Sequential data assimilation techniques in oceanography

Laurent BERTINO[1], Geir EVENSEN[1], Hans WACKERNAGEL[2]

[1]Nansen Environmental and Remote Sensing Center, Norway

[2]Centre de Géostatistique, Ecole des Mines de Paris, France

Email: Laurent.Bertino@nersc.no

Summary

We review recent developments of sequential data assimilation techniques used in oceanography to integrate spatio-temporal observations into numerical models describing physical and ecological dynamics. Theoretical aspects from the simple case of linear dynamics to the general case of nonlinear dynamics are described from a geostatistical point-of-view. Current methods derived from the Kalman filter are presented from the least complex to the most general and perspectives for nonlinear estimation by sequential importance resampling filters are discussed. Furthermore an extension of the ensemble Kalman filter to transformed Gaussian variables is presented and illustrated using a simplified ecological model. The described methods are designed for predicting over geographical regions using a high spatial resolution under the practical constraint of keeping computing time sufficiently low to obtain the prediction before the fact. Therefore the paper focuses on widely used and computationally efficient methods.

Key words: Data Assimilation; Geostatistics; Kalman Filter; Non-Linear Dynamical Systems; State-Space Models; Ecological Model.

1 Introduction

Data assimilation (DA) can be defined as the incorporation of observations into a dynamical model to improve forecasts. The techniques of DA have found many applications in the fields of meteorology and oceanography. They have in particular been successfully applied in operational weather forecasting, first with Optimal Interpolation (Gandin, 1963 ; Daley, 1991), then later with the adjoint method (Le Dimet and Talagrand, 1986 ; Courtier et al., 1998) in its conceptually most complex form. Predictions of physical parameters of the oceans (tidal currents, sea surface elevation, temperature) are presently in operational use and in the near future predictions of ocean biological parameters should be utilized operationally. The latter are critical to efficient policy making in the coastal zones environments and for fish stock management. We shall concentrate hereafter on DA methods applied in oceanography, that are nevertheless of interest for studies of any application having variations in space and time.

We denote $0:n$ the sequence $\{0,\ldots,n\}$. The DA state space model is basically a hidden Markov chain and the hidden state at time n is a vector $\mathbf{x}_n$ carrying several biogeochemical water parameters at all points of the domain. The two main inputs for its estimation are:

1. Physical processes can be described by a set of differential equations. The equations are most often discretized spatially and temporally, which means that the components of the vector $\mathbf{x}_n$ are average values of variables of interest on a grid cell and during a model integration time step. The spatial resolution has to be high enough to let the model resolve the relevant processes, therefore the state dimension N in oceanography is usually in the range from 10^5 to 10^6 parameters, which leads to severe numerical problems in DA. The discretized model gives the transition between two successive states $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$.

2. The other source of information is the data $\mathbf{y}_n$. The data measures imperfectly the spatiotemporal domain, for example satellite data only inform about the ocean surface while

buoys can provide dense vertical observations but only at few points in the horizontal plane. The observations carry a random error about which distributional assumptions are made. The number of observations at a given time is denoted m, it can be less than 10 in the case of in-situ data, or larger than 10^5 in the case of satellite data.

DA methods are divided into two classes: variational and sequential methods. Variational DA is based on the optimal control theory. Optimization is performed on unknown parameters (for example the initial state $\mathbf{x}_0$) by minimizing a given cost function that measures the model to data misfit. Among variational methods, the representer method is used to solve the full problem (Bennett et al., 1996 ; Ngodock et al., 2000) while the adjoint method (Le Dimet and Talagrand, 1986) has been successfully used in meteorology and oceanography (Luong et al., 1998 ; Nechaev and Yaremchuk, 1994 ; Thacker and Long, 1988), but assuming a perfect model. They were mainly developed for meteorological forecasting.

In this article, we will focus on sequential DA methods. These methods use a probabilistic framework and give estimates of the whole system state sequentially by propagating information only forward in time. They therefore avoid deriving an inverse or an adjoint model and make sequential methods easier to adapt for all models. Further, the probabilistic framework is necessary to quantify the uncertainty associated with the results.

It is common to assume that the state $\mathbf{x}_{n+1}$ depends only on the state $\mathbf{x}_n$ but not on previous ones $\mathbf{x}_{0:n-1}$. The observations $\mathbf{y}_n$ depend only on the state $\mathbf{x}_n$ according to the scheme:

$$\begin{array}{ccccccccc} \mathbf{x}_0 & \longrightarrow & \dots & \longrightarrow & \mathbf{x}_n & \longrightarrow & \mathbf{x}_{n+1} & \longrightarrow & \dots \\ & & & & \downarrow & & \downarrow & & \\ & & & & \mathbf{y}_n & & \mathbf{y}_{n+1} & & \end{array} \tag{1}$$

We look for the most likely state trajectory $\mathbf{x}_{0:n}$ knowing the state dynamics and the observations. We estimate the posterior density $\phi(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$, whose marginal is the well-known filtering density $\phi(\mathbf{x}_n|\mathbf{y}_{0:n})$. The latter is often used for prediction purposes. Sequential DA methods es-

timate the latter density recursively in two steps: first the propagation step uses the dynamical model to determine the prior distribution $\phi(\mathbf{x}_n|\mathbf{y}_{0:n-1})$ then a statistical analysis of the observations $\mathbf{y}_n$ updates the prior distribution and provides the posterior distribution $\phi(\mathbf{x}_n|\mathbf{y}_{0:n})$.

Many sequential DA methods have been proposed over the last decades and developed in oceanography and air pollution. Some of them have been compared experimentally to simulated test cases (Cañizares, 1999 ; Verlaan, 1998 ; Pham, 2000 ; Verlaan and Heemink, 1999) focusing on computer time, robustness of the algorithms to nonlinearities (Verlaan and Heemink, 1999), and robustness to systematic errors and to incorrect specification of error statistics (Cañizares, 1999). Brusdal et al. (2003) have compared the feasibility of three methods for operational marine monitoring and forecasting systems, assimilating remote sensing data in a model of the North Atlantic Ocean. However these papers typically mask the discussion of the underlying stochastic models by matrix algebra considerations. To our point-of-view, the core of the DA problem lies in the stochastic model of the state variable $\mathbf{x}_n$ used at the interface between the propagation and analysis steps. It should be consistent with the physical properties of the system (for example the geostrophic balance or the non-divergence of the flow) so that the estimates obtained by statistical analysis can be used to restart the following propagation step.

We present here the most common sequential DA methods in their probabilistic formulation, described in order of increasing complexity. In the linear case, under Gaussian assumptions on the error terms, the optimal solution is given by the classical Kalman filter (Section 2). We compare the Kalman filter to the Optimal Interpolation methods. For nonlinear dynamics, an Extended Kalman filter can be derived by linearization at each time step but requires simplifications to make it computationally tractable for oceanographical applications (Section 3). The Ensemble Kalman filter (Section 4) uses Monte Carlo sampling to approximate the prior distribution while still applying a linear analysis step. An original contribution of this paper is the application of the Ensemble Kalman filter to transformed Gaussian (or "anamorphosed") variables. Improvements

are illustrated on a simplified ecological model. In Section 5, we leave Kalman filters with the description of other Monte Carlo methods that perform nonlinear analysis and do not require any Gaussian assumption.

2 The Kalman filter

We consider the case of linear dynamics. The optimal solution is classically given by the Kalman filter (KF), against which the Optimal Interpolation appears as a crude approximation. The efforts related to the practical application of the KF in high-dimensional systems are reviewed. Possible improvements of the KF by geostatistical tools are then discussed.

2.1 Comparison to Optimal Interpolation

The simplest model for DA assumes linear dynamics. Although purely academic, this model has inspired most of the DA methods currently used. The associated state space model is:

$$\mathbf{x}_n = F_n \mathbf{x}_{n-1} + \varepsilon_n^m \tag{2}$$

$$\mathbf{y}_n = H \mathbf{x}_n + \varepsilon_n^o, \tag{3}$$

where F_n is the $N \times N$ dynamical model matrix and H the $m \times N$ observation matrix. Dynamics and observations are supposed imperfect, so ε_n^m and ε_n^o are respectively model and observation random errors.

In linear DA methods, the model and observation error processes are zero mean Gaussian white noise in time but may be correlated in space. Their time-invariant covariance matrices are respectively Σ^m and Σ^o. The independence of ε_n^m and ε_n^o is also assumed.

Under these assumptions, the Kalman filter estimates $\mathbf{x}_n$ by its first two moments, the mean and the variance-covariance matrix C_n. In DA notations, the results of the propagation step are labeled with f since they are interpreted as a "forecast" of the system variables and the forecasts

updated by the data are labeled with a as in "analysis". The two forecast moments are:

$$\mathbf{x}_n^f = F\mathbf{x}_{n-1}^a \tag{4}$$

$$C_n^f = FC_{n-1}^a F^\top + \Sigma^m. \tag{5}$$

The distributions of $\mathbf{x}_n$ and $\mathbf{y}_n$ are Gaussian, the optimal estimator is the linear combination that minimizes the estimation variance. As the model is assumed to be unbiased, the mean of $\mathbf{x}_n$ is known. This estimation in geostatistical terminology is a simple kriging with a matrix of weights:

$$K_n = C_n^f H^\top (HC_n^f H^\top + \Sigma^o)^{-1}, \tag{6}$$

known as the Kalman gain, and the analyzed moments are then:

$$\mathbf{x}_n^a = \mathbf{x}_n^f + K_n(\mathbf{y}_n - H\mathbf{x}_n^f) \tag{7}$$

$$C_n^a = C_n^f - K_n H C_n^f. \tag{8}$$

By comparison, Optimal Interpolation is more rudimentary. The covariance matrices C_n^f are computed under second order stationarity assumption, and are not propagated in the dynamical model. Therefore, a serious difficulty of the Optimal Interpolation is the need for a stationary spatial covariances that can reasonably represent the ocean variability throughout the whole domain at all times. To this aim, we find two specific difficulties in ocean modeling: first the ocean has a layered structure (the variability of ocean parameters is higher above the level of the thermocline, which is often sharp and has a variable depth), then the multivariate relationships between the ocean variables are dependent both on location and time. For example the correlation between sea-ice thickness and the upper sea salinity can take either positive or negative sign depending on the dominant process in sea-ice dynamics (personal communication from Knut Arild Lisæther). Because of such particularities of the ocean parameters, the design of the covariance matrix C_n^f in Optimal Interpolation is a very complex task although the related stochastic model is the simplest. In contrast, the multivariate covariance matrix obtained by the KF propagation step (5)

can automatically account for the system dynamics and it is spatially not stationary. Thus the KF should be preferred.

2.2 Practical application in oceanography

The KF, even in the linear case, already presents serious practical difficulties related to computational costs. Physical ocean models need a dense spatial discretization in order to reproduce correctly the mesoscale dynamics. This discretization leads to a state dimension superior to $N = 10^5$ parameters. The storage of a single associated covariance matrix of size N^2 requires around ten gigabytes and the matrix product in (5) involves $2N$ model propagations. Such a propagation step would require years of CPU time and simplified methods are needed. Numerous methods have been developed for reducing this computational load, reviewed in Ghil and Malanotte-Rizzoli (1991). Among them Fukumori and Malanotte-Rizzoli (1995) use a coarser grid to define the covariance, which restricts the effect of the Kalman filter to large-scale corrections only. This might be inefficient in oceanography since the physical models often produce small-scale errors, as for example the misplacement of some eddies or of ocean meanders, and the error covariance needs small-scale structures.

For operational forecasting, some authors obtain a Kalman gain matrix K_∞ from a previous run and use it to update the forecasts. The method is known as a "constant gain Kalman filter" (Verlaan and Heemink, 1997). The heavy computation of covariances is performed off-line and is avoided for operational use. This approach resembles an Optimal Interpolation with a covariance matrix inherited from a KF run. The advantage over Optimal Interpolation is that the covariance is in agreement with the physics of the system, however only for a given period of time.

An important simplification of the KF algorithm has been introduced in the DA community by Cane et al. (1996), similar in spirit to the KF by Wikle and Cressie (1999) and the so-called "kriged Kalman filter" discussed by Mardia et al. (1998). The latter terminology by the way

tends to distinguish between kriging and Kalman filtering as two distinct techniques, although the linear update of the KF is nothing but kriging. Cane et al. (1996) proposed a projection of the state space on a reduced basis of r principal components, termed *multivariate Empirical Orthogonal Functions* (EOFs), $\mathbf{x}_n = U\tilde{\mathbf{x}}_n$, where U is the matrix whose columns are selected multivariate EOFs and $\tilde{\mathbf{x}}_n$ is the time-varying principal component that weights these r EOFs. Approximating F_n and H in the new basis, the KF algorithm is applied in the reduced space to estimate $\tilde{\mathbf{x}}_n$. In the case of a model F constant in time, the EOFs can be computed by a singular value decomposition of the matrix F. Approximate recursive methods can be used as this matrix is generally too large to be expressed in an explicit format. The computational load is dramatically reduced if the number r of dominant eigenmodes to be retained is low. Cane et al. (1996) obtained satisfactory results with $r = 5$, and the computational load was divided by 20000. This result indicated that the full computation of equation (5) is not necessary.

Cane et al. (1996) applied this method to the circulation of the tropical oceans, that can be modeled by linear processes controlled by time-varying wind fields. Instead of computing time-dependent EOFs related to the singular values of the matrix F_n, Cane et al. (1996) used a fixed basis taken as dominant EOFs of a historical sequence of model-generated states. The decomposition $\mathbf{x}_n = U\tilde{\mathbf{x}}_n$ shows that, whatever the corrections brought to $\tilde{\mathbf{x}}_n^f$, the solution of the KF will belong to the state spanned by the fixed EOFs computed on the historical sequence. Therefore the success of the algorithm essentially depends on the ability of the model to simulate all likely states of the system in a given period of time without the help of DA: “The unhappy case is when [the multivariate EOFs], which are derived from a limited simulation with an imperfect model, miss the true signal” (Cane et al., 1996).

2.3 Geostatistical perspectives

The observation matrix H is often a projection from the state space to the observation space, which means that model output and observations have to be taken on the same support (i.e. integration time and volume). Since in real cases the size and acquisition times of sensors usually do not match model grid cells or model time steps, one needs a change-of-support model of geostatistics (Chilès and Delfiner, 1999) to describe how distributions vary depending on support. However DA is a different framework for nonlinear geostatistics because of scarce data and nonstationary variables, and change-of-support corrections in DA have until now been very rudimentary, limited to interpolating observations onto the model grid and adding spatial noise to the observation error to account for change-of-support variability.

The computation of EOFs of ocean modeled fields needs some precautions due to their spacial and temporal correlations. We point towards von Storch and Zwiers (1999) for illustrations of the use of EOFs in natural sciences and towards Wackernagel (2003) for approaches to compute EOFs of multivariate autocorrelated data using linear coregionalization models.

The KF assumes an unbiased model but most of the ocean models have biases due to the spatial discretization and to inaccurate parameterization. In geostatistics, a classical way to filter an unknown model bias is to perform the estimation with respect to universality constraints (Chilès and Delfiner, 1999), under the assumption that the bias is uniform on a certain domain, or that it is a combination of translation-invariant base functions. Filtering the bias would change the simple kriging in equation (6) into alternative kriging methods (ordinary or universal kriging) without any significant increase of the computational cost of the KF. At the moment the bias filtering has only been experimented in the Optimal Interpolation community but not with the KF.

3 The extended Kalman filter

The extended Kalman filter (EKF) is derived from the basic KF to nonlinear dynamics by linearization at each time step. Both theoretical and practical problems of the use of the EKF are discussed. Computationally efficient variations of the EKF are reviewed with their applications. Among them, two popular methods (the RRSQRT and SEEK filters) are described in more details and compared.

3.1 The nonlinear state space model

The degree of nonlinearity depends not only on the physics of the system, but also on the data sampling frequency. For example the Lorenz equations are broadly used as the prototype of a nonlinear system for testing methods. With frequent sampling in time, the nonlinear evolution of the Lorenz system between two consecutive observations can satisfactorily be approximated as linear. In most oceanographical cases the measurements cannot be that intensive. We study the nonlinear model, that will also be used in Section 4:

$$\mathbf{x}_n = \mathbf{f}_{n-1}(\mathbf{x}_{n-1}, \varepsilon_n^m) \tag{9}$$

$$\mathbf{y}_n = H\mathbf{x}_n + \varepsilon_n^o. \tag{10}$$

The model error in (9) is introduced directly into the physical model **f**, which is adequate for oceanographical DA since the main model error comes from a lack of knowledge of internal model parameters or of model boundary conditions rather than from a misspecification of the physical processes. A practical advantage of this approach is that the dimension of ε_n^m needs not to be equal to the huge state space dimension. Note that the observation operator H remains linear and constant. The filters presented in the two following sections could be expressed with a nonlinear observation operator. However this is avoided in practice because the linear estimator is inefficient with nonlinear observations. In the case when the observations are nonlinear func-

tions of the state variables, as in remote sensing of sea surface color, the observed variables are generally integrated to the state space and the matrix H is used to select them from the space vector.

3.2 Equations of the extended Kalman filter

The idea behind extensions of the KF to nonlinear dynamics (Jazwinski, 1970) is to perform a Taylor expansion of the model at each time step and propagate the error covariance matrix along the truncated series. We note $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}|_{\mathbf{x}^a_{n-1}}$ the linear tangent model (Jacobian matrix of partial derivatives of $\mathbf{f}_{n-1}$ taken at $\mathbf{x}^a_{n-1}$). Truncation at first order provides the following propagation equations:

$$\mathbf{x}^f_n = \mathbf{f}_{n-1}(\mathbf{x}^a_{n-1}, 0) \tag{11}$$

$$C^f_n = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right|_{\mathbf{x}^a_{n-1}} C^a_{n-1} \left.\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right|^{\top}_{\mathbf{x}^a_{n-1}} + \left.\frac{\partial \mathbf{f}}{\partial \varepsilon^m}\right|_{\mathbf{x}^a_{n-1}} \Sigma^m \left.\frac{\partial \mathbf{f}}{\partial \varepsilon^m}\right|^{\top}_{\mathbf{x}^a_{n-1}} . \tag{12}$$

There is a systematic estimation bias inherent to the use of the above propagation equations: if at the previous time step the expected state is $\mathbf{x}^a_{n-1} = E(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1})$ then during the propagation equation (11) $\mathbf{x}^f_n$ becomes a biased estimate of the mean $E(\mathbf{x}_n|\mathbf{y}_{0:n})$ since $\mathbf{f}$ is a nonlinear function that does not commute with the expectation. This problem is solved with the Ensemble KF that uses Monte-Carlo sampling for the propagation. Truncations at second or following orders produce an Extended KF that is still biased, but empirically more robust against nonlinearities (Jazwinski, 1970 ; Maybeck, 1979). However, the equations of the second order EKF become so complex that their application is not feasible in DA. We will therefore consider only first order truncation, for which the analysis equations remain of the form (6), (7) and (8). Yet these linear analysis equations have two other drawbacks when used with nonlinear models. First the linear estimation is not optimal any more because the variables propagated in a nonlinear model are generally not Gaussian. Second, the estimation of the first two moments does not provide the full probability density. There have been several attempts to apply the EKF directly

in different simulated ocean circulation models (Evensen, 1992 ; Gauthier et al., 1993 ; Miller et al., 1994). But the application resulted in an unbounded error growth as soon as the system entered an unstable regime. The role of error propagation linearization in the divergence has been identified by Evensen (1992). Furthermore, the enormous computational effort required is a serious disadvantage of the EKF and numerous suboptimal schemes of the EKF have been developed for implementation in oceanographic studies. They are generally also more robust against nonlinearities than the crude full scale EKF.

3.3 Sub-optimal schemes

Most of the efforts to adapt the EKF to real high-dimensional problems have concentrated on working on approximations of the covariance matrix C_n^f.

Cohn and Todling (1996) propose three suboptimal schemes. The first one is a coarse grid approximation of C_n as in Fukumori and Malanotte-Rizzoli (1995), but they show that a minimum grid resolution is necessary to avoid filter divergence. The second suboptimal scheme is an approximation by singular value decomposition of the tangent linear model and the third is an approximation by eigendecomposition of C_n^f. The latter two suboptimal schemes were found equally efficient. Similar to the third method, the Reduced Rank Square Root (RRSQRT) Kalman filter (Verlaan and Heemink, 1997) and the Singular Evolutive Extended Kalman (SEEK) filter (Pham et al., 1998) both use an eigenvalue decomposition of the covariance matrix C_n^f and approximate the covariance by a projection on a low number of dominant eigenmodes. They have become popular in the oceanographical community and have been applied in real cases. The RRSQRT Kalman filter has been applied to storm surge forecasting (Verlaan and Heemink, 1997), two and three dimensional coastal hydrodynamics (Verlaan, 1998 ; Wolf et al., 2001 ; Cañizares et al., 2001 ; Bertino et al., 2002) and air pollution (van Loon and Heemink, 1997 ; Segers et al., 2000 ; Elbern et al., 2000 ; Segers, 2002 ; Heemink and Segers, 2002). The SEEK filter has

been applied to ocean circulation models (Brasseur et al., 1999), within the European projects DIADEM and TOPAZ (Brusdal et al., 2003) and will soon be applied in the French operational oceanography program MERCATOR.

The sub-optimal schemes of the EKF have been applied satisfactorily in nonlinear cases, but we should remember that they are appoximations of the EKF that originally has an estimation bias in equation (11). These methods can therefore not be considered as theoretical solutions to the nonlinear model (9), although they are found empirically efficient.

3.4 Implementation of the RRSQRT and SEEK filters

In the RRSQRT and the SEEK filters, the authors consider an eigendecomposition of the error covariance matrix $C_n = S_n S_n^\top = U_n D_n U_n^\top$, where S_n is the square root of C_n. The columns of U_n are the r dominant eigenvectors of C_n and D_n the diagonal matrix of the r highest eigenvalues. Propagating the square root covariance reduced to the first r eigenvalues alleviates considerably the computational burden since r is generally smaller than 100 and the square root covariance is an $N \times r$ matrix instead of $N \times N$. If their principle is similar, there are practical differences in the filter initialization, propagation and analysis.

The initial covariance S_0 is seldom known accurately, it is either approximated on the basis of a model output analysis or a data analysis. Brasseur et al. (1999) assume that the model without DA is able to reach well all possible states of the system at the initial time and suggest the use of an EOF analysis of a historical free run of the model. In practice, the covariance obtained by EOF analysis often has long range structures related to the seasonal cycles in the historical run. These long range structures are not realistic and it is not obvious how to remove them. Alternatively we can assume that the initial error is second order stationary, model the covariance by structural analysis of the observations and compute the dominant eigenvectors of the covariance matrix by a Lanczos algorithm. But measurements under the ocean surface are still insufficient for this

task.

The propagation steps of these two suboptimal schemes have some similarities (Verlaan and Heemink, 1997 ; Brasseur et al., 1999). The equation (12) when applied to the square root S_n becomes the concatenation of two matrices, inherited respectively from the propagation of previous analysis errors and from the introduction of new model error:

$$S_n^f = \left[\left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\mathbf{x}_{n-1}^a} S_{n-1}^a, \left. \frac{\partial \mathbf{f}}{\partial \varepsilon^m} \right|_{\mathbf{x}_{n-1}^a} (\Sigma^m)^{1/2} \right]. \tag{13}$$

The eigenvectors from the previous analysis $\mathbf{s}_{n-1}^{i,a}$ (the i-th column of S_{n-1}^a) evolve in time according to an approximate linearized model:

$$\left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\mathbf{x}_{n-1}^a} \mathbf{s}_{n-1}^{i,a} = \lim_{\epsilon \to 0} \frac{\mathbf{f}_{n-1}(\mathbf{x}_{n-1}^a + \epsilon \mathbf{s}_{n-1}^{i,a}, 0) - \mathbf{f}_{n-1}(\mathbf{x}_{n-1}^a, 0)}{\epsilon} \tag{14}$$

$$\mathbf{s}_n^{i,f} \approx \frac{\mathbf{f}_{n-1}(\mathbf{x}_{n-1}^a + \epsilon \mathbf{s}_{n-1}^{i,a}, 0) - \mathbf{f}_{n-1}(\mathbf{x}_{n-1}^a, 0)}{\epsilon}. \tag{15}$$

The linearization parameter ϵ should in principle be very small. However, due to strong model nonlinearities or discontinuities, very small values may initiate model disruptions and filter instabilities (Verlaan, 1998 ; Bertino et al., 2002) and the ϵ has to be fine-tuned to obtain a trade-off between accuracy of model linearization and filter stability. Furthermore, the physical model is supposed to process only valid physical states, and it is implicitly assumed in (15) that the combination of two outputs from statistical procedures (the analysis estimate $\mathbf{x}_{n-1}^a$ and the weighted eigenvector of the error covariance matrix $\mathbf{s}_{n-1}^{i,a}$) can be processed by a realistic ocean model. The potential conflict between physics and statistics will be adressed in Section 4.2, but the problem remains open in the present methods. Another disadvantage of the linearization is its incompatibility with the nonlinear model (9): for example if we replace the covariance square root S_n by its opposite $-S_n$, the propagation in the nonlinear equation (15) produces a different result. Verlaan (1998) solves the problem by gathering in S_n the dominant eigenvectors with both signs and obtains a more robust filter, but multiplies by two the computational cost of the

forecast step.

The model error can be introduced by the same means. If we denote by $(\Sigma^m)^{1/2}$ a Cholesky decomposition of the model error covariance matrix, and $\mathbf{s}_n^{i,m}$ its i-th column, the model error is incorporated in the last columns of S_n^f:

$$\left.\frac{\partial \mathbf{f}}{\partial \varepsilon^m}\right|_{\mathbf{x}_{n-1}^a} \mathbf{s}^{i,m} = \lim_{\epsilon \to 0} \frac{\mathbf{f}_{n-1}(\mathbf{x}_{n-1}^a, \epsilon \mathbf{s}^{i,m}) - \mathbf{f}_{n-1}(\mathbf{x}_{n-1}^a, 0)}{\epsilon} \tag{16}$$

$$\mathbf{s}_n^{r+i,f} \approx \frac{\mathbf{f}_{n-1}(\mathbf{x}_{n-1}^a, \epsilon \mathbf{s}^{i,m}) - \mathbf{f}_{n-1}(\mathbf{x}_{n-1}^a, 0)}{\epsilon}. \tag{17}$$

In the propagation step, the SEEK filter does not use any model error and instead amplifies previous error estimates by a *forgetting factor* (Pham et al., 1998). We find this approach dangerous since all the error analysis relies on the initial error. Since Pham et al. (1998) approximate the initial covariance by an EOF analysis of model generated states, the above reservations by Cane et al. (1996) also apply here.

Purely formal differences between the SEEK and the RRSQRT Kalman filters occur in the update step. In the RRSQRT KF, the number of vectors $\mathbf{s}_n^{i,f}$ is reduced at every step along the r dominant eigenvalues of the approximate covariance matrix and then each vector is processed by Potter's scalar update. In the SEEK, the reduction step is avoided by the use of the forgetting factor, then the matrix D_n^f is updated by the classical KF equations as the image of C_n^f in the reduced state space. These two methods are equivalent since they both simply translate the idea of least squares linear interpolation. However, the SEEK is less general than the RRSQRT since it assumes that model errors only amplify previous errors whereas the RRSQRT filter allows any kind of model errors and makes it possible to track their origin and design further improvements of the physical model. It should therefore be preferred for the easier interpretation of the results.

4 Ensemble Kalman filters

We review here adaptations of the KF to highly nonlinear systems by using Monte Carlo sampling in the propagation step while still applying a linear update. This is the principle of three DA methods, the Ensemble Kalman filter (EnKF) (Evensen, 1994 ; Evensen, 2003), the Harvard scheme (Lermusiaux and Robinson, 1999a) and the Singular Evolutive Interpolated Kalman (SEIK) filter (Pham, 2000) presented as an improved version of the SEEK filter. These filters are the most widely used in oceanography. An extension of the EnKF algorithm is proposed here with the use of a distributional transformation, called an *anamorphosis* in geostatistics (Chilès and Delfiner, 1999). It is illustrated with a simple ecological model.

4.1 Principle

We use the previous nonlinear model (9), the EnKF propagation step uses a Monte Carlo sampling to approximate the forecast density $\phi(\mathbf{x}_n|\mathbf{y}_{0:n-1})$ by r realizations:

$$\forall i = 1:r,\ \mathbf{x}_n^{f,i} = \mathbf{f}_{n-1}(\mathbf{x}_{n-1}^{a,i}, \varepsilon_n^{m,i}). \tag{18}$$

The Monte Carlo method does not require the linearization of $\mathbf{f}$. Thus the algorithm becomes much simpler than the EKF and its derivatives. The r realizations at time $n-1$ are processed and r simulations of the model error $\varepsilon_n^{m,i}$ are provided by classical tools. The computational cost of the propagation is directly proportional to the number r of realizations (also called "members" of the ensemble). Therefore, DA for oceanographical applications cannot afford much more than a hundred realizations, which is very low compared to many Monte Carlo methods that require an order of ten thousands realizations to produce stable results. However, simulated case studies did not show the interest of using more than a hundred members (Evensen, 1994), and the EnKF with only $r = 100$ members has been successfully applied in many real cases.

The update step simulates the posterior density $\phi(\mathbf{x}_n|\mathbf{y}_{0:n})$ by conditioning each forecast

member (18) to the new observations $\mathbf{y}_n$ by a linear update. The EnKF is therefore optimal for Gaussian prior and observation distributions, resp. $\phi(\mathbf{x}_n|\mathbf{y}_{0:n-1})$ and $\phi(\mathbf{y}_n|\mathbf{x}_n)$. In practice, the mean and covariance of the prior are approximated from the ensemble statistics $\mathbf{x}_n^f = 1/r \sum_i \mathbf{x}_n^{f,i}$ and $C_n^f = \frac{1}{r-1} \sum_i \mathbf{x}_n^{f,i}(\mathbf{x}_n^{f,i})^\top - \mathbf{x}_n^f(\mathbf{x}_n^f)^\top$ and r simulations of the observational noise $\varepsilon_n^o \sim \mathcal{N}(0, \Sigma^o)$ are also easily obtained. The kriging linear interpolator conditions the forecast members:

$$\mathbf{x}_n^{a,i} = \mathbf{x}_n^{f,i} + K_n(\mathbf{y}_n - H\mathbf{x}_n^{f,i} + \varepsilon_n^{o,i}), \tag{19}$$

with the classical Kalman gain matrix K_n as in (7). The three filters reduce the computational cost of the above update scheme by performing the update only along the dominant eigenmodes of the covariance matrix C_n^f (Evensen, 2003). They have been applied to a broad range of coupled physical and biological models (Allen et al., 2002 ; Lermusiaux and Robinson, 1999b ; Evensen, 2003) and have proven better adequation for highly nonlinear systems than the various suboptimal schemes of the EKF (Evensen, 1994 ; Verlaan and Heemink, 1999 ; Bertino, 2001).

4.2 Limits of the sequential linear estimator

When applying sequential DA to the nonlinear model (9) the successive state estimates of $\mathbf{x}_n$ have to satisfy conditions both of physical and statistical order:

- The state vector $\mathbf{x}_n^a$ produced by statistical analysis should be physically consistent to be further propagated into the model $\mathbf{f}$. The latter is designed to restart only from valid physical states and might crash or give non-physical results if this condition is not respected.

- The forecast state vector $\mathbf{x}_n^f$ generated by a perturbed model run should have multivariate Gaussian distribution for optimal use of the linear statistical analysis. However, the analysis will still produce a variance-minimizing result in the opposite case, though not being the optimal estimate.

In sequential DA, both conditions are linked: failure to satisfy one condition actually increases the risks of missing the other at the following step. The linear estimation in (19) is sensitive to extreme values in the data $\mathbf{y}_n$, but also to extreme values in the ensemble of forecast states $\mathbf{x}_n^{f,i}$ from which the prior covariance C_n^f is computed. In the particular case of distributions with a longer distribution tail than the normal, the posterior members $\mathbf{x}_n^{a,i}$ can contain extremely high or low values, depending on the sign of the covariances, that may be physically unrealistic. As another effect of an unappropriate use of a linear estimator, the posterior distribution is nearer to the normal distribution than that of the original variables (Miller et al. (1999) provides a simplified example in the case of sequential DA). The normal distribution, despite the name, does not fit all ocean variables. For example it has infinite distribution tails that many physical variables do not have: concentrations of water constituent are positive and the water temperature should always lie between the freezing and boiling points. Estimates that exceed these physical thresholds should not be processed further in the next propagation step as they could start disruptions in the physical model. The idea that the ocean model should be able with time to correct the non-physical features produced by the statistical analysis is dangerous. If sophisticated ocean models are able to simulate mesoscale processes, for example eddies that cross large ocean basins without losing their energy nor exchanging their contents with the surrounding waters, then they might also maintain the non-physical features for a very long time.

Until the present paper, there has been little concern so far in the DA community about the lack of Gaussianity of the forecast variables, although it is known that the variance minimizing estimate is no longer optimal in the non-Gaussian case. Our point is that the physical variables under consideration do not need to be Gaussian but only need to be subject of a Gaussian transformation in order to benefit from the good properties of this distribution. Still, we need a hypothesis that the random function is Gaussian after transformation of the marginal distribution. In particular, we assume that the bivariate distribution of the random function are bi-Gaussian.

If the above reservations apply to all DA methods mentioned in Sections 3 and 4.1, only the EnKF will be used to propose an alternative method since it provides a more convenient framework for the use of nonlinear transformations than most other methods.

4.3 The Gaussian anamorphosis function

The case of Gaussian anamorphosed variables allows to apply the two above mentioned conditions to two distinct variables, $\tilde{\mathbf{x}}_n = \psi_n(\mathbf{x}_n)$, ψ_n being the anamorphosis function, the raw variable $\mathbf{x}_n$ being physically consistent and the transformed variable $\tilde{\mathbf{x}}_n$ suitable for linear estimation. This is less constraining than expecting the same variable to have both physical and statistical properties together.

However the objective computation of the transformation ψ_n leads to practical difficulties in oceanographical applications because of the high dimensions of the state variables and of the operational aspects that need an automatic procedure. We are therefore looking for nonlinear bijections ψ_n defined on $\mathbb{R}^N$ that can turn the state vector into a Gaussian vector and, after the analysis is performed, back-transform the updated vectors into physically consistent states, at every assimilation time step n. Unfortunately this is not possible in practice. Indeed, evaluating the physical consistency of a given ocean state is not a straightforward task, neither checking for the Gaussiannity of a multivariate spatial law. The latter would formally require an order of fact(N) tests at each step n to check that all linear combinations of subsets of $\mathbf{x}_n^f$ are Gaussian. We shall therefore not look for the best transformation but rather concentrate on those that respect simpler conditions while already improving the original method.

The usual assumption is that the variables at different locations are identically distributed, conditionally to the past data and to the physics. Even if this assumption cannot be met, the variable minimum and maximum values remain the same all over the domain and we can build a transformation that respects these bounds. Within this framework the anamorphosis function ψ_n

is homogeneous all over the spatial domain and it is sufficient to model one distribution for each physical variable, which amounts to an order of ten instead of N. The modeling of the marginal distributions is computationally tractable for use in operational methods.

In practice, the Gaussian anamorphosis function can be obtained from the empirical marginal distribution by its expansion into Hermite polynomials. Chilès and Delfiner (1999) and Wackernagel (2003) give accounts of other possible anamorphosis functions, corresponding bivariate distribution models and their applications in nonlinear geostatistics. At time n, the propagation step (18) provides an ensemble of states $\{\mathbf{x}_n^{f,i}, i = 1 : r\}$ ideally drawn from the prior density $\phi(\mathbf{x}_n|\mathbf{y}_{1:n-1})$. Under the previous assumption, the density of each variable can be approximated by the histogram of their values all over the domain and for all members. In typical EnKF applications in oceanography, the number of grid cells times the number of members is of an order of 10^7 samples, which is sufficient for modeling graphically the prior density of $\mathbf{x}_n^f$.

4.4 An EnKF variant with Gaussian anamorphosis

We suggest a modification of the EnKF algorithm above. All other sequential DA methods can be adapted similarly if the update is performed by Monte Carlo sampling. The propagation step (18) remains unchanged since the law of ε_n^m and the nonlinear relationship between ε_n^m and $\mathbf{x}_n$ are both predefined and can already integrate any nonlinear transformation. The anamorphosis is used in the update step.

The observation density $\phi(\mathbf{y}_n|\mathbf{x}_n)$ is predetermined and therefore easy to transform. We denote χ_n the anamorphosis function for the measurements: $\tilde{\mathbf{y}}_n = \chi_n(\mathbf{y}_n)$. χ_n is related to ψ_n through the observation operator H. In practice it can be chosen so that the operator $\tilde{H} = \chi_n \circ H \circ \psi_n$ is linear. Then the change of variables redefines the nonlinear state space model (9) for the transformed variables $\tilde{\mathbf{x}}_n$ and $\tilde{\mathbf{y}}_n$, to which we can apply sequential DA. The classical KF matrices associated to the transformed variables will be noted $\tilde{C}$, $\tilde{K}$, $\tilde{\varepsilon}^o$, $\tilde{\Sigma}^m$ and $\tilde{\Sigma}^o$.

The first and second moments of the Gaussian prior density $\phi(\tilde{\mathbf{x}}_n|\tilde{\mathbf{y}}_{0:n-1})$ are approximated after the forward transformation $\tilde{\mathbf{x}}_n^{f,i} = \psi_n(\mathbf{x}_n^{f,i})$ by their experimental moments on the ensemble. The covariance of the transformed Gaussian observations $\phi(\tilde{\mathbf{y}}_n|\tilde{\mathbf{x}}_n)$ is deduced from $\phi(\mathbf{y}_n|\mathbf{x}_n)$, so we can simulate the transformed observation error $\tilde{\varepsilon}_n^o \sim \mathcal{N}(0, \tilde{\Sigma}^o)$.

In order to apply a linear update to $\tilde{\mathbf{x}}_n$, a second assumption has to be made about its multivariate Gaussian properties. Indeed the bivariate distribution of a couple of Gaussian variables is not necessarily bi-Gaussian but may be only Hermitian, and so on at higher orders. Here we will suppose $\tilde{\mathbf{x}}_n$ multi-Gaussian, but as previously mentioned, this assumption is practically very difficult to check. Yet we believe that the anamorphosed variables are more likely to fulfill it than the original physical variables $\mathbf{x}_n$.

The posterior density $\phi(\tilde{\mathbf{x}}_n|\tilde{\mathbf{y}}_{0:n})$ is simulated using the conditional simulation algorithm (19) applied to the transformed vectors :

$$\tilde{\mathbf{x}}_n^{a,i} = \tilde{\mathbf{x}}_n^{f,i} + \tilde{K}_n(\tilde{\mathbf{y}}_n - \tilde{H}\tilde{\mathbf{x}}_n^{f,i} + \tilde{\varepsilon}_n^{o,i}), \tag{20}$$

Back-transformation of the above conditional simulations $\mathbf{x}_n^{a,i} = \psi_n^{-1}(\tilde{\mathbf{x}}_n^{a,i})$ provides samples of the posterior probability density $\phi(\mathbf{x}_n|\mathbf{y}_{0:n})$ from which we can draw the unbiased statistics of interest, which would be difficult to obtain without performing the update by Monte Carlo sampling. The latter samples sequentially feed the physical model in the next ensemble propagation step. For application with other Kalman filtering methods, the analyzed mean $\mathbf{x}_n^a$ and covariance matrix C_n^a can be approximated by the ensemble average and covariance of $\mathbf{x}_n^{a,i}, i = 1 : r$.

4.5 Illustration: a simplified ecological model

We consider for **f** the simplified ecological model introduced by Evans and Parslow (1985) to reproduce yearly cycles of phytoplankton blooms. Phytoplankton (P) feeds on nutrients (N) and is grazed by herbivores (H). Eknes and Evensen (2002) extended the original nondimensional model to 1-D (a vertical water column) and implemented the EnKF to show the potential use

of remotely sensed ocean color data for DA in ecological models. This model is also used to compare the EnKF to the RRSQRT Kalman filter (Bertino et al., 2002). The state vector $\mathbf{x}_n$ contains the N, P and H values on all vertical grid cells, and the observations $\mathbf{y}_n$ are a subset of all variables every fourth grid cell. More details of the model and DA setup are given in Bertino (2001). The yearly cycle simulated by the model is presented in Figure 1: at the beginning of the year, nutrients are present at the bottom of the water column and brought by diffusion to the top of the water column. During the spring (days 100 to 150) the solar radiative energy lets the phytoplankton take up the nutrients; the "spring bloom". Their growth is stopped after day 150 by the herbivores grazing.

(Figure 1 around here)

This simplified ecological model has some interesting characteristics for DA: it is nonlinear, very sensitive to perturbations in the initial and boundary conditions and all three water constituents need positive values. Negative concentrations would cause disruptions in the biological equations. Here we will simply show the interest of modeling the water constituents by positive random variables with a lognormal model.

The twin experiment consists in generating a reference "true" yearly cycle by the model, from which we subsample a few synthetic observations perturbed by a random measurement noise. We assume that the measurements have lognormal distributions. The DA methods are then tested with respect to their ability in regenerating the reference solution. The model is driven by noisy initial and boundary conditions but corrected every 10 days by DA of the noisy synthetic observations. The ensemble size for the normal and lognormal versions of the EnKF will be $r = 100$, a larger ensemble size does not improve significantly the estimation (Bertino, 2001).

We present the results of the EnKF as in equations (18) and (19) that assumes all error noises Gaussian: the initial distribution of $\mathbf{x}_0$, the model errors ε_n^o and measurement errors ε_n^o. We

compare these results to those of the lognormal version (equation (20) with $\tilde{\mathbf{x}}_n = \log(\mathbf{x}_n)$ and $\tilde{\mathbf{y}}_n = \log(\mathbf{y}_n)$). In Figure 2 we represent the three histograms of the raw biological variables modeled in the course of the reference run, for all depths and all time steps, while the histograms of their logarithms are shown in Figure 3 for comparison and tend to indicate that the Gaussian assumption should be more appropriate for the logarithms than for the raw biological variables. The raw histograms are indeed more strongly asymmetric than the logarithms but the logarithm might not be the optimal transformation.

(Figure 2 about here) (Figure 3 about here)

Both filter error statistics have been set arbitrarily to realistic values and adjusted by a rule of thumb so that both approaches give equivalent results. We therefore briefly comment on their qualitative behavior.

(Figure 4 about here)

The errors in Figure 4 show no significant differences between the two versions for the N variable, but differences are obvious both for P and H. The results of the normal EnKF show repeated episodes with underestimation of P when overestimating H (see the stripes on middle left and bottom left graphs in Figure 4) during the whole spring bloom.

This result can be interpreted as an effect of negative simulated values. During the spring, light conditions are good enough to let phytoplankton and herbivores grow, however the growth of herbivores is limited by their own low concentration. The update in the normal EnKF simulates Gaussian posterior distributions with a mean near to zero and generates many negative values. As these values are not allowed to feed the ecological model they are cleared from the analysed vectors and set to zero. The truncation biases the H estimates by a systematical overestimation. During the following propagation step, the model lets the falsely numerous herbivores graze the phytoplankton, thus increases the overestimation of H and jointly underestimates P. At the next assimilation time, the estimates are corrected to their observed value and the estimation error is

pulled back to zero. These repeated biases and corrections form the sequence of vertical stripes on the left graphs of Figure 4.

Looking at the results from the lognormal EnKF, most of the vertical stripes have disappeared from the spring bloom period. Other errors occur later on for P and for H, but are apparently not linked to each other. We can further improve the estimates by the use of a more appropriate anamorphosis function than the logarithm, based on fine modeling of the ensemble histograms of $\mathbf{x}_n \sim \phi(\mathbf{x}_n|\mathbf{y}_{0:n-1})$ and $\mathbf{y}_n \sim \phi(\mathbf{y}_n|\mathbf{x}_n)$ as discussed in Section 4.3. The above results have been reproduced accurately with three different realizations of the random measurement perturbations (Bertino, 2001).

5 Other Monte Carlo methods

We point towards alternative Monte Carlo methods for the update step. Importance sampling is used to sample from the posterior distribution given by the Bayes formula. The algorithm is made sequential and an additional resampling step is added to prevent algorithm degeneracy. These Sequential Importance Resampling (SIR) methods are more general than Kalman filtering since they do not need to assume Gaussianity but their use in real DA applications is still new — the only realistic application at this date is in van Leeuwen (2002) — while they are already current practice in the statistical community (Doucet et al., 2001).

5.1 The general nonlinear model

We consider here a fully nonlinear state space model:

$$\mathbf{x}_n = \mathbf{f}_{n-1}(\mathbf{x}_{n-1}, \varepsilon_n^m) \tag{21}$$

$$\mathbf{y}_n = \mathbf{h}(\mathbf{x}_n, \varepsilon_n^o). \tag{22}$$

It has been demonstrated (Evensen, 1994) that Monte Carlo sampling represents progress compared to linear methods to derive the prior density. From the above example, we can guess

that the use of a linear estimator in the analysis step is also problematic when variables assimilated are obviously non-Gaussian. In principle the filter density can be computed by the general Bayes formula:

$$\phi(\mathbf{x}_n|\mathbf{y}_{0:n}) = \frac{\phi(\mathbf{x}_n|\mathbf{y}_{0:n-1})\phi(\mathbf{y}_n|\mathbf{x}_n)}{\int_{\mathbb{R}^N} \phi(\mathbf{x}|\mathbf{y}_{0:n-1})\phi(\mathbf{y}_n|\mathbf{x})d\mathbf{x}}. \tag{23}$$

However, in the general form of equation (23) the densities are complex and have high dimensions. They cannot be computed directly except for simplistic cases or under Gaussian assumptions that lead to Kalman filtering methods. Alternative Monte Carlo sampling techniques have therefore been considered to sample from the posterior distribution. Among them, the use of importance sampling has recently gained attention in the signal processing community.

5.2 Sequential importance sampling

We first give the general principle of importance sampling of a given variable **x** before discussing the application to the posterior density from equation (23). Importance sampling allows computing statistics of a variable **x** having a complex density $\phi(\mathbf{x})$ by the means of an arbitrary density $\pi(\mathbf{x})$, easier to simulate, called the "importance function". For any integrable function $\mathcal{M}$, we get:

$$\begin{aligned} E_\phi(\mathcal{M}(\mathbf{x})) &= \int \mathcal{M}(\mathbf{x})\phi(\mathbf{x})d\mathbf{x} && (24)\\ &= \int \mathcal{M}(\mathbf{x})\frac{\phi(\mathbf{x})}{\pi(\mathbf{x})}\pi(\mathbf{x})d\mathbf{x} && (25)\\ &= E_\pi\left(\mathcal{M}(\mathbf{x})\frac{\phi(\mathbf{x})}{\pi(\mathbf{x})}\right). && (26) \end{aligned}$$

Then samples are taken from the density π: $\{\mathbf{x}^i \sim \pi(\mathbf{x}),\, i = 1:r\}$ and we can build an estimator of the statistics of interest based on equation (24) by weighting the samples by the importance weights $w^i = \phi(\mathbf{x}^i)/\pi(\mathbf{x}^i)$:

$$E^*_\phi(\mathcal{M}(\mathbf{x})) = \frac{1/r\sum_{i=1}^r \mathcal{M}(\mathbf{x}^i)w^i}{1/r\sum_{j=1}^r w^j} = \sum_{i=1}^r \tilde{w}^i\mathcal{M}(\mathbf{x}^i), \tag{27}$$

with the normalized importance weights $\tilde{w}^i = \frac{w^i}{\sum_{j=1}^{r} w^j}$.

By selecting the identity function for $\mathcal{M}$, equation (27) provides an estimate of the expectation of $\mathbf{x}$. We obtain all other statistics with different choices of $\mathcal{M}$. The estimate in (27) is asymptotically unbiased under weak assumptions on the importance function (Doucet et al., 2001). A nice feature of the algorithm is that the speed of convergence does not depend on the state space dimension.

The importance sampling is applied to the filter density $\phi(\mathbf{x}_n|\mathbf{y}_{0:n})$ but can more generally be applied to the density $\phi(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$. It can be made adequate for recursive estimation under some assumption on the importance function (Doucet et al., 2001). The resulting algorithm is then called *Sequential Importance Sampling* (SIS). A common choice of the importance function, also followed by van Leeuwen (2002), is the prior distribution $\phi(\mathbf{x}_n|\mathbf{x}^i_{n-1})$ - simply obtained by model propagation. Then the weights of the samples are modified, but not the sample values themselves so that they can be fed into the physical model at next step without losing any physical property. Following the Bayes equation, the normalized importance weights become:

$$\tilde{w}^i_n \propto \tilde{w}^i_{n-1} \frac{\phi(\mathbf{x}^i_n|\mathbf{y}_n)}{\phi(\mathbf{x}^i_n|\mathbf{x}^i_{n-1})} \propto \tilde{w}^i_{n-1} \phi(\mathbf{y}_n|\mathbf{x}^i_n). \tag{28}$$

The above importance function is not necessarily the best one, especially when the prior and posterior probability densities do not have a significant overlap. This happens when the model error is incorrectly specified. Then the SIS computes the relative weights of ensemble states that are all unlikely (*i.e.* they all have a low likelihood in $\phi(\mathbf{y}_n|\mathbf{x}^i_n)$). Another extreme case is when the model error is assumed excessively large. Then the prior ensemble states can be simulated very far off the domain of likely states. It needs a lot more realizations to get a chance to hit this domain and the SIS algorithm is computationally inefficient. Zaritskii et al. (1975) proved that the optimal importance function is the conditional density $\phi(\mathbf{x}_n|\mathbf{x}^i_n, \mathbf{y}_n)$. However in the oceanographical studies the latter distribution is difficult to simulate, especially because

the simulations conditioned on $\mathbf{y}_n$ have to preserve some physical properties. A possible way to obtain conditional simulations that respect the physical properties of the system could be based on the ensemble Kalman filter techniques as in equation (19), or its variant with anamorphosis (equation 20). The conditioning by means of a Gibbs sampler has apparently not been attempted yet.

Whatever the importance function, the SIS algorithm generally becomes inefficient because the weights tend to be extremely unbalanced. After a large number of steps, only one member receives the full weight and all the others are uselessly propagated with zero weight. This problem is called the "degeneracy" of the filter by analogy with genetics.

5.3 Resampling

To avoid the degeneracy of the SIS algorithm, a resampling step can be added after the sampling (Doucet et al., 2001). The most common procedure is a bootstrap: at each step, the discrete distribution is resampled according to the importance weights $\tilde{w}_n^i$. It eliminates the members (called "particles" in the SIS/SIR literature) having low importance weights and multiplies those having high importance weights. It is possible to work with a constant number of particles r. The drawback is that it requires a large number of samples to make sure that all the particles with heavy weights will be well represented after resampling. This requirement is serious if we consider applications in oceanography that demand computational efficiency.

Variants of the bootstrap are then used and recent efforts in the field of sequential importance resampling (SIR) are reviewed in Doucet et al. (2001). Among these SIR filters, a computationally efficient filter imposes a minimum number of replications of the particles with high weights. The number of replications is r times the normalized importance weight $\tilde{w}_n^i$, rounded to its integer part. The remainders are resampled randomly, which may lead to an increase of the number of samples. The above procedure is also discussed in van Leeuwen (2002) and applied

both to a simple test equation and a real physical ocean model. 500 particles have been used in the latter case. Other variants use a fixed kernel method to perturb the distributions after resampling (Miller et al., 1999) but the method has only been tested for low dimensional models. The authors doubted the feasibility of the approach for real case studies but the encouraging results in van Leeuwen (2002) should draw more interest on SIR methods. The most promising resampling method seems to be a tree-based approach, as suggested in Doucet et al. (2001). All SIR algorithms are suitable for parallel implementation and will certainly benefit from progresses in computational techniques.

The idea of updating the probability density by the general Bayes formula has been discussed in the scope of oceanographical applications (van Leeuwen and Evensen, 1996 ; Miller et al., 1999 ; Anderson and Anderson, 1999 ; Pham, 2000), but only demonstrated on low dimensional problems such as the classical Lorenz equations. Applications to high-dimensional systems are still recent (van Leeuwen, 2002).

6 Conclusion

Sequential Data Assimilation has been built within the physical modeling community during the last twenty years. The physical models benefit from rapid improvements and they are now very performant and complex codes. By comparison the stochastic models used at the interface between the physical model and the data remained strikingly poor. This contrast can be explained by the fact that stochastic models are invisible in the setup of data assimilation: for example the need for Gaussian variables is not obvious at the sight of the Kalman filter equations. Therefore the optimal estimators minimizing a mean quadratic error are often believed to be absolute optima, whereas they are only optimal with respect to a simple linear stochastic model.

In the present review, we classified the current sequential data assimilation methods by classes of stochastic model to which they apply. In practice, the most rudimentary assimilation

methods (namely Optimal Interpolation) are often preferred for their low CPU cost. However, the simplicity of the stochastic model has a counterpart in the huge efforts required to ensure that the so-called optimal estimate is a valid physical solution. In our point-of-view, the lack of physical properties of the estimates is a consequence of the inconsistency between the stochastic and the physical model. This explains the paradox that simple data assimilation methods become extremely complex once applied to real operational forecasting. We therefore advise the use of methods that refer to a stochastic model of sufficient complexity compared to the physical model, both for the scientific interest and the practical implementation of the experiment. For example, the comparison of the extended Kalman filter to the ensemble Kalman filter in the case of nonlinear dynamics illustrates how the use of Monte Carlo sampling in the propagation step simplifies both the algorithm and the interpretation of the results, and also removes the estimation bias in the nonlinear propagation.

Ensemble Kalman filters repeat the model propagations a large number of times and they are often criticised for their computational cost, but their use considerably alleviates the engineering tasks. They are therefore more general and portable tools than the more rudimentary methods. Similarly, the use of Monte Carlo sampling in the analysis step is certainly a progress since it opens data assimilation methods to nonlinear estimation techniques (estimation of Gaussian anamorphosed variables or sequential importance resampling). Nonlinear estimation ought to be considered in many applications where linear combinations of state vectors and observations do not necessarily make a valid physical state. The ongoing applications of data assimilation to marine ecological variables which densities differ strongly from the Gaussian make a natural experimentation field for nonlinear estimation methods. These methods will probably be commonly applied in the oceanographic community and ready to use in other fields studying strongly nonlinear dynamical processes.

Acknowledgments

The present work was performed in relation with the EC funded MAST III project *Pioneer* and has been partly supported by a grant of computer time from the Norwegian Supercomputing Committee (TRU). We wish to thank two anonymous reviewers for their constructive comments.

References

Allen, J., M. Eknes, and G. Evensen (2002). An ensemble Kalman filter with a complex marine ecosystem model: hindcasting phytoplankton in the Cretan sea. *Ann. Geophys. 20*, 1–13.

Anderson, J. and S. Anderson (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Weather Rev. 127*, 2969–2983.

Bennett, A. F., B. S. Chua, and L. M. Leslie (1996). Generalized inversion of a global numerical weather prediction model. *Meteorol. Atmos. Phys. 60*, 165–178.

Bertino, L. (2001). *Assimilation de données pour la prédiction de paramètres hydrodynamiques et écologiques : le cas de la lagune de l'Oder*. Ph. D. thesis, Ecole des Mines de Paris, Fontainebleau, ftp://cg.ensmp.fr/pub/theses/bertino.pdf.

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

Brasseur, P., J. Ballabrera-Poy, and J. Verron (1999). Assimilation of altimetric data in the mid-latitude oceans using the Singular Evolutive Extended Kalman filter with an eddy-resolving, primitive equation model. *J. Marine Syst. 22*, 269–294.

Brusdal, K., J. Brankart, G. Halberstadt, G. Evensen, P. Brasseur, P. J. van Leeuwen, E. Dom-

browsky, and J. Verron (2003). A demonstration of ensemble based assimilation methods with a layered OGCM. *J. Marine Syst.*. In print.

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

Cañizares, R., D. Madsen, H. Jensen, and H. J. Vested (2001). Developments in operational shelf sea modelling in danish waters. *Estuar. Coast. Shelf S. 53*(4), 595–605.

Cane, M., A. Kaplan, R. N. Miller, B. Tang, E. Hackert, and A. Busalacchi (1996). Mapping tropical Pacific sea level: Data assimilation via a reduced state space Kalman filter. *J. of Geophys. Res. 101*(C10), 22599–22617.

Chilès, J. P. and P. Delfiner (1999). *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Cohn, S. and R. Todling (1996). Appropriate data assimilation schemes for stable and unstable dynamics. *J. Meteorol. Soc. Jpn. 74*(1), 63–75.

Courtier, P., E. Andersson, W. Heckley, J. Pailleux, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, and M. Fisher (1998). The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Q. J. R. Meteorol. Soc. 124*(550), 1783–1808.

Daley, R. (1991). *Atmospheric Data Analysis*. Cambridge: Cambridge University Press.

Doucet, A., N. de Freitas, and N. Gordon (Eds.) (2001). *Sequential Monte Carlo methods in practice*. New York: Springer.

Eknes, M. and G. Evensen (2002). An ensemble Kalman filter with a 1-D marine ecosystem model. *J. Marine Syst. 36*, 75–100.

Elbern, H., H. Schmidt, O. Talagrand, and A. Ebel (2000). 4D-variational data assimilation

with an adjoint air quality model for emission analysis. *Environ. Modell. Softw. 15*(6-7), 539–548.

Evans, G. and J. Parslow (1985). A model for annual plankton cycles. *Biol. Oceanogr. 3*, 327–347.

Evensen, G. (1992). Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model. *J. of Geophys. Res. 97*(C11), 17905–17924.

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

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical inplementation. *Ocean dynamics*. In print.

Fukumori, I. and P. Malanotte-Rizzoli (1995). An appropriate Kalman filter for ocean data assimilation: An example with an idealized Gulf Stream model. *J. of Geophys. Res. 100*(C4), 6777–6793.

Gandin, L. (1963). *Objective Analysis of Meteorological Fields*. Leningrad: Hydrometeoizdat.

Gauthier, P., P. Courtier, and P. Moll (1993). Assimilation of simulated wind lidar data with a Kalman filter. *Mon. Weather Rev. 121*, 1803–1820.

Ghil, M. and P. Malanotte-Rizzoli (1991). Data assimilation in meteorology and oceanography. *Adv. Geophys. 33*, 141–266.

Heemink, A. and A. Segers (2002). Modeling and prediction of environmental data in space and time using Kalman filtering. *Stoch. Env. Res. Risk A. 16*(3), 225–240.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press.

Le Dimet, F. X. and O. Talagrand (1986). Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus 38A*, 97–110.

Lermusiaux, P. F. J. and A. R. Robinson (1999a). Data assimilation via error subspace statistical estimation, Part I: Theory and schemes. *Mon. Weather Rev. 127*(8), 1385–1407.

Lermusiaux, P. F. J. and A. R. Robinson (1999b). Data assimilation via error subspace statistical estimation, Part II: Middle atlantic bight shelfbreak front simulations and ESSE validation. *Mon. Weather Rev. 127*(8), 1408–1432.

Luong, B., J. Blum, and J. Verron (1998). A variational method for the resolution of a data assimilation problem in oceanography. *Inverse Problems 14*, 979–997.

Mardia, K., C. Goodall, E. Redfern, and F. Alonso (1998). The kriged Kalman filter. *Test 7*(2), 217–285.

Maybeck, P. (1979). *Stochastic Models, Estimation, and Control*. New York: Academic Press.

Miller, R. N., E. F. Carter, Jr., and S. T. Blue (1999). Data assimilation into nonlinear stochastic models. *Tellus 51A*, 167–194.

Miller, R. N., M. Ghil, and F. Gauthiez (1994). Advanced data assimilation in strongly nonlinear dynamical systems. *J. of Atmos. Sci. 51*, 1037–1056.

Natvik, L.-J. and G. Evensen (2003). Assimilation of ocean colour data into a biochemical model of the North Atlantic. Part 1. Data assimilation experiments. *J. Marine Syst.*. In print.

Nechaev, V. and M. Yaremchuk (1994). Applications of the adjoint technique to processing of a standard section data set: world ocean circulation experiment section S4 along 67 deg S in the Pacific ocean. *J. of Geophys. Res. 100*(C1), 875–79.

Ngodock, H. E., B. S. Chua, and A. F. Bennett (2000). Generalized inverse of a reduced grav-

ity primitive equation model and tropical atmosphere-ocean data. *Mon. Weather Rev. 128*, 1757–1777.

Pham, D., J. Verron, and M. Roubaud (1998). A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Marine Syst. 16*(3-4), 323–340.

Pham, D. T. (2000). Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Mon. Weather Rev. 129*(5), 1194–1207.

Segers, A. (2002). *Data assimilation in atmospheric chemistry models using Kalman filtering*. Ph. D. thesis, TU Delft, Delft, http://www.library.tudelft.nl/dissertations/.

Segers, A., A. Heemink, M. Verlaan, and M. van Loon (2000). A modified RRSQRT-filter for assimilating data in atmospheric chemistry models. *Environ. Modell. Softw. 15*(6-7), 663–671.

Thacker, W. and R. Long (1988). Fitting dynamics to data. *J. of Geophys. Res. 93*, 1227–40.

van Leeuwen, P. J. (2002). A variance-minimizing filter for large-scale applications. *Mon. Weather Rev.*. Submitted, available on ftp://ftp.phys.uu.nl/pub/leeuwen/publ/SIR.ps.

van Leeuwen, P. J. and G. Evensen (1996). Data assimilation and inverse methods in terms of a probabilistic formulation. *Mon. Weather Rev. 124*, 2898–2913.

van Loon, M. and A. Heemink (1997). Kalman filtering for nonlinear atmospheric chemistry models : first experiences. Technical report MAS-R9711, CWI, TU Delft, the Netherlands.

Verlaan, M. (1998). *Efficient Kalman filtering algorithms for hydrodynamic models*. Ph. D. thesis, TU Delft, Delft, http://ta.twi.tudelft.nl/users/verlaan/artikelen/thesis.ps.gz.

Verlaan, M. and A. W. Heemink (1997). Tidal flow forecasting using reduced rank square root filters. *Stoch. Hydrol. Hydraul. 11*(5), 349–368.

Verlaan, M. and A. W. Heemink (1999). Non-linearity in data assimilation applications: a practical method for analysis. *Mon. Weather Rev.*. submitted.

von Storch, H. and F. Zwiers (1999). *Statistical Analysis in Climate Research*. Cambridge: Cambridge University Press.

Wackernagel, H. (2003). *Multivariate Geostatistics* (3rd ed.). Berlin: Springer Verlag.

Wikle, C. K. and N. Cressie (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika 86*(4), 815–829.

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

Zaritskii, V. S., V. B. Svetnik, and L. I. Shimelevich (1975). Monte Carlo techniques in problems of optimal data processing. *Auto. Remo. Cont. 12*, 95–103.

Résumé

Nous recensons quelques développements récents de techniques d'assimilation séquentielle utilisées en océanographie, qui intègrent des observations spatio-temporelles dans des modèles numériques décrivant des dynamiques physiques et écologiques. Les aspects théoriques allant du cas simple d'une dynamique linéaire au cas général d'une dynamique non-linéaire sont examinées du point de vue géostatistique. Des méthodes usuelles dérivées du filtre de Kalman sont présentées en partant du cas le moins complexe au cas le plus général et des perspectives pour une estimation non-linéaire sont discutées. Nous présentons en outre une extension du filtre de Kalman d'ensemble au cas de variables ayant subi une transformation gaussienne et nous l'illustrons en utilisant un modèle écologique simplifié. Les méthodes exposées sont conçues

pour prédire dans une région géographique avec une haute résolution spatiale sous la contrainte pratique que les temps de calcul soient suffisamment courts pour obtenir une prédiction avant l'heure. Ainsi l'article se concentre sur des méthodes couramment utilisées et de grande efficacité calculatoire.

List of Figures

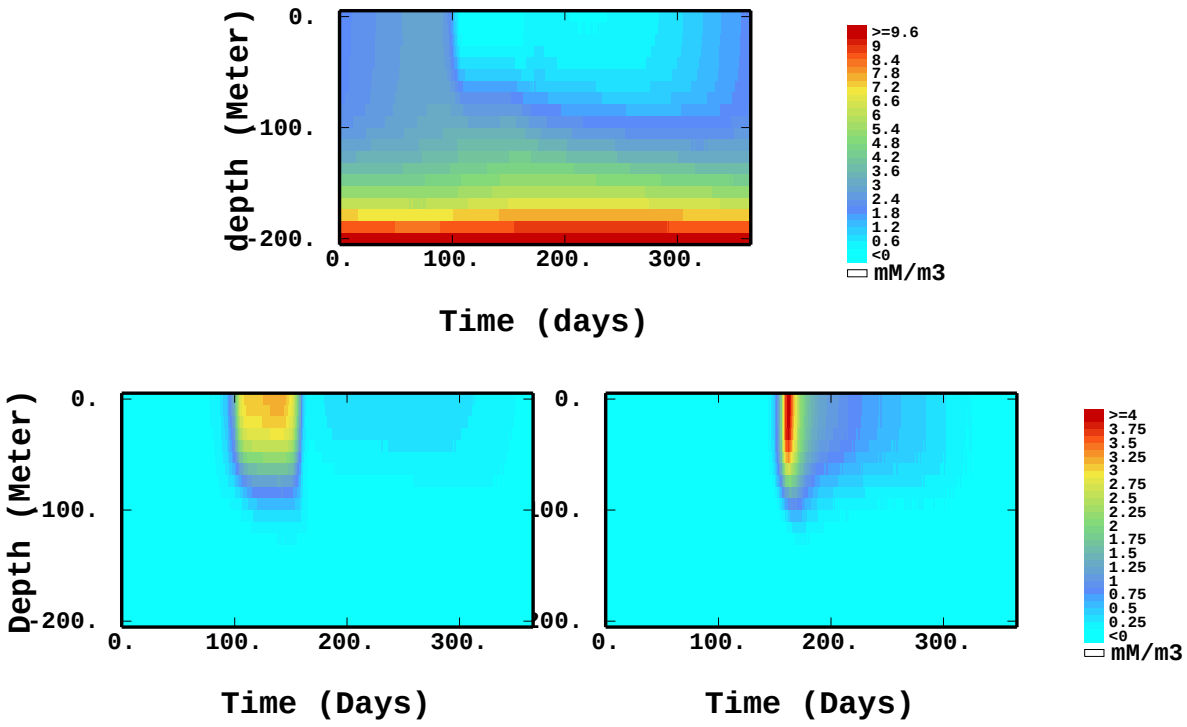

Figure 1: Yearly annual ecological cycle: temporal evolution of the Nutrients (top), Phytoplankton (bottom left) and Herbivores (bottom right) concentrations in a water column, the unit is the milimole of equivalent Nitrogen per cubic meter.

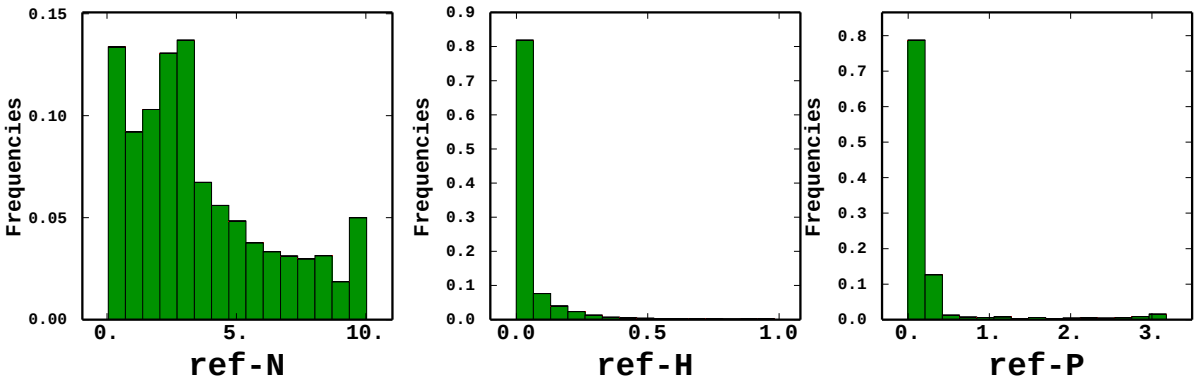

Figure 2: Histograms of the raw reference concentrations: Nutrients (left), Herbivores (middle) and Phytoplankton (right), modelled at all time steps of the yearly ecological cycle and all cells of the water column. All concentrations are positive.

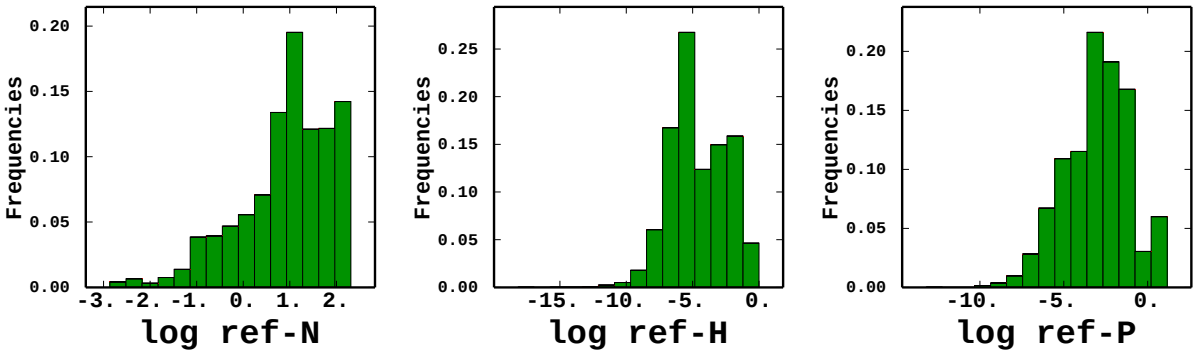

Figure 3: Histograms of the logarithms of the reference concentrations: Nutrients (left), Herbivores (middle) and Phytoplankton (right), modelled at all time steps of the yearly ecological cycle and all cells of the water column.

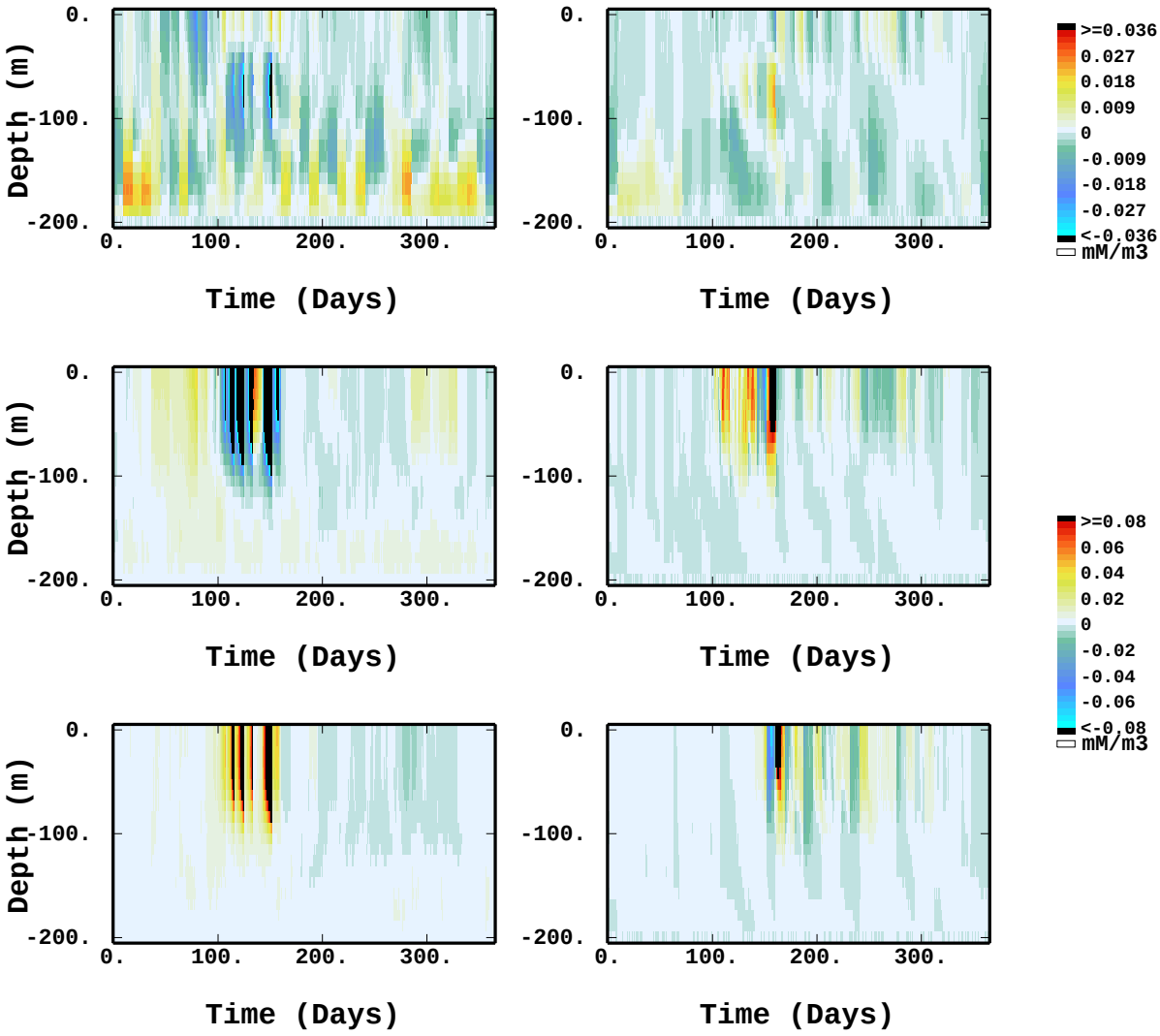

Figure 4: Forecast errors for variables N, P and H (resp. top, middle and bottom). Left: errors from the gaussian estimator EnKF. Right: errors from the lognormal anamorphosed EnKF, all "false starts" of spring bloom have disappeared. Note that the color scale for N has increased resolution since the errors are much lower than for P and H.