and Computational Technologies, Novosibirsk, Lavrentyev Ave. 6, Russia, 630090
E-mail: klimova@ict.nsc.ru

Abstract. In the ensemble Kalman filter, a currently popular data assimilation algorithm, it is assumed that the density of the estimation error distribution is a Gauss function, which is not fulfilled in the general case of a nonlinear model and a nonlinear observation operator. Statement of the optimal filtering problem in the general case is based on the Bayesian approach. One way to approximate the nonlinear optimal filtration problem is to represent the distribution density as a sum of Gaussian distributions with given mean values and covariance matrices. Then the optimal estimate is the weighted sum of estimates obtained in the Kalman filter for the corresponding parameters of the Gaussian distribution in this sum. The paper considers the algorithm for implementing this approach, which uses the previously developed effective local ensemble data assimilation algorithm (ensemble pi-algorithm). The results of numerical experiments to evaluate the properties of the proposed algorithm with the 1-dimensional nonlinear model are presented.

1. Introduction

The data assimilation problem is the optimal estimation of geophysical fields based on observations and mathematical models. Algorithms for data assimilation use the different approaches depending on the specific problem statement [1]. All of data assimilation methods can be unified through Bayesian theorem. An ensemble of forecasts and observations can be used to approximate the covariance matrices considered in the data assimilation algorithms [2]. If the optimal estimate is searched for the end of a given time interval, the problem is reduced to the ensemble Kalman filter. If the optimal estimation is considered for a given time interval, such a problem is called the ensemble smoothing problem [2]. If the considered random variables are Gaussian, and the forecast and observation models are linear, Kalman filter is equivalent to the variational approach to the data assimilation problem (4DVAR) [1]. In the strictly nonlinear non-Gaussian case, the particle filter is considered, which is based on the Bayesian approach. This method also can use ensembles [3, 4].

The article is devoted to solving the nonlinear filtration problem based on the approximation of the forecast density distribution by mixtures of Gaussian distributions. To implement the algorithm, the
ensemble pi-algorithm proposed in [5, 6, 7] is used. The results of numerical experiments with 1-dimensional nonlinear model are presented.

2. The problem

Let us assume that the forecast model describes the change in time of the estimated value $x^k$

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

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

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

Where $\eta^k$, $\varepsilon^k$ - random errors of the forecast and observations, respectively, $f_{k+1|x}$ - the forecast model operator, $h_k$ - the operator of transformation of the predicted variable into the observed one. $\varepsilon^k$ and $\eta^k$ - Gaussian random variables: $E[\varepsilon_k (\varepsilon_k)^T] = R_x, E[\eta_k (\eta_k)^T] = Q_x$.

In the nonlinear case it is customary to apply Bayes's theorem to obtain an optimal estimate based on observations and model forecast of the state:

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

There are various options for assessing the state of the data and the model forecast $p(x_k | y_{k-1}), k > l$ - forecast, $p(x_k | y_{k-1})$ - filtration, $p(x_{k-0} | y_{k-1})$ - smoothing, where $x_{k-0} = \{x_k, x_{k-1}, \ldots, x_0\}$, $y_{k-1} = \{y_1, \ldots, y_l\}$. Notation and definition taken from the review [1].

In the linear Gaussian case, the solution to the filtration problem is the Kalman filter, and the solution to the smoothing problem is the Kalman smoothing [4]. In [2] a Monte Carlo-based approach to forecasting and data assimilation is proposed. The Monte Carlo method is used 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 calculated using a nonlinear model, and the Gaussian condition is violated. In addition, in this case the estimate at the analysis step approximates the minimum variance estimate (linear variation minimizing). In the nonlinear non-Gaussian case, the particle filter is used, which is based on the Bayesian theorem [3, 4].

3. The algorithm of optimal nonlinear filtration based on the representation of the density distribution as a mixtures of Gaussian distributions

It is known that the density distribution can be approximated using the mixtures of Gaussian distributions [8, 9]. Let

$$p(x) = \sum_{i=1}^{L} \alpha_i g(m_i, P_i),$$

where $\sum_{i=1}^{L} \alpha_i = 1$, $g$ - Gauss function, $m_i$ - n-vectors, $P_i$ - positive definite matrices. It is known from the theory that in this case the mean value and covariance matrix of this random process will be equal, respectively [8]

$$\hat{x} = \sum_{i=1}^{L} \alpha_i m_i, \tag{1}$$

$$\hat{P} = \sum_{i=1}^{L} \alpha_i [P_i + (\hat{x} - m_i)(\hat{x} - m_i)^T]. \tag{2}$$

The data assimilation algorithm is usually represented as a forecast step based on the model forecast and an analysis step (estimation of values based on observations and forecast). If we assume the density, distribution at the forecast step has the form
when observations are available, the distribution density will have the form [8]:

\[ p(x | y) = \sum_{i=1}^{L} \alpha_i g(m_i^f, P_i^f), \]

where

\[ m_i^f = m_j^f + K_i [y - h(m_j^f)], \]

\[ K_i = P_i^f H_i^T (H_i^T P_i^f H_i^T + R_i)^{-1}, \]

\[ H_i^f = \frac{\partial h(x)}{\partial x} |_{x=m_j^f}, \]

\[ P_i^f = (I - K_i H_i) P_i^f, \]

\[ \alpha_i = \frac{g[y - h(x_j^f), HP_i^f H_i^T + R]}{\sum_{i=1}^{L} g[y - h(x_j^f), HP_i^f H_i^T + R]}. \]

In other words, the optimal estimate (mean value) will be equal to the sum with the weights of the values obtained in the L Kalman filter algorithms.

An algorithm based on these formulas is proposed in [10]. A simplified version of the representation of the distribution density is considered. In this variant for a given sample of the N, independent random variables with a density distribution \( p \) the density distribution estimate can be obtained using the representation as the mixtures of \( N \) Gaussian distributions:

\[ \hat{\rho}(x) = \frac{1}{N} \sum_{i=1}^{N} g(x_i, P), \]

where \( P \) is a positive definite matrix. The matrix \( P \) is approximated using a sample (ensemble) \( \{x_i, i = 1, \ldots, N\} [10]: \hat{P} = h'\hat{P}, \) where \( \hat{P} = \frac{1}{N-1} XX^T, X = \{x_1 - \bar{x}, \ldots, x_N - \bar{x}\}, \) \( \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i, \) \( h \) - parameter.

If the density distribution considered at the forecast step is given by (8), then at the correction step based on observations \( y \) (analysis step), the density distribution will be given by (4)-(7), where \( P_f = \hat{P} \) and \( m_f = x. \) Then, at the forecast step, in the algorithm proposed in [10], the ensemble of forecasts of mean values and the covariance matrix are calculated using the model.

One of the major problems of the particle filter, a variant of which is considered here, is its "divergence" over time, when one of the weight coefficients \( \alpha_i \) becomes close to one, and the others to zero. In this case, the so-called resampling is used, when a sample is generated according to the distribution density (3), and after the forecast step, the algorithm is similar to the first step, when the covariance matrix is approximated by the sample (8).

4. Implementation of the optimal nonlinear filtering algorithm based on the ensemble approach

In [12], the algorithm described above was implemented using ensembles, and the formula for representing the distribution density as a mixture of Gaussian distributions was considered. As noted in [12], the ensemble member corresponding to the density distribution at the analysis step (3) has the form

\[ x_{i,j}^e = x_{i,j}^f + K_i (y + \varepsilon_i - H_i x_{i,j}^f), \]

where \( i \) is the number of the Gauss function in the sum (3) \( (i = 1, \ldots, L), j \) is the number of the ensemble member corresponding to the \( i \)-th Gaussian distribution and the coefficient \( \alpha_i, \varepsilon_i \) - a Gaussian random.
variable: $E[e_i (e_i)^T] = R_i$. According to the algorithm proposed in [11], the number of random variables corresponding to the $i$-th Gaussian distribution is proportional to the coefficient $\alpha_i$.

In [5, 6, 7] a stochastic ensemble Kalman filter is proposed, which is implemented locally (ensemble $\pi$-algorithm), as well as the deterministic LETKF algorithm [13]. In this algorithm, the ensemble member has the form (9).

Let us consider the applicability of the ensemble $\pi$-algorithm for implementing the described above algorithm to the approximate solution of the nonlinear filtration problem. Let the initial sample corresponding to the covariance matrix be set at time $t=0$: $\frac{1}{1+h^2} P_{ij}$ [10]. The method based on the $\pi$-algorithm consists of several steps.

Step 1. After the forecast step, we consider the density distribution is given by $p(x) = \frac{1}{N} \sum_{i=1}^{N} g(x'_i, P_{ij})$, where $P_f = h^2 \hat{P}$, $\hat{P}$ is the sample covariance matrix of the ensemble of forecasts.

Step 2. According to the theory the density distribution at the analysis step is given by

$$\hat{p}(x) = \sum_{i=1}^{N} \alpha_i g(x'_i, P_a)$$

where $x'_i = x'_i + K(y - Hx'_i)$, $K = P_a H^T R^{-1}$, $P_a = (I - KH)P_f$ is the analysis for the mean values of Gaussian distribution.

To perform the forecast step, we need to set a sample corresponding to the density distribution (10). In [11], it is proposed to generate the number of sample elements “$i$” ($i$ is the number of the Gaussian distribution in the sum (10)) in accordance with the weight coefficient $\alpha_i$. Then the $j$-th member of the ensemble “$i$” has the form (9). The deviation from the $i$-th mean will be equal to

$$dx'_i = dx'_j + K(e_j - Hx'_j).$$

We assume that the covariance matrix of analysis errors is $P_a = \frac{1}{N-1} D_x D_x^T$, where $D_x = \{dx'_1, \ldots, dx'_N\}$. The equation for $dx'_i$ is similar to the equation of the ensemble $\pi$-algorithm [5, 6, 7]. Therefore, you can calculate $dx'_j$: $dx'_j = (I + \Pi^T) dx'_j$. Detailed calculations are given in [5, 6, 7].

Thus, the analysis step consists of the following procedures:

1) the ensemble of perturbations $dx'_i$ is calculated from the ensemble of perturbations $dx'_j$;

2) knowing the matrix $P_a$, we calculate $N$ mean analyses for the corresponding Gaussian distribution $\{x'_i, i = 1, \ldots, N\}$;

3) the sample $x'_{i,j}$ is calculated according to the formula for the density distribution (10).

Next, the forecast of the ensemble members is performed, and the first step of the procedure is repeated. The locality of the algorithm is determined by the properties of the $\pi$-algorithm: the analysis step can be performed for all grid nodes independently. All arithmetic operations are performed with matrices whose dimension is equal to the number of the ensemble members.

5. Numerical experiments with the Lorenz-96 model

5.1 Lorenz-96 Model

The Lorenz-96 model [12] is used for the numerical experiments. The equations of the model are
\[ \frac{dx_j}{dt} = (x_{j+1} - x_{j-2})x_{j-1} - x_j + F_0, \quad j = 1, \ldots, J \]
\[ x_1 = x_{J+1}, \quad x_{j+1} = x_j, \]

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, \( x'_0 \approx N(F_0/4; F_0/2) \) is taken to be an initial value, as in [14], and \( N_t \) time steps are made. In the numerical experiments, the results of a forecast for \( N_t = 1000 \) time steps are taken as initial data for the ‘truth’. To obtain initial data for forecasting of \( \hat{x}_n \) by using the model, a perturbation is added to the ‘true’ initial data \( x_d(0) = x_i(0) + \delta, \quad \delta \sim N(0, s_0) \)

### 5.2 Numerical experiments

To organize numerical experiments, the following parameters were set: ensemble of initial fields \( x^0 = x_i(0) + \delta^0, \delta^0 \sim N(0, s_0), n=1, \ldots, N_{ens} \); observations \( y_0 = x_i(0) + \delta_0, \delta_0 \sim N(0, e_0) \); ensemble of observations with perturbations \( y'_0 = y_0 + \delta'_0, \delta'_0 \sim N(0, e_0), n=1, \ldots, N_{out} \). In all experiments, “noise model” \( \eta^0 = 0 \), in the simulation “truth” \( \eta^0 = 0.01 \). The number of the ensemble members is indicated by \( N_{ens} \). The observations are assumed to be known in the whole region of integration. The forecast was made within \( N_t = 3000 \) steps in time, the assimilation was carried out in four steps in time. Numerical experiments were performed for values \( s_0 = e_0 = 1 \), \( N_{ens} = 50 \).

In all numerical experiments, \( R = \delta^2 I \) was considered. In the analysis at the grid node \( l \), observations from the interval \((l - id, l + id)\) were taken. In this case, when analyzing the grid node \( l \), instead of the matrix \( R \), a matrix \( \hat{R} = R \circ e^{-0.5(\rho_\beta/\delta)^2} \) was taken, where \( \rho_\beta \) - the distance between the grid node and the observation, "\( \circ \)" - the sign of element multiplication. In the numerical experiments, \( id = 5, \beta c = 5 \Delta x \) values were taken \( (\Delta x \text{ is the grid spacing}) \).

To prevent divergence of the proposed algorithm, which is a variant of the particle filter, we considered the modification of the weight coefficients proposed in [9]:
\[ \hat{\alpha} = \alpha_l \times w + (1 - w) \times N^{-1}, \quad (11) \]

\( w \) is a parameter selected empirically, and \( N \) is the number of ensemble members.

The following numerical experiments were performed. The parameter \( h \) in the estimation formula for \( P_\gamma \) was set to 0.7. The weight coefficients were modified according to the formula (11). The parameter \( w \) was set to 0.5.

Calculations were performed for 3000 time steps, and the root-mean-square error (compared to the "true" value) was calculated for time steps from 2001 to 3000. For comparison, the data assimilation procedure was also performed using a variant of the ensemble Kalman filter, which occurs when setting the parameters \( h = 1, w = 0 \) [9]. The results of numerical experiments are shown in figure 1. Figure 1 shows the time behavior of the RMS (root-mean-square) error for the proposed algorithm at \( h = 0.7, w = 0.5 \) (PF) and \( h = 1, w = 0 \) (KF).

As you can see from the figure, on average, the new algorithm wins compared to EnKF.
6. Conclusion
The article proposes a new data assimilation algorithm that can be applied in the nonlinear case. The algorithm is based on representation of the density distribution as a mixture of Gaussian distributions and application of a local stochastic ensemble Kalman filter with transformation of the covariance matrix. Numerical experiments have been performed in which the algorithm is tested. Experiments have been implemented for a nonlinear one-dimensional model.

References
[1] Carrassi A, Bocquet M, Bertino L and 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] Nakamura G and Potthast R. 2015 *Inverse Modeling* (IOP Publishing)
[4] van Leeuwen P J, Kunsch H R, Nerger L, Potthast R and Reich S Particle filters for high-dimensional geoscience applications 2019 *Quarterly Journal of the Royal Meteorological Society* **145** 2335-2365
[5] Klimova E 2012 A suboptimal data assimilation algorithm based on the ensemble Kalman filter *Quarterly Journal of the Royal Meteorological Society* **138** 2079-2085
[6] Klimova E G 2019 The Kalman stochastic ensemble filter with transformation of perturbation ensemble *Siberian. J. Num. Math.* **22** 1 27-40
[7] Klimova E G 2019 Bayesian approach to data assimilation based on ensembles of forecasts and observations *IOP Conf. Series: Earth and Environmental Science* **386** 012038
[8] Anderson B D and Moore J B 1979 *Optimal filtering* (Prentice-Hall, Inc., Englewood Cliffs, New Jersey)
[9] Stordal A S, Karlsen H A, Nævdal G, Skaug H J and Valles B 2011 Bridging the ensemble Kalman filter and particle filter: the adaptive Gaussian mixture filter *Comput Geosci* **15** 293-305
[10] Hoteit I, Pham D -T, Triantafyllou G and Korres G 2008 A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography *Monthly Weather Review* **138** 317-334
[11] Carpenter J, Clifford P and Fearnhead P 1999 Improved particle filter for nonlinear problems *IEE Proc.- Radar, Sonar navig.* **146** 1-7
[12] Bengtsson T, Snyder C and Nychka D 2003 Toward a nonlinear ensemble filter for high-dimensional systems *Journal of Geophysical Research* **108** D24 8775

[13] Houtekamer H L and Zhang F 2016 Review of the ensemble Kalman filter for atmospheric data assimilation *Mon. Wea. Rev.* **144** 4489-532

[14] Lorenz E N and Emanuel K A 1998 Optimal sites for supplementary weather observations: simulation with a small model *Journal of the Atmospheric Sciences* **55** 399-414