MATHEMATICAL FOUNDATION, APPLICATION, AND COMPARISON OF GENERAL DATA ASSIMILATION METHOD BASED ON DIFFUSION APPROXIMATION WITH OTHER DATA ASSIMILATION SCHEMES

K. P. Belyaev[1], C. A. S. Tanajura[2], and N. P. Tuchkova[3]

Аннотация: Data assimilation methods commonly used in numerical ocean and atmospheric circulation models for weather and climate prediction produce approximations of state variables in terms of stochastic processes. This approximation consists of random sequences of Markov chains, which converge to a diffusion-type process. The conditions for this convergence are investigated. The optimization problem associated with the search of the best possible approximation of the state variable and the results of a numerical experiment are discussed. It is shown that the data assimilation method can be used in practical applications in meteorology and oceanography. Several applications of the methods as an example of the modern operational data processing system with the ocean circulation model HYCOM and data from ARGO drifters are performed and the results as well as comparisons with other assimilation schemes are presented.

Ключевые слова: sequence of Markov chains; diffusion stochastic process; data assimilation methods; HYCOM; ARGO drifters

1 Introduction

In numerical modeling of geophysical systems, such as an ocean, an atmosphere, or a climate system, the data assimilation is a common and popular approach. It produces the initial conditions for weather and climate models and complements monitoring by correcting model variables in the direction of observations. Therefore, data assimilation methods have a substantial impact in weather forecast modeling and, consequently, in several human activities that rely on weather and climate forecasts, such as agriculture, water resources, and others.

Actually, the observed data such as the sea level, the ocean temperature, the salinity, and other tracers are collected from different sources including satellite measurements, merchant ships reports, specially designed ocean moorings and drifters, scientific expeditions reports, and so on. The processing of such enormous number of observations in the on-line regime requires to develop both the technology of data transfer and methods of their analysis. It implies a further advance in information technologies and their applications, parallel computations, mathematical methods of data analysis, etc.

In practice, the data inflow from all sources are subject to the data quality control which filters out poorly known and/or wrong observations, their preprocess analysis, further transfer to computers, and then their parallel calculations. This complex information system requires reliable and fast data assimilation methods in conjunction with circulation models.

The correction of the model output by observational data is generally based on a scheme of the following type: a system of partial differential equations usually represented in the finite-difference or finite-element form is considered within a time interval (t_0, T). In general, t_0 can be associated with 0, while T can be considered as infinity provided it is large enough. The interval (t_0, T) can be divided into intervals $(t_0, t_1), (t_1, t_2), \ldots, (t_n, t_n + 1)$. Within any time interval $(t_n, t_n + 1)$, the model starts at the moment t_n with the initial state-vector Θ_n, integrates forward until the moment t_{n+1}, when it produces (predicts) the state-vector Θ^m_{n+1}. Here and further, the superscript m indicates that the system state has been obtained only by the model integration without any other source of information. During the time interval $(t_n, t_n + 1)$, a series of observations enters, represented by the vectors $\xi^n_1, \ldots, \xi^n_k$ independent of the model, where the index k denotes the vector within the nth time

*Работа выполнена при частичной поддержке РФФИ (гранты 10-07-00487-а и 12-07-00108) и Рособразования (проект 020619-1-174).

[1] Shirshov Institute of Oceanology, Russian Academy of Science, Moscow, Russia, kb@sail.msk.ru

[2] Federal University of Bahia, Salvador, Brazil, cast@ufba.br

[3] Dorodnicyn Computer Centre, Russian Academy of Science, Moscow, Russia, tuchkova@ccas.ru

interval. Usually, the vector ξ_k^n is a subset of the state vector Θ_n, since only a part of the model state can be observed. Then, the model output Θ_{n+1}^m is corrected by observations according to

$$\Theta_{n+1} = \Theta_{n+1}^m + \sum_{n=1}^{k} \alpha_i(\xi_i^n - \xi_i^m)\,.$$

Here, ξ_i^m denotes the model variables, corresponding to observations, calculated at the observation time frame. The weight functions α_i also depend on time and may be either known *a priori* or determined by some appropriate algorithm. The corrected state vector Θ_{n+1}, the so-called objective analysis, is taken as the new initial condition, and the model integration resumes.

Under the name of data assimilation methods, various versions of this scheme are commonly used in geophysics. In particular, the Kalman filter approach determines the optimal weight-coefficient utilizing the statistical properties of observational data [1, 2]. Alternatively, the observations may be not considered as random. In this case, the optimal weights can be determined according to the variational or adjoint data assimilation technique [3]. The integration intervals can be considered as given, e. g., the model produces forecasts every 24 h or at random. For instance, in modern coupled ocean-atmosphere models the correction may be applied as soon as the model temperature difference between the ocean and the atmosphere exceeds a certain a priori chosen limit.

Despite these and other differences, all these methods in essence follow the scheme outlined in the above. Alternatively, it could be interesting to review these techniques from a different point of view. At the “moment of assimilations,” the time series of variables Θ_n undergoes a jump of its trajectories. What will happen if the interval between consecutive assimilations approaches to zero along with the values of the jump? How does the limiting behavior of the trajectories depend on the number of state variables and their distributions? Under which conditions does the limiting distribution of Θ_n exist as T goes to infinity and what is it? If it exists, it is called a stationary distribution. These and similar questions attract interest not only from the theoretical point of view, but also for quite practical reasons. For instance, knowing the limiting behavior of the time series for $\Delta t_k = t_{k+1} - t_k \to 0$, it becomes easy to calculate various parameters needed for the weather forecast, while the knowledge of stationary distribution enhances the reliability of a climate prediction. In addition, the knowledge of this limit can simplify the optimization problem for weight coefficients, which generally are the extremum of a given function, e. g., the error variance.

One of the aims of the present paper is to prove that under appropriate conditions, the trajectories of the objective analysis Θ_n as a function of time converge to the trajectories of the stochastic diffusion process. These are continuous functions satisfying the Fokker–Planck (FP) equation. Their characteristics provide a tool for the determination of properties of the limiting trajectories, such as their maxima. Furthermore, the optimization problem of the best weight coefficients satisfying the unbiased and minimum variance estimator is solved. Finally, to illustrate the feasibility and usefulness of this method, several numerical experiments are performed. The main idea of this study is based on the classical theorem of convergence of Markov chains to diffusion process [4], and its recent generalizations [5]. It is a continuation of the works by the authors [6] and it generalizes some results previously obtained in [7].

2 Main Definitions and Notations

Let the system of equations

$$\frac{\partial\Theta(t)}{\partial t} = \Lambda(\Theta, t) \qquad (1)$$

be considered on the time interval (t_0, T). Without loss of generality, t_0 in further references will be associated with 0 while T may be both finite and infinite. In (1), $\Theta(t)$ represents the random state-vector of dimension r defined on a given probability space, $\Lambda(x, t)$ denotes nonrandom generally nonlinear operator acting in R^r, which does not explicitly involve temporal derivatives, and η denotes the random variable with zero average and finite covariance function. Symbol { ’ } denotes a transpose of a corresponding vector and/or matrix, the symbols | | or || || represent a vector or a matrix norm, respectively. A sequence of time series is considered such that in each series the interval $(0, T)$ is divided by time points $(0 = t_{1,n}, t_{2,n}, \ldots, t_{k,n}, \ldots)$, where the first index denotes the order number of the corresponding point while the second index refers to the time series.

It is supposed that in each series within the interval $\Delta t_{k,n} = t_{k+1,n} - t_{k,n}$, a number of random vectors $(\xi_1^n, \ldots, \xi_l^n)$, $l = \overline{1, \nu^n}$, with dimension $q, q \leq r$, are observed, where ν^n is also a random integer variable with given distribution $p_l^n = P(\nu^n = l)$ independent of the vectors ξ_l^n. Knowing the solution of the system (1), $\Theta_k^{n,m}(t)$ for initial vector $\Theta_k^n(t)$ within entire interval $\Delta t_{k,n}$, the observed variables ξ_l^n, as well as a realization of the random index ν^n, the newly constrained variables are introduced by the formula:

$$\zeta_k^n = \sum_{l=0}^{\nu^n} \alpha_{l,n}(\xi_l^n - \Theta_l^{n,m})\,. \qquad (2)$$

In (2), the matrices $\alpha_{l,n}$ with dimension $r \times r$ referred to as weight coefficients are supposed to be known and depend on the time series.

Remark. The matrices $\alpha_{l,n}$ in (2) are supposed to be known, but arbitrary enough. Their specific determination with respect to some criteria is a matter of another problem.

For the consistency of (2), it is necessary that in (2) the vectors ξ_l^n and $\Theta_l^{n,m}(t)$ have the same dimension q which may differ from the dimension of the vector $\Theta(t)$ of the dimension r. However, without loss of a generality, it will be assumed that ζ_k^n has the same dimension as Θ_k^n setting the "dummy" components of vectors ξ_l^n and $\Theta_l^{n,m}(t)$ to zero. Ultimately, the new state variables Θ_n^{k+1} are defined by formula

$$\Theta_n^{k+1} = \Theta_n^k + \int\limits_{t_k}^{t_{k+1}} \Lambda(\Theta_n^k, \tau)\, d\tau + \zeta_k^n \qquad (3)$$

and Θ_n^{k+1} is taken as the initial conditions in (1) for the continuation of the integrations.

In this manner, it becomes possible to obtain the sequence of trajectories $\Theta^n(t)$ defined over the entire interval $(0, T)$. Starting from some known random vector $\Theta_0^n(t)$, the solution of (1) for each interval $\Delta t_{k,n} = t_{k+1,n} - t_{k,n}$ with breaks at moments $t_{k,n}$ can be evaluated. The goal of the present paper is to determine the limiting behavior of the solution (3) when $n \to \infty$.

3 Formulations

The following notions are introduced below.

A0. The time lattice $\Delta t_{k,n} = t_{k+1,n} - t_{k,n}$ is considered as nonequidistant, with real values in each series.

A1. The intervals $\Delta t_{k,n} \to 0$ approach zero uniformly with respect to k, i. e., $\max\limits_k \Delta t_{k,n} \to 0$, $n \to \infty$.

A2. The probability distribution for random variables ν^n satisfies the conditions: $p_{l,n} = P(\nu^n = l) = p_l \Delta t_{k,n} + o(t_{k,n})$, $l > 0$, for any k, and $\mu = \sum\limits_{l=1}^{\infty} l p_l < \infty$.

A3. Random vectors $\xi_1^n, \dots \xi_l^n, \dots$ have $2 + \delta$ moments for positive δ for each n, i. e., $E\xi_i^n = \lambda_{i1}^n$, $E(\xi_i^n \xi_j'^n) = \gamma_{ij}^n$, $E|\xi_i^n|^{2+\delta} < \infty$, and these variables are uniformly bounded with respect to n, i. e., $\overline{\lim}|\lambda_{i1}^n| < \infty$, $\overline{\lim}||\gamma_{ij}^n|| < \infty$, $\gamma_{i2}^n = \gamma_{ii}^n$, $i, j = 1, 2, \dots$

A4. The operator $\Lambda(x, t)$ is a continuous function of its arguments.

A5. The set of weight coefficients $\alpha_{l,n}$ is uniformly bounded with respect to n, i. e., $\overline{\lim}||\alpha_{l,n}|| < \infty$, $l = 1, 2, \dots$

A6. The sequence of distributions of random variables θ_0^n converges to the distribution of some random variable θ_0, i. e., $P(\theta_0^n < x) \to P(\theta_0 < x)$, $n \to \infty$, for each x.

Without loss of generality, the limit values of variables $\lambda_{i1}^n, \gamma_{ij}^n$, and $\alpha_{l,n}$ when n tends to infinity are supposed to exist and to be equal to $\lambda_{i1}, \gamma_{ij}$, and α_l, respectively. Otherwise, the corresponding subsequence can be chosen to provide the convergence to the limit points, which exist as a consequence of the conditions A3 and A5.

Theorem 1. *Let the conditions* A0–A6 *hold. Then, the sequence of finite-dimensional distributions of random processes $\theta^n(t)$ converges to the stochastic process, which will be a solution of the stochastic differential equation*

$$\theta(t) = \theta_0 + \int\limits_0^t a(s, \theta(s))\, ds + \int\limits_0^t b(s, \theta(s))\, dw(s)$$

where

$$a(t, x) = \sum_{l=1}^{\infty} p_l \left[\sum_{j=1}^{l} \alpha_j (\lambda_{j1} - x) \right] + \Lambda(x, t)\,; \qquad (4)$$

$$b^2(t, x) = \sum_{m=1}^{\infty} p_m \left\{ \sum_{i,j=1}^{m(m+1)} \alpha_i \left[\gamma_{ij} - (x\lambda'_{1,i} + \lambda_{1,j} x') + x x' \right] \alpha'_j \right\}. \qquad (5)$$

The Wiener process $w(t)$ is the defined on interval $(0\,, T)$ and it is independent of the random variable θ_0.

Proof. The sequence of random variables θ_k^n, $k = 0, 1, \dots$, forms the Markov chain for each series n. Hence, the general statements about the convergence of Markov chains can be applied. The proof of theorem is based on the results [4], which have been generalized in [5]. To examine the conditions of convergence given in [4, 5], the following statements have to be proved:

I There are vectors $a_k(x)$ and matrices $B_k(x)$ such that

$$\left| \frac{1}{\Delta t_{k,n}} \int\limits_{-\infty}^{\infty} (y - x)\, dp_{k,n}\left(\frac{y}{x}\right) - a_k(x) \right| \to 0\,;$$

$$\left\| \frac{1}{\Delta t_{k,n}} \int\limits_{-\infty}^{\infty} (y - x)(y - x)'\, dp_{k,n}\left(\frac{y}{x}\right) - B_k(x) \right\| \to 0$$

for each k, when $n \to \infty$.

Here and further, $p_{k,n}(y/x)$ denotes the transitional probabilities of Markov chains in each time series, i. e., $P(\theta_{k+1}^n = y/\theta_k^n = x) = p_{k,n}(y/x)$.

This equality is correct for both discrete and absolutely continuous random variables $\xi_l^n$and $\theta_k^n(t)$. In the latter case, this equality applies for the corresponding probability densities.

II The convergence of sequences

$$E(\theta_{k+1}^n - \theta_k^n)^{2+\delta} \overset{n\to\infty}{\longrightarrow} 0\,;$$
$$P\left(\frac{|\xi_k^n - \xi_{k+p}^n|^{2+\delta}}{\xi_k^n}\right) \longrightarrow 0$$

for any k, p, when $n \to \infty$, has to be shown.

III The sequence $\{|a_k(x)|/||B_k(x)||\}$ is uniformly bounded.

As in [4, 5], these conditions provide the basis of the proof of the theorem. The transitional probabilities $p_{k,n}(y/x)$ are directly calculated as

$$P(\theta_{k+1}^n = y/\theta_k^n = x) = p_n\left(\frac{y}{x}\right)$$
$$= P\left(\theta_k^n + \int\limits_{t_k}^{t_{k+1}} \Lambda(\theta_k^n, \tau)\,d\tau + \zeta_k^n = y/\theta_k^n = x\right)$$
$$= P\left(\zeta_k^n + x + \int\limits_{t_k}^{t_{k+1}} \Lambda(x,\tau)\,d\tau = y\right)$$
$$= P\left(\sum_{l=0}^{\nu^n} \alpha_{l,n}(\xi_l^n - \theta_l^{n,m}) = y - x - \int\limits_{t_k}^{t_{k+1}} \Lambda(x,\tau) d\tau\right).$$

Then, this equality may be extended to

$$p_n(y/x) = p\left(\sum_{l=0}^{\nu^n} \alpha_{l,n}\left(\xi_l^n - x - \int\limits_{t_k}^{s_l} \Lambda(x,\tau)\,d\tau\right)\right.$$
$$\left. = y - x - \int\limits_{t_k}^{t_{k+1}} \Lambda(x,\tau)\,d\tau\right) \qquad (6)$$

where $s_l, s_1 < s_2 < \cdots < s_l$, $l = 0, 1, \ldots, \nu^n$, are chronologically ordered moments within the interval (t_k, t_{k+1}). Using the continuity condition A4, the integrals in (6) can be represented as $\int\limits_{t_k}^{s_l} \Lambda(x,\tau)\,d\tau = \Lambda(x,\tau_*)(s_l - t_k)$. On account of this relation, one obtains from the conditional probabilities (6):

$$p\left(\sum_{l=0}^{\nu^n} \alpha_{l,n}(\xi_l^n - x)\right.$$
$$\left. = y - x - \Delta t_{k,n}\Lambda(x,\tau_\bullet) + \sum_{l=0}^{\nu^n}(s_l - t_{k,n})\Lambda(x,\tau_{*,k})\right)$$

where $\tau_* \in (t_k, s_l)$. Using this relation, the conditional average $E(y-x)p(y/x)$ may be written as

$$E(y-x)p_n(y/x) = \int\limits_{-\infty}^{\infty} (y-x)\,d_y p$$
$$\times \left(\sum_{l=0}^{\nu^n} \alpha_{l,n}(\xi_l^n - x) = y - x - \Delta t_{k,n}\Lambda(x,\tau_\bullet)\right.$$
$$\left. + \sum_{l=0}^{\nu^n}(s_l - t_{k,n})\Lambda(x,\tau_{*,k})\right)$$
$$= \int\limits_{-\infty}^{\infty}\left(z + \Delta t_{k,n}\Lambda(x,\tau_\bullet) - \sum_{l=0}^{\nu^n}(s_l - t_{k,n})\Lambda(x,\tau_{*,k})\right)$$
$$\times\, d_z p\left(\sum_{l=0}^{\nu^n}\alpha_{l,n}(\xi_l^n - x) = z\right)$$
$$= \int\limits_{-\infty}^{\infty} z\,d_z p\left(\sum_{l=0}^{\nu^n}\alpha_{l,n}(\xi_l^n - x) = z\right) + \Delta t_{k,n}\Lambda(x,\tau_\bullet)$$
$$- \sum_{l=0}^{\nu^n}(s_l - t_{k,n})\Lambda(x,\tau_{*,k}) \qquad (7)$$

Since the last two terms are independent of z, they can be taken out of the integral. The first term on the right-hand side of Eq. (7) is the expectation value of a series of random terms. It may be expanded as

$$E\left(\sum_{l=0}^{\nu^n}\alpha_{l,n}(\xi_l^n - x)\right) = \sum_{m=0}^{\infty} p_{m,n}\sum_{l=0}^{m}\alpha_{l,n}(\xi_l^n - x)$$
$$= \sum_{m=1}^{\infty} p_{m,n}\sum_{l=1}^{m}\alpha_{l,n}(\xi_l^n - x)\,,$$

because the term with $m = 0$ vanishes. Inserting this expression into Eq. (7), the equality is obtained:

$$E(y-x)p_n(y/x)$$
$$= \sum_{m=1}^{\infty} p_{m,n}\sum_{l=1}^{m}\alpha_{l,n}(\xi_l^n - x) + \Delta t_{k,n}\Lambda(x,\tau_\bullet)$$
$$- \sum_{m=1}^{\infty} p_{m,n}\sum_{l=1}^{m}(s_l - t_{k,n})\Lambda(x,\tau_{*,k})\,. \qquad (8)$$

Similarly, for conditional variance $E(y-x)(y-x)'p_n(y/x)$, one arrives at the relation (omitting intermediate calculations):

$$E(y-x)(y-x)'p_n(y/x) = \sum_{m=1}^{\infty} p_{m,n}$$
$$\times\left\{\sum_{i,j=1}^{m(m+1)} \alpha_{i,n}\left[\gamma_{ij}^n - x\lambda_{1,i}^{n'} - \lambda_{1,j}^n x' + xx'\right]\alpha'_{j,n}\right\}$$

$$+2\left(\Delta t_{k,n}\Lambda(x,\tau_\bullet)-\sum_{m=1}^{\infty}p_{m,l}\sum_{l=1}^{m}(s_l-t_{k,n})\Lambda(x,\tau_{*,k})\right)\times E(y-x)p_n(y/x)+\left(\Delta t_{k,n}\Lambda(x,\tau_\bullet)-\sum_{m=1}^{\infty}p_{m,l}\sum_{l=1}^{m}(s_l-t_{k,n})\Lambda(x,\tau_{*,k})\right)^2. \qquad (9)$$

Now, the last term in right-hand side of (8) can be estimated as

$$\sum_{m=1}^{\infty}p_{m,n}\sum_{l=1}^{m}(s_l-t_{k,n})\Lambda(x,\tau_{*,k}) \le \sum_{m=1}^{\infty}p_{m,n}\mathrm{max}_{t_k\le\tau\le t_{k+1}}|\Lambda(x,\tau)|m(s_m-t_k) \le \Delta t_{k,n}\mathrm{max}_{t_k\le\tau\le t_{k+1}}|\Lambda(x,\tau)|(\mu\Delta_{k,n}+o(\Delta t_{k,n}))\,.$$

Similar, in (9),

$$(\Delta t_{k,n}\Lambda(x,\tau_\bullet)-\sum_{m=1}^{\infty}p_{m,n}\sum_{l=1}^{m}(s_l-t_{k,n})\Lambda(x,\tau_{*,k}))^2 \le (\Delta t_{k,n}\mathrm{max}_{t_k\le\tau\le t_{k+1}}|\Lambda(x,\tau)|\mu\Delta t_{k,n})^2\,.$$

From (8), (9), conditions A2, A3, A4, and the last inequalities, one finally yields when $n\to\infty$ that

$$(\Delta t_{k,n})^{-1}E(y-x)p_n(y/x) \to \sum_{m=1}^{\infty}p_m\sum_{l=1}^{m}\alpha_l(\xi_l-x)+\Lambda(x,t)\,;$$

$$(\Delta t_{k,n})^{-1}E(y-x)(y-x)'p_n(y/x) - \Delta t_{k,n}(E(y-x)p_n(y/x))^2 \to \sum_{m=1}^{\infty}p_m\left\{\sum_{i,j=1}^{m(m+1)}\alpha_i\left[\gamma_{ij}-x'\lambda_{1,i}-\lambda_{1,j}x'+xx'\right]\alpha'_j\right\}.$$

Hence, condition I is satisfied.

For the proof of the entire theorem, it is necessary to verify the convergence conditions II and III. However, it is readily seen that the condition II is satisfied as a consequence of the continuity of the operator and conditions A2 and A3. Finally, the condition III follows from the existence of a uniform bound of weight coefficients (A5) and the nonzero variance, if the distribution of the random index ν is not zero with probability 1. This proves the theorem completely.

Remark. All parameters p, α, λ, and γ depend, in general, on both t and x. To avoid the overloading of the text and to simplify the notation, this dependence is not explicitly shown.

This theorem may be generalized on the case when a random index ν is a vector with different distributions for each component, i. e.,

$$p(\nu=l)=p(\nu_1=l_1,\dots v_r=l_r)\,.$$

Corollary 1. If no observations are assimilated, $a(t,x)=\Lambda(t,x)$ and $b(t,x)=0$. Thus, the limiting diffusion process coincides with the initial model state, which should be expected.

Corollary 2. The probability distribution of the trajectory is determined by the Fokker–Planck equation (Kolmogorov's second equation)

$$\frac{\partial p(t,x)}{\partial t} = -\frac{\partial(a(t,x)p(t,x))}{\partial x}+\frac{1}{2}\,\frac{\partial^2(b^2(t,x)p(t,x))}{\partial x^2} \qquad (10)$$

with the initial conditions $p(0,x)=p(\theta_0=x)$ and boundary conditions $p(t,\pm\infty)=0$.

In Eq. (10), the drift vector $a(t,x)$and the diffusion matrix $b^2(t,x)$ are given by (4) and (5).

4 Optimization Problem

Above, the convergence of a sequence of random variables was considered with all parameters fixed. However, in practical applications, some parameters may be unknown, but rather be sought according to given criteria. The most relevant physical problem is the determination of optimal weight coefficients α_l. By Theorem 1, the limiting trajectories will be those of a diffusion process with the drift vector $a(t,x)$ and the diffusion matrix $b^2(t,x)$, which determine the mean and its variance. More precisely, the following equalities are valid:

$$\frac{\partial}{\partial t}E\left(\frac{\theta(t)}{\theta_0}\right)=\int\limits_{-\infty}^{\infty}a(t,x)\,d_xp(t,x)\,;$$

$$\frac{\partial}{\partial t}\,\mathrm{var}\left(\frac{\theta(t)}{\theta_0}\right)=\int\limits_{-\infty}^{\infty}b^2(t,x)\,d_xp(t,x)\,.$$

These relations render it sensible to investigate the optimization problem under the following constrains.

Let the unknown "true" field θ_t be estimated by observations and by the model from the aforementioned scheme. It is supposed that the unknown field is governed by the equation $d\theta_t=C\,dt+\eta\,dW$ with the initial value θ_0 being known. The problem is to define the coefficients so that the variance of the estimator is minimized while the average remains unchanged. This is the well-known problem of constraining the unbiased estimator with minimum variance.

Theorem 2. *Let the conditions of theorem* 1 A1–A6 *hold. Also, let $\alpha^*_{l,n}$ be the coefficients that provide the minimum norm of the conditional variance $(\Delta t_{k,n})^{-1}||E[(\theta^n_k) - E\theta^n_k)(\theta^n_k - E\theta^n_k)'/\theta_0]||$ while the conditional average $(\Delta t_{k,n})^{-1}E[(\theta^n_k)/\theta_0]$ remains fixed and equals to C in each series. The numerical sequence $\alpha^*_{l,n}$ will then converge to coefficients α^*_l, satisfying the system of equations with the unknown vector $\Phi = (\varphi_1, \ldots, \varphi_r)$:*

$$\sum_{j=1}^{\infty} p_j \left(\sum_{l=1}^{j} \alpha^*_l g_{li} + \Phi \lambda'_{1,i} - E\xi_j \eta \, dW \right) = 0 \quad \text{for each } i = 1, \ldots, \qquad (11)$$

where $g_{ij} = \gamma_{ij} - (x\lambda'_{1,i} - \lambda_{1,j}x') + xx'$, and

$$\sum_{l=1}^{\infty} p_l \left[\sum_{j=1}^{l} \alpha^*_j (\lambda_{j1} - x) \right] + \Lambda(x,t) = C\,. \qquad (12)$$

Equation (11) is a matrix equation with unknown matrices α_i, the equality is valid for each element of the corresponding matrices. At any given constant C, the system (11), (12) has a unique solution for α^*_l and Φ if and only if the observations as random vectors are linearly independent of each other, i. e., no observation can be represented as a linear combination of the others.

Remark. Statistically, the constant C measures the model bias with respect to the observations. This bias can be eliminated from the algorithm if it is known and/or previously determined.

Proof. The expression for the extremes is obtained by utilizing the method of Lagrange multipliers. For this, it is necessary to determine the minimum of the following function of α and Φ:

$$F(\alpha, \Phi) = ||E(\theta - E\theta - (\theta_t - E\theta_t))(\theta - E\theta - \theta_t - E\theta_t))'|| + \Phi(E\theta - E\theta_t) \qquad (13)$$

where Φ is a r-dimensional unknown vector. The minimum of this function requires a minimum of the sum of the elements in each row of matrices α_i for all $i = 1, \ldots$

The scheme of the proof is as follows: (i) inserting into Eq. (13) the explicit expression for θ and θ_t according to (4) and (5); (ii) taking the derivatives of Eq. (13) with respect to α_i and Φ and setting them to zero; (iii) substituting of matrices g_{ij} and g_{ji} because they are identical; and $(i\nu)$ performing ordinary calculations which lead to Eqs. (11) and (12). The convergence of the extremes is proven along the same line of arguments as in the proof of Theorem 1. Actually, it is only necessary to prove the existence and uniqueness of the solution of (11), (12) if all matrices g_{ij} are linearly independent. Indeed, system (11), (12) is a linear one. For large m, the residual of the sums $\sum\limits_{i=m}^{\infty} p_i$ go to zero. Thus, it is sufficient to consider only a finite number of equations in (11), (12). Let l equations be considered. The matrix in system (11), (12) has the rang $r^2l + r$. Namely, it consists of matrices $g_{kn} = g_{nk}$, $k, n = 1, \ldots, l$, with symmetrically attached last row and column:

$$A = \begin{pmatrix} g_{11} \; g_{12} \cdots g_{1l} \; f_1 \\ g_{21} \; g_{22} \cdots g_{2l} \; f_2 \\ \cdots \cdots \cdots \cdots \cdots \\ g_{l1} \; g_{l2} \cdots g_{ll} \; f_l \\ f_1 \;\; f_2 \; \cdots \; f_l \;\; 0 \end{pmatrix}.$$

Therefore, the determinant of A does not vanish, if rows or columns of matrix A are linearly independent. The variances form the diagonal of this matrix while covariances are placed outside of the diagonal. Because of linear independence of observations, all rows are linearly independent. This completes the proof of Theorem 2.

Theorem 2 states that at each small time interval, once the initial conditions and observation statistics are known, the optimal weight coefficients are to be determined by formulae (11) and (12). However, in practical applications, it is desirable to define these coefficients once and forever, knowing only the initial values and the temporal properties of observations, such as the forecast of their mean values and covariance. The most relevant physical problem in this case is to minimize the functional T^{-1}

$$\int\limits_0^T \left|\left| E\left[\frac{(\theta(t) - E\theta)(\theta(t) - E\theta)'}{\theta_0} \right] \right|\right| dt \qquad (14)$$

with conditions $T^{-1} \int\limits_0^T E\left[(\theta(t))/\theta_0\right] dt = C$.

Theorem 3. *Let the same scheme of constraining the trajectories be considered and all the conditions* A0–A6 *hold. If the limiting average vectors $\lambda_{1,i}$ and covariance matrices γ_{ij} are continuous in time, then, starting from the known vector θ_0, the optimal trajectories with respect to criteria of Theorem 2 will be given by formulae* (11) and (12).

Proof. After passing the limit under integral, the functional (14) takes the form of (13) where all parameters will be considered as functions of time t. The classical Euler equation for optimal trajectory of functional (14) with respect to $\alpha_i(t)$ coincides with the system (11), (12). The limiting transform under the integral is proven by the continuity of limiting functions and upper bound conditions of functions α_i. Ultimately, the initial conditions $\theta(0) = \theta_0$ simply $\alpha_i(0) = 0$ and, hence, no constants should be added to obtain system (13). This proves Theorem 3.

5 Applications to the Ocean Circulation Model HYCOM and ARGO Drifter Data

In this section, several numerical experiments are conducted as illustrations of the applicability of Theorems 2 and 3 .To realize the assimilation scheme, the Hybrid Coordinate Ocean Model (HYCOM) was used. HYCOM solves five prognostic equations: two for the horizontal motion, one for the continuity equation, and two for the thermodynamic conservation that can be the salinity and the potential temperatures [8, 9]. The model is structured with a hybrid vertical coordinate system to solve the prognostic equations associated with the shallow water physics. It uses isopycnic coordinates for the open stratified ocean, which reverts to terrain-following coordinates in shallow coastal regions, and z-level coordinates in the mixed layer and over unstratified ocean regions.

The data from ARGO drifters available from the site http://www.coriolis.eu.org were assimilated according to the aforementioned scheme. The data selected for the experiments were those daily assimilated during January 2008 in the Atlantic Ocean. The assimilation technique based on the covariance evolution is rather applicable to each model vertical layer than in a fixed depth. It is consistent to suggest that the real link between the physical variables is propagated mostly along isopycnic layers. This allows exploiting the similarity of physics within the same layer due to the concept of identity of water masses. Therefore, the covariance among physical variables in the same water mass depends mostly on the properties of this water mass and on the distance in isopycnic layer between two considered points. This makes the physical justification of the mathematical formalism described above.

The computations were performed on the IBM cluster "Neptune" by using multiparallel computation scheme. The entire Atlantic was divided into 64 subdomains, and all computations were carried out independently in each domain with a parallel technology from MPI library with the exchange of the current information among processors.

Two functions have to be determined prior to solving the system. These are the bias C and the covariance $g_{ij} = \gamma_{ij} - (x_j E\xi_i' - E\xi_j x_i') + x_i x_j'$ or simply the matrix G of the covariance of the error between a pair of observed vectors $\xi_i - H_i\theta$ and $\xi_j - H_j\theta$ where $H_i\theta$ and $H_j\theta$ are the projections of the prognostic model state θ onto the observational points $i, j = 1, \ldots, N$. The number of observations N has been given at the day of assimilation so that the probability distribution of the random index ν was set up as $P(\nu(t) = N) = 1$, $P(\nu(t) = L \neq N) = 0$. Equations (11) and (12) are similar to the standard Kalman filter theory, but include explicitly the model bias. If the bias is zero, the covariance g_{ij} will be the error covariance matrix widely used in the Kalman filter theory. Additionally, it is necessary to set up the covariance $E\xi_i\eta\, dW$ between an arbitrary grid point and observations at the point *i*.

Three different approaches were used to constrain the matrix of covariance G and the covariance $E\xi_i\eta\, dW$.

(A) A common way to define this matrix is through the extended Kalman filter approach (EKF), which is a version of the Monte-Carlo scheme [2]. The observations are supposed to follow Eq. (2) with zero average and the variance of the noise part being known. Ensemble experiments are conducted with different initial conditions and the error covariance matrix is estimated from the series of model outputs. This approach requires substantial computational costs and its accuracy increases with the square root of M where M is the number of ensemble members. However, the generation of an adequate ensemble set is a challenge, since for practical applications, only a relatively small number of members may be used, and this may not lead to a good estimation.

In the present work, 10 ensemble members were set to realize the EKF scheme to constrain the covariance. The method of their contraction was the following: the model was initialized with the World Ocean Atlas temperature and salinity fields at rest. It was integrated for 40 years with atmospheric forcing from the Comprehensive Ocean-Atmosphere Data Set (COADS) climatological monthly mean fields available at http://icoads.noaa.gov as the spin-up run. After that, each January 1 of the last 10 years of the spin-up was used as initial condition to create a 10-member ensemble. Ten members were forced by 6-hour reanalysis data from the U.S. National Oceanic and Atmospheric Administration/National Centers for Environmental Prediction (NOAA/NCEP) with 1 degree horizontal resolution during 13 months, from January 1, 2007 until January 31, 2008.

Knowing the model output for all grid points after 10 independent runs, it is easily to estimate the covariance matrix simply as

$$g_{ij} = \frac{1}{10}\sum_{l=1}^{10}\left(\theta_i^l - E\theta_i\right)\left(\theta_j^l - E\theta_j\right)'; \;\; E\theta_i = \frac{1}{10}\sum_{l=1}^{10}\theta_i^l$$

where θ_i^l is the value of the lth model output at the grid point i, $i = 1, L$, $l = 1, \ldots, 1, 0$, and L is the total number of grid points. Since the model temperature and salinity are considered, $r = 2$. The numerical consumption regarding this scheme is very high due to the huge size of this matrix, however this scheme can be realized. The issue of the numerical realization and its cost is out of the scope of this paper. The bias value C was zero for this scheme.

(B) The scheme of the creation of the matrix G stems from the limit approximation as diffusion-type process of the model output θ and, hence, the representation of this matrix through the solution of the FP equation

$$\frac{\partial p(t,u)}{\partial t} = -\frac{\partial (Ap)}{\partial u} + \frac{1}{2}\,\frac{\partial^2 (Bp)}{\partial u^2}\,; \qquad (15)$$

$$g_{ij} = \int\limits_{-\infty}^{\infty} (u_i - E\theta_i)(u_j - E\theta_j)p(t,u_i,u_j)\,du_i du_j$$

where $p(t,u) = p(t,u_i,u_j)$ is the joint probability density of the error between points i, j, $i, j = 1, \ldots, N$; t is the time; and A and B are the drift and diffusion coefficients, respectively. All pairs for temperature–temperature, temperature–salinity, and salinity–salinity are considered. These coefficients are defined through the model output and data as it is done in [6]. Briefly, their calculation scheme is the following: at time moment t, all grid points where temperature and/or salinity values are equaled to $u = (u_i, u_j)$ are marked. Let this value be $L(u)$. Then, at the time moment $t + dt$, among the marked grid points, take all grid points in which the considered variable is equal to $v = (v_i, v_j)$ and let this number be $L(v)$. Then, the ratio $p(v/u) = L(v)/L(u)$ is taken as the estimation of the conditional probability $p(t, v/u) = p(\theta(t + dt) = v/\theta(t) = u)$. Following [4], the drift coefficient is $A(t,u) = (dt)^{-1} \int\limits_{-\infty}^{\infty} (v - u)p(v/u)\,dv$ and diffusion matrix is $B(t,u) = (dt)^{-1} \int\limits_{-\infty}^{\infty} (v - u)(v - u)'p(v/u)\,dv - dtA^2(t,u)$. In the current work, the diffusion coefficient was prescribed and equaled to $B = (1/(N-1)) \sum\limits_{l,j=1}^{N} (\xi_i - H_i\theta)(\xi_j - H_j\theta)'$. Once the coefficients are determined, Eq. (15) are solved with the boundary conditions $p(t,u) = 0, u \to \pm\infty$ and the known initial conditions, $p(0,u) = \delta(u - u^0)$, $\delta(u)$ is the Dirac delta-function. The latter condition means that at the initial moment, the error was known and equaled to u^0. The bias C at an arbitrary grid point was set up as $C = (1/\bar{N}) \sum\limits_{l=1}^{\bar{N}} (\xi_i - H_i\theta)$ if the distance between the location of ξ_i and the considered grid point did not exceed the selected cut-off radius and $\bar{N}$ denotes the quantity of observed points within this circle. For remote grid points where the locations of observed points exceeded the cut-off radios, the bias was set to zero.

(C) The matrix G was prescribed and set up as

$$g_{ij} = \sigma^2 \exp(-\lambda d_{ij})$$

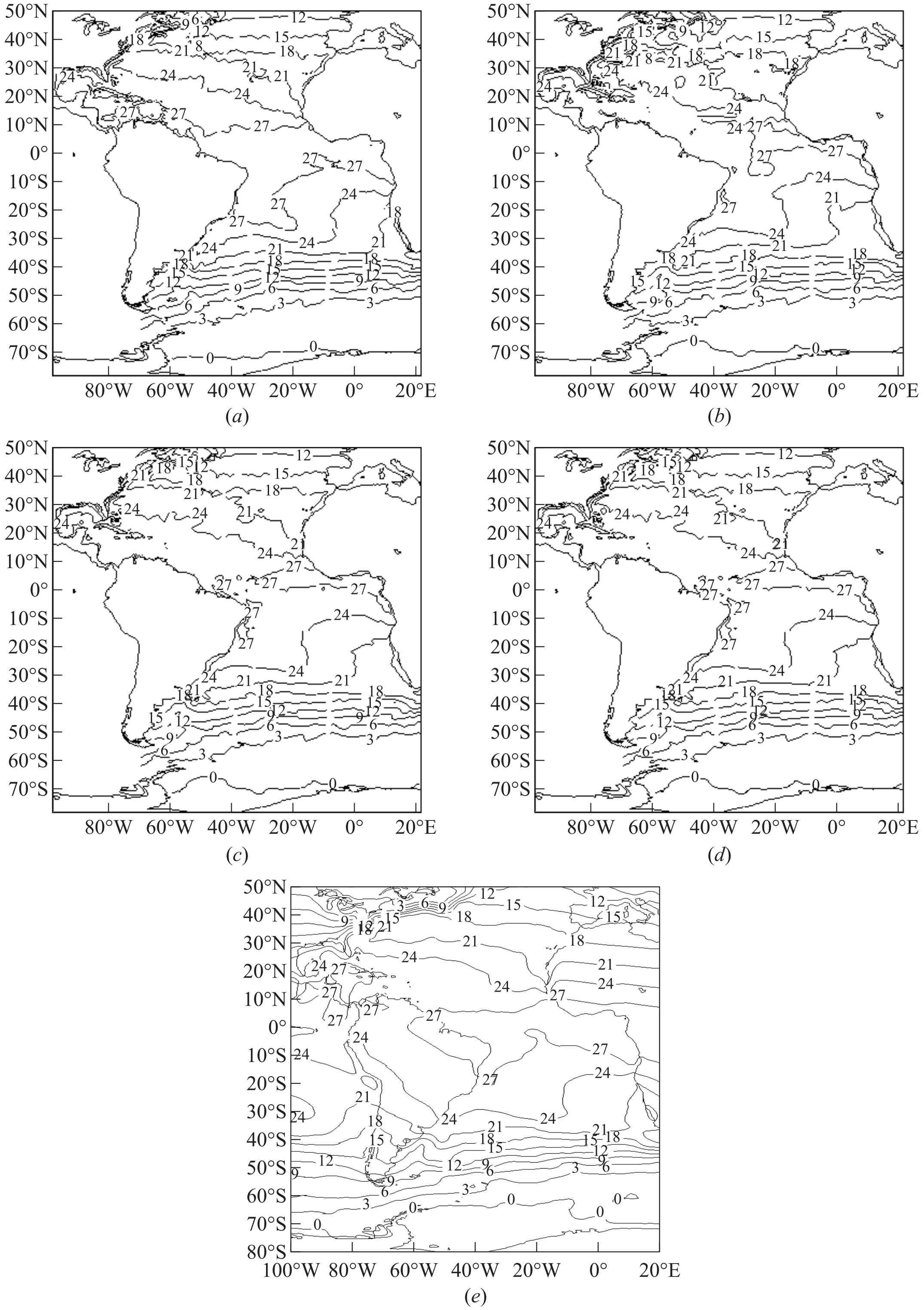
where d_{ij} denotes the grid distance between points i and j, i. e., all neighboring points in any horizontal direction have the distance equaled to 1, and λ is a known dimensionless factor (normally varying between 0.1 and 0.3). The bias was set up as in scheme (B).

6 Results of Experiments and the Comparison of the Data Assimilation Scheme

For illustration of the feasibility of the considered assimilation schemes, Fig. 1 presents the results of the modeling with and without assimilation. This figure shows the snapshot of the sea surface temperature (SST) taken on January 30, 2008 for all four assimilation schemes and for the control run, i. e., a model run without assimilation. It also contains the Reynolds SST analysis from remote sensed and *in situ* data on the same day. Data were retrieved from the site http://www.nhc.noaa.gov/aboutsst.shtml. The control run overestimated the temperature in the tropics and in the North Atlantic mid-latitude. This can be seen by the area covered by the 27, 24, and the 21 °C isotherms. Also, the 21 °C isotherm to the west of Southern Africa associated with the Benguela Current does not match observations. All assimilation runs produced substantial reduction of temperature in the tropics and cooling to the west of southern place of Africa around 30°S. This correction makes all analysis closer to the Reynolds SST in many regions. However, the correction in the equatorial western Atlantic and in the Caribbean Sea was intense, and the analyses produced cooler values than Reynolds SST in these regions. The EKF produced the strongest cooling. The FP result is in between the EKF and OI run.

In order to numerically compare the skill of all methods, the time behavior of the variance for each method is presented in Fig. 2. Let $\sigma = \sqrt{\sum\limits_{i=1}^{N} (\xi_i - H_i\theta)^2/(N-1)}$ be the root mean squared error (RMS) averaged over all observations. Along with the variable σ, two other variables, σ_b and σ_a, are considered where σ_b is the RMS for the one-day forecast error, i. e., the deviation is considered between the one-day forecast run made after correction and the data at this moment; and σ_a is the RMS for the analysis error, i. e., the deviation is considered after correction at the instant of the correction. These two variables are calculated for all assimilation methods independently and for each day from January 2 until January 30, 2008.

Figures 2*a*–2*c* show the time behavior of σ, σ_b, and σ_a for the temperature according to the EKF, the OI, and the FP schemes, respectively. The EKF forecast error in the beginning of the assimilation run is larger



(*a*)

(*b*)

(*c*)

(*d*)

(*e*)

Figure 1 Sea surface temperature (in °C) modeled by HYCOM with and without correction on January 30, 2008 and SST observed by satellite on the same day: (*a*) control; (*b*) OI; (*c*) EKF; (*d*) FP; and (*e*) Reynolds SST

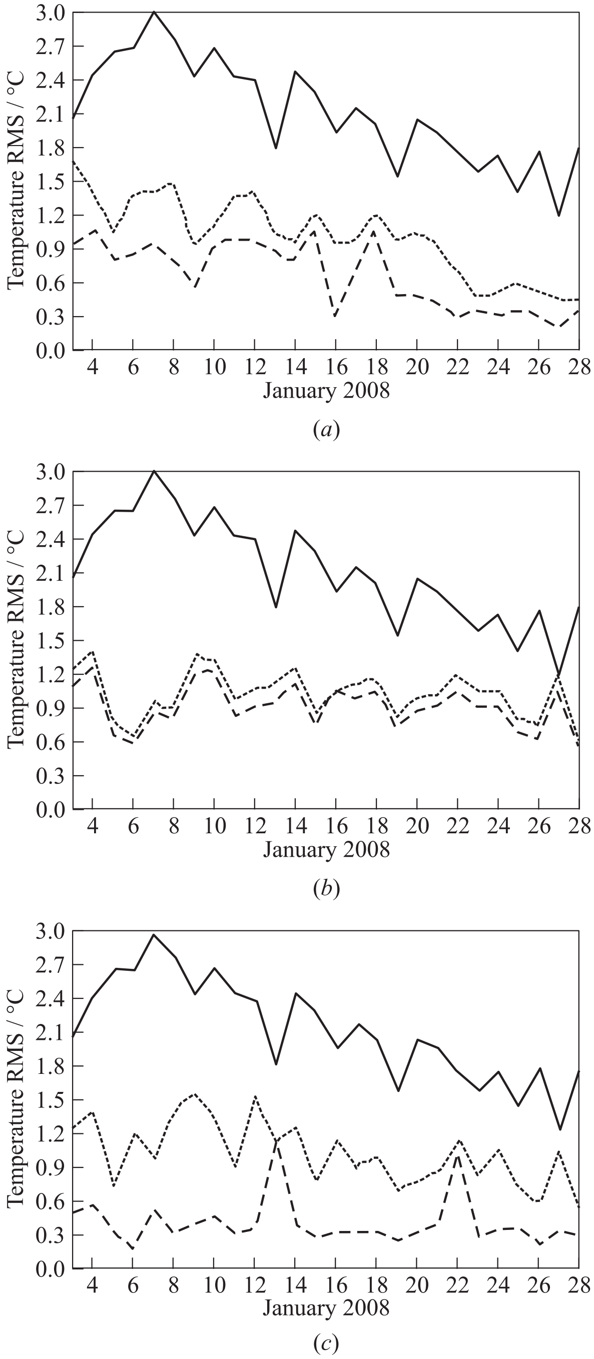


Figure 2 Root mean squared error of temperature over the period January 2–30, 2008 for all assimilation schemes: solid curves — control run; dotted curves — the one-day-forecast error, and dashed curves — the analysis error: (*a*) OI; (*b*) EKF; and (*c*) FP

than the error of the control run. In the second day, the one-day forecast error is already smaller than the control run error, and there is a substantial reduction of the forecast error in time from around 1.5 °C in the first day to 0.8 °C in the end of the month. The EKF analysis errors have similar behavior, starting around 1.0 °C and reaching a minimum smaller than 0.3 °C few days before the end of the experiment. The OI forecast errors (see Fig. 2*b*) oscillated in the beginning of the experiment between 0.6 and 1.4 °C. However, it does not decrease in time, so that the forecast error by the end of the experiment is of the same magnitude as in the beginning. The OI analysis error has almost exactly the same variability of its forecast errors, but with values about 0.1 °C lower along the whole period. The FP forecast errors (see Fig. 2*c*) oscillated from 0.7 to 1.5 °C in the first 15 days and then it went down to lower values by the end of the experiment. The FP analysis error is the smallest among the three assimilation runs, despite the three pronounced spikes in the days 13, 22, and 29. Without considering these spikes, it oscillated around 0.3 °C.

7 Discussion and Concluding Remarks

The presented results are of both theoretical and practical interest. From the theoretical point of view, it is important to know when and under which conditions the solution of differential equations (1) along with the correction (2) may be approximated by a continuous function and how the probability of limiting function could be calculated. In practical applications, the scheme can be utilized to determine a variety of relevant parameters such as the probability of extremes, the probability for crossing some levels, and similar issues. Also, the optimization scheme generalizes the conventional schemes in the assimilation problem; besides, it includes the bias and random quantity of observations.

Nevertheless, several issues still have to be addressed specifically. The system of Eqs. (10) solves the assimilation problem ultimately and uniquely, once the observational statistics, averaged values, covariance, and their time evolution are known. However, in practice, it is not an easy task. Data insufficiency, irregularity of measurements in space and time as well as their inhomogeneous spatial-temporal distribution often make corresponding estimations highly unreliable.

The application of proposed methods was presented in [10] in a climate research with a couple model EGMAM, and this application demonstrated the ability and fruitfulness of the used approach.

Acknowledgments

This work was financially supported by PETROBRAS and Agencia Nacional do Petroleo, Gas Natural e Biocombustiveis (ANP), Brazil, via the Oceanographic

Modeling and Observation Research Network (REMO); Russian Foundation for Basic Research (project 11-07-00161-a).

References

1. *Gill M., Malanotte-Rizzoli P.* Data assimilation in meteorology and oceanography // Adv. Geophys., 1991. Vol. 33. P. 141–266.
2. *Evensen G.* Sequential data assimilation with a non-linear quasi geostrophic model using Monte-Carlo methods to forecast error statistics // J. Geophys. Res., 1994. Vol. 6. P. 10143–11062.
3. *Cohn S.* An introduction to estimation theory // J. Meteor. Soc. Japan, 1997. Vol. 75. P. 257–288.
4. *Gikhman I. I., Skorokhod A. S.* Introduction to the theory of random processes. — Dover Publications, 1996.
5. *Strook D., Varadhan S. R. S.* Multidimensional random processes. — Berlin: Springer-Verlag, 1995.
6. *Belyaev K., Tanajura C. A. S., O'Brien J. J.* A data assimilation technique with an ocean circulation model and its application to the tropical Atlantic // Appl. Math. Model., 2001. Vol. 25. P. 655–670.
7. *Tanajura C. A. S., Belyaev K.* A sequential data assimilation method based on the properties of diffusion-type process // Appl. Math. Model., 2009. Vol. 33. P. 2165–2174.
8. *Bleck R., Boudra D. B.* Initial testing of a numerical ocean circulation model using a hybrid quasi-isopycnal vertical coordinate // J. Phys. Oceanogr., 1981. Vol. 11. P. 750–770.
9. *Bleck R.* An oceanic general circulation model framed in hybrid isopycnic Cartesian coordinates // Ocean Model., 2002. Vol. 4. P. 55–88.
10. *Belyaev K. P., Tuchkova N. P., Cubasch U.* Response of a coupled ocean-ice–atmosphere model to data assimilation in the tropical zone of the Pacific Ocean // J. Oceanology, 2010. Vol. 50. No. 3. P. 306–316.

МАТЕМАТИЧЕСКОЕ ОБОСНОВАНИЕ, ПРИМЕНЕНИЕ И СРАВНЕНИЕ ОБОБЩЕННОГО МЕТОДА УСВОЕНИЯ ДАННЫХ НАБЛЮДЕНИЙ, ОСНОВАННОГО НА МЕТОДАХ ДИФФУЗИОННОЙ АППРОКСИМАЦИИ, С ДРУГИМИ МЕТОДАМИ УСВОЕНИЯ ДАННЫХ

К. П. Беляев[1], К. А. С. Танажура[2], Н. П. Тучкова[3]

[1] Институт океанологии им. П.П. Ширшова Российской академии наук, kb@sail.msk.ru
[2] Федеральный университет штата Баийя, Бразилия, cast@ufba.br
[3] Вычислительный центр им. А. А. Дородницына Российской академии наук, tuchkova@ccas.ru

Аннотация: Многие методы усвоения данных, применяемые в численных океанских и атмосферных моделях, базируются на теории случайных процессов. Предложен метод усвоения, основанный на построении специальной последовательности цепей Маркова, с помощью которой строится сходимость к состоянию модели. Исследуются условия этой сходимости. Решается проблема оптимизации параметров этой цепи для наилучшего приближения, и обсуждаются результаты численных экспериментов. Показано, что предложенный метод усвоения данных может использоваться в практическом применении в метеорологии и океанографии. В данном исследовании метод применялся для океанской модели HYCOM и данных наблюдений с дрифтеров АРГО. В работе также выполнялись эксперименты с другими методами усвоения. Представлены результаты сравнения и анализа.

Ключевые слова: последовательность цепей Маркова; диффузионный случайный процесс; методы усвоения данных наблюдений; НУСОМ (Гибридная модель циркуляции океана); дрифтеры АРГО