Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Marine Systems 40–41 (2003) 155–169
JOURNAL OF MARINE SYSTEMS
www.elsevier.com/locate/jmarsys

Assimilation of ocean colour data into a biochemical model of the North Atlantic Part 2. Statistical analysis

L.-J. Natvik*, G. Evensen

Nansen Environmental and Remote Sensing Center, Edv. Griegsvei 3A, N-5059 Bergen, Norway

Received 7 December 2001; accepted 9 August 2002

Abstract

In a companion paper [J. Mar. Syst. 40/41 (2003)], hereafter referred to as Part 1, we investigated an advanced data assimilation technique, the ensemble Kalman filter, for sequentially updating the biochemical state of a three-dimensional coupled physical–biochemical model of the North Atlantic. Within the methodology, an ensemble of model states is integrated forward to a measurement time, where an estimate based on information from both the model and the observations is calculated. The ensemble of states can provide estimates of any statistical moment, although moments of order three and higher are discarded in the analysis. In the Part 1 paper, we presented a simple demonstration experiment for the months April and May 1998, with some additional sensitivity tests at the first measurement time. The simulation included the early part of the spring bloom, which is characterized by strong nonlinear biochemical activity. It was concluded that the ensemble Kalman filter was able to provide an updated state consistent with the observations, and it was seen that the ensemble variance of the different biochemical components decreased during the analysis.

In this paper, we make some important remarks about linear versus nonlinear systems, emphasizing the fact that a data assimilation problem may become extremely complicated for strongly nonlinear problems. Statistical moments of any order may develop from Gaussian initial conditions during nonlinear evolution, and important information may be discarded by calculating an estimate based on only the Gaussian part of the full probability distribution. We demonstrate that a Monte Carlo approach can provide information about the system under consideration. For example, an ensemble of states, which is a representative of the true probability density function, can be visualized in one, two or three dimensions. Also, one can find estimates for the degree of nonnormality of the ensemble, which may act as indicators of the validity of performing a data assimilation based on the Gaussian part of the full probability distribution.

Keywords: Data assimilation; Biochemical model; North Atlantic

1. Introduction

In a companion paper (Natvik and Evensen, 2003, this issue), hereafter referred to as Part 1, we presented a data assimilation system for the biochemical part of a three-dimensional coupled physical–biochemical

* Corresponding author. Tel.: +47-55-29-72-88; fax: +47-55-20-00-50.
E-mail addresses: larsj@nersc.no (L.-J. Natvik), geir@nersc.no (G. Evensen).

0924-7963/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0924-7963(03)00017-4

model of the North Atlantic, utilizing real chlorophyll data from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS). An advanced multivariate methodology, the ensemble Kalman filter, was used to update the full discrete state vector of the biochemical model.

To describe the physics of the ocean, we used a version of the Miami Isopycnic Coordinate Ocean Model (MICOM), by Bleck et al. (1992). This model uses isopycnal coordinates in the vertical, i.e. the ocean interior is divided into a finite number of layers of constant and predefined potential density, and above them there is a single turbulent mixed layer of the Kraus–Turner type. The output from the ocean circulation model is used to force the 11-component Fasham–Ducklow–McKelvie (FDM)-type biochemical model, as described by Drange (1994, 1996). For more information on the coupled model, see Part 1.

The ensemble Kalman filter uses a finite ensemble of model states to represent the true probability density. Thus, by integrating each member due to the model dynamics, one can estimate any statistical moment from the ensemble. Note that since we use a finite number of members, an infinite number of appropriate ensembles exist. The methodology is sequential, i.e. the ensemble is integrated forward to a time where measurements are available, and the members are then updated due to an analysis scheme (see Part 1 for details).

In the Part 1 paper, we presented a simple demonstration experiment for April and May 1998, assimilating SeaWiFS chlorophyll data every 10th day. In addition, some sensitivity tests were done at the first assimilation time, i.e. at day 96. Note that the assimilation experiment includes the early part of the North Atlantic spring bloom. Since this is a period of great biochemical activity, nonlinear effects are probably relatively strong. Thus, the data assimilation problem is nontrivial. A first observation in the Part 1 paper was that the model system was not capable of producing a spatial distribution of phytoplankton consistent with the data at day 96. While the model predicted a large phytoplankton bloom in the eastern part of the North Atlantic basin, the SeaWiFS data showed large concentrations near the western boundary. There may be several reasons for the poor prediction, e.g. the model was set up with only two biochemical sublayers in the physical mixed layer, which may not be sufficient for resolving the ecosystem properly. Also, important biochemical dynamics may be missing, or the parameter values may not be optimal. A very positive result was that the ensemble Kalman filter analysis, assuming 35% errors for the observations and using 100 members in the ensemble, provided an estimate consistent with the data. Further, the ensemble variances of the different biochemical components decreased significantly during the analysis. The sensitivity with respect to the number of ensemble members was also investigated. By performing the ensemble Kalman filter analysis grid point by grid point horizontally (see Part 1), it is sufficient to ensure that we have reliable estimates of the covariances locally, i.e. within a defined influence radius for the measurements. We concluded that at least 60–80 members are needed. Finally, we studied the sensitivity with respect to the measurement errors; even with 60% errors, the poor model prediction was greatly improved during the analysis.

In the ensemble Kalman filter, a finite ensemble of model states is used to sample the probability density (see Part 1). Thus, instead of working with theoretical distributions, one can consider an ensemble containing any statistical information. Mardia (1970) presented various methods to extract statistical information from a general sample or ensemble. Further, Stephenson and Doblas-Reyes (2000) demonstrated how some of these methods could be used to monitor an evolving ensemble of meteorological forecast fields. In this paper, we illustrate the relevance of these methods for data assimilation, i.e. we show that by using an ensemble-based assimilation technique, one can obtain important information about the underlying statistics. For example, the validity of calculating an analysis based on the Gaussian part of the full probability distribution can be tested. Also, one can project the ensemble to a one-, two-, or three-dimensional space, i.e. to be able to visualize its evolution during the sequential assimilation. The experiment from the Part 1 paper will be used to illustrate the theory. Although we are studying data assimilation in a biochemical model, the methods should be of general interest for the data assimilation community.

We start by discussing linear versus nonlinear sequential estimation in Section 2. That is, we illustrate the fact that a data assimilation method may become extremely complicated for strongly nonlinear dynamics. Thus, the extension of the traditional

sequential Kalman filter for linear dynamics to nonlinear problems is nontrivial. Then, the theory to analyze an ensemble of states is outlined in Section 3, with application to the data assimilation experiment from the Part 1 paper. To be more specific, Section 3 is subdivided as follows. A simple approach to project the ensemble to a one-dimensional space is presented in Section 3.1. Further, a theoretical discussion of the covariances is given in Section 3.2, followed by a method to interpret ensemble member distances in a reduced space in Section 3.3. In Section 3.4, moment measures of skewness and kurtosis are used to monitor the degree of nonnormality of the ensemble. Finally, a summary is given in Section 4.

2. Nonlinear estimation and Gaussian versus non-Gaussian statistics

As in Part 1, let us define an n-dimensional vector $\boldsymbol{\psi}(t)\in\Re^n$, describing the state of the ocean at a particular time t. For our biochemical model, it is practical to let $\boldsymbol{\psi}$ contain all the discrete model compartments, i.e. one realization of the entire space discretized biochemical state. Note that a realization of $\boldsymbol{\psi}$ represents a single point in state space $\mathscr{P}\subseteq\Re^n$. A probability density function may be defined as

$$\phi(\boldsymbol{\psi}) = \frac{\mathrm{d}N}{N}, \tag{1}$$

where $\mathrm{d}N$ is the number of points per volume increment and N is the number of points altogether. Thus, $\phi(\boldsymbol{\psi})\mathrm{d}\boldsymbol{\psi}$ is the probability of a realization of the state located inside the volume element or n-ball $\mathrm{d}\boldsymbol{\psi}$ around the point $\boldsymbol{\psi}$. Note that instead of working with a theoretical distribution, one can represent it by defining a finite ensemble of model states (see Part 1).

For a linear model where the initial condition is taken from a Gaussian distribution, the probability density will remain Gaussian at any time (Jazwinski, 1970). Thus, in the linear case, one can find exact expressions for the evolution of the mean and covariance, and this approach is used in the standard Kalman filter (KF) (e.g. Bennett, 1992). However, note that a nonlinear model may cause the probability density function to become non-Gaussian during model evolution. In this case, a popular approach to get a closed system of equations has been to discard statistical moments of order three and higher in the equation for the evolving covariance. However, this approximation, which is used in the extended Kalman filter (EKF), has been shown to be unrealistic in many systems (see Part 1). Thus, the extension of the Kalman filter to nonlinear problems is nontrivial. Alternatively, one can integrate an ensemble of model states (e.g. EnKF), which will describe the correct statistics at any time in the limit of infinitely many members. For more information about the differences between KF, EKF and EnKF, please refer to Part 1.

At measurement times, the above methods all calculate an analysis using only the Gaussian part of the full probability density. Thus, the analysis will be less relevant for highly non-Gaussian statistics. One important property of ensemble-based methods is that one can extract any statistical information from the ensemble. For example, Mardia et al. (1979) introduced multivariate measures of skewness (third-order moment) and kurtosis (fourth-order moment), i.e. indicating the degree of nonnormality of an ensemble. In Section 3.4, we illustrate that this can be used as an independent check on the assimilation, i.e. to study the relevance of the EnKF analysis scheme.

3. Analyzing and monitoring the ensemble of model states

In this section, we demonstrate that ensemble-based methods can provide valuable information about the underlying statistics, and that the data assimilation experiments can be monitored in an effective way. The general theory can be found in Mardia (1970) and Mardia et al. (1979). Our work is also based on the paper by Stephenson and Doblas-Reyes (2000), who investigated an ensemble of 51 weather forecasts, concentrating on the height of the 500-hPa geopotential surface; we mainly follow their notation.

As said in Section 2, the state vector, which contains the 11 discrete biochemical variables, represents a single point in state space $\mathscr{P}\subseteq\Re^n$. It is not a straightforward task to extract the most relevant information from the huge amount of data in $\mathscr{P}$. For example, to visualize the ensemble of points, one will have to make a projection of $\mathscr{P}$ to some one-, two- or three- dimensional space.

3.1. State vector averages

One simple approach is to consider some kind of state vector average. For example, by taking the mean value of a variable over a specific domain, one can plot each ensemble member as a single point evolving with time. For this to be a true domain average, it is necessary to use the sizes of the grid cells as weights. As an even simpler alternative, one can consider grid point averages, which means that domains with a high resolution contribute more to the mean. Note that a state vector including several model variables makes it practical to work with dimensionless scaled variables. In this way, the contributions from different ecosystem components are of the same order. For convenience, we used scaled variables also when considering the mean of a single component. Fig. 1 shows the ensemble of averages of phytoplankton and nitrate over the horizontal grid in the surface layer, using the ensemble mean surface layer mean of each variable after the first assimilation (at day 96) as scaling factors. Thus, the ensemble mean of the grid averages of the analysis estimate at day 96 are exactly 1, as seen in the figure. As expected, it is observed that the ensemble of grid averages spreads out during model integration, while a convergence is seen at analysis times (the EnKF analyses are at days 96, 106, 116, 126, 136 and 147). Note also that the phytoplankton average ensemble experiences a higher degree of convergence at analysis times than the nitrate ensemble. This is also expected, since phytoplankton is the observed (measured) variable of the multivariate system. Further, the spread of both ensembles increases for each integration cycle throughout the entire experiment. This could indicate an unstable model. However, the duration of the experiment is very short and covers only the early part of the spring bloom, which is the most unstable period of the year for the biochemical system. Thus, to study the stability of the model, one would have to increase the time period of the experiment, preferably to a year or more.

The use of domain averages or state vector averages are very simple, but not necessarily very reliable indices of ensemble member distances in state space. Important information is probably lost by averaging each member into a single point in a reduced space. For example, one has not taken into consideration which directions represent the major axes of variability in state space $\mathscr{P}$, and different contributions may cancel each other when taking the average. A more thorough analysis is given in the following sections.

3.2. Covariance matrices

In this section, two different covariance matrices will be defined, one by averaging with respect to the ensemble members, i.e. the covariance used in the EnKF, and one by averaging with respect to the control variables, respectively (see below). To reduce the computational requirements, we only include a single variable in the state vector. A general state containing all the ecosystem components could be used, but one should then define dimensionless scaled variables. The following theory is based on Mardia et al. (1979) and Stephenson and Doblas-Reyes (2000).

Let i be an ensemble member counter and $\mathbf{x}_i$ a vector having one discrete model variable (one ensemble member) as its components. For example, let $\mathbf{x}_i^{\mathrm{T}}=(P_{i,1},\ldots,P_{i,n_g})$, where P is phytoplankton and the number of grid points $n_g = n_x n_y n_z$. The theory will be outlined using this state vector including phytoplankton only. However, the expressions should also be valid for some other ecosystem variable, or even for a general state vector containing all the (dimensionless) compartments. In the discussion in the following sections, it should be clear from the context whether we are speaking about a general state or some substate containing only one variable.

Since we are considering an ensemble of phytoplankton states in "phytoplankton state space" $\mathscr{P}_P$, it is convenient to define a matrix $\mathbf{X}$ containing the ensemble of phytoplankton states, i.e.

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_{n_{\mathrm{ens}}}^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} P_{1,1} & \cdots & P_{1,n_g} \\ \vdots & \ddots & \vdots \\ P_{n_{\mathrm{ens}},1} & \cdots & P_{n_{\mathrm{ens}},n_g} \end{pmatrix}, \qquad (2)$$

where n_{ens} is the number of ensemble members. The number of grid points is normally much larger than the number of ensemble members, and $\mathbf{X}$ has therefore normally many more columns than rows.

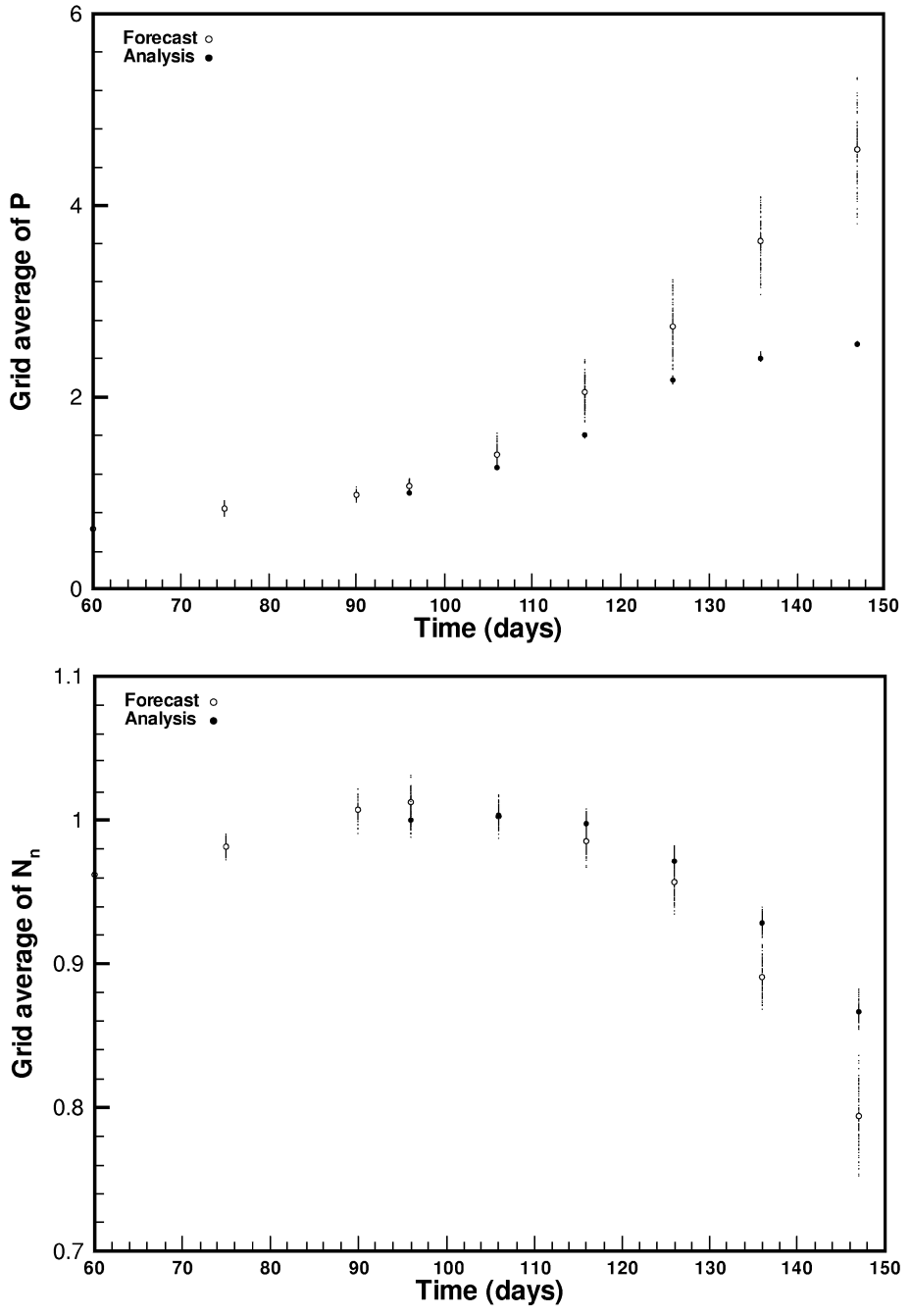

Fig. 1. Grid point average (over horizontal grid in surface layer) of phytoplankton (top) and nitrate (bottom) for each ensemble member. The large circles represent the ensemble mean for the forecast ensemble (open circle) and for the analysis ensemble (filled circle), respectively. Note that dimensionless variables are assumed, using averages of the analyzed variables at day 96, i.e. over the ensemble and surface grid, as scaling factors (see text).

The center of mass of the cloud of points in state space can be found by taking the average of the rows of **X**, i.e.

$$\bar{\mathbf{x}} = \frac{1}{n_{\text{ens}}} \sum_{i=1}^{n_{\text{ens}}} \mathbf{x}_i. \tag{3}$$

Further, by defining deviations from the ensemble mean, $\mathbf{y}_i = \mathbf{x}_i - \bar{\mathbf{x}}$, one can also form a matrix of deviations,

$$\mathbf{Y}^{\mathrm{T}} = (\mathbf{y}_1, \ldots, \mathbf{y}_{n_{\text{ens}}}) = \mathbf{X}^{\mathrm{T}}\mathbf{G}, \tag{4}$$

where $\mathbf{G} = \mathbf{I} - (1/n_{\text{ens}})\mathbf{1}$. Further, **I** is the identity matrix and all elements of **1** are 1.

By averaging with respect to the ensemble members, one can find estimates for covariances between the different control variables in state space. In our case, $P_1, \ldots, P_{n_g}$ are the control variables, and an $n_g \times n_g$ "control variable" or "state space" covariance matrix **C** (for phytoplankton) can be evaluated as follows:

$$\mathbf{C} = \overline{(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^{\mathrm{T}}} = \frac{1}{n_{\text{ens}}} \sum_{i=1}^{n_{\text{ens}}} \mathbf{y}_i \mathbf{y}_i^{\mathrm{T}}$$

$$= \frac{1}{n_{\text{ens}}} \mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \frac{1}{n_{\text{ens}}} \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X}, \tag{5}$$

where it has been used that $\mathbf{G} = \mathbf{G}^{\mathrm{T}}$ and $\mathbf{G}\mathbf{G}^{\mathrm{T}} = \mathbf{G}$. Thus, the state space covariance matrix can be expressed as an outer product of **Y** with itself. An unbiased estimate of the true covariance can be defined by dividing by $n_{\text{ens}} - 1$ instead of n_{ens}, i.e.

$$\mathbf{C}_{\mathrm{u}} = \frac{1}{n_{\text{ens}} - 1} \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X} = \frac{n_{\text{ens}}}{n_{\text{ens}} - 1} \mathbf{C}. \tag{6}$$

By averaging with respect to the control variables, one can find an $n_{\text{ens}} \times n_{\text{ens}}$ ensemble covariance matrix **B** (for phytoplankton), expressed as an inner product of **Y** with itself (Stephenson and Doblas-Reyes, 2000), i.e.

$$\mathbf{B} = \frac{1}{n_{\mathrm{g}}} \mathbf{Y}\mathbf{Y}^{\mathrm{T}} = \frac{1}{n_{\mathrm{g}}} \mathbf{G}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{G}. \tag{7}$$

The eigenvectors of **C** provide information about how the state space is spanned by the ensemble cloud, and about which spatial patterns contribute the most to the second-order statistical moments. Further, the eigenvectors of **B** provide information about which ensemble members contribute most to the variance in state space. Since the covariance matrices are symmetric, they must be orthogonally diagonalizable (Anton, 1991). That is, if we let **V** be an orthogonal matrix containing the eigenvectors of **C** as columns, the covariance can be diagonalized as (Stephenson and Doblas-Reyes, 2000)

$$\mathbf{V}^{\mathrm{T}}\mathbf{C}\mathbf{V} = \frac{1}{n_{\text{ens}}} \mathbf{D}_{\mathbf{C}} = \frac{1}{n_{\text{ens}}} \mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma}, \tag{8}$$

where $\mathbf{D}_{\mathbf{C}}$ is a diagonal matrix having the eigenvalues of $n_{\text{ens}}\mathbf{C}$ on the diagonal, and $\mathbf{\Sigma}$ is zero except on the diagonal containing the square root of the eigenvalues of $n_{\text{ens}}\mathbf{C}$. The diagonal matrix $\mathbf{D}_{\mathbf{C}}$ is $n_{\text{ng}} \times n_{\text{ng}}$, while $\mathbf{\Sigma}$ is $n_{\text{ens}} \times n_{\text{ng}}$, respectively. In a similar manner, the matrix **B** can be diagonalized as

$$\mathbf{U}^{\mathrm{T}}\mathbf{B}\mathbf{U} = \frac{1}{n_{\text{ng}}} \mathbf{D}_{\mathbf{B}} = \frac{1}{n_{\text{ng}}} \mathbf{\Sigma}\mathbf{\Sigma}^{\mathrm{T}}, \tag{9}$$

where $\mathbf{D}_{\mathbf{B}}$ contains the eigenvalues of $n_{\text{ng}}\mathbf{B}$ and **U** have the eigenvectors of **B** as columns, respectively. Note that $\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma}$ and $\mathbf{\Sigma}\mathbf{\Sigma}^{\mathrm{T}}$ have the same leading values on the diagonal, and therefore the two matrices $n_{\text{ens}}\mathbf{C}$ and $n_{\text{ng}}\mathbf{B}$ share the same leading eigenvalues.

The matrix $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ is a symmetric matrix. Further, it must be positive semidefinite, since for some vector $\boldsymbol{a}$, $\boldsymbol{a}^{\mathrm{T}}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})\boldsymbol{a} = (\mathbf{Y}\boldsymbol{a})^{\mathrm{T}}(\mathbf{Y}\boldsymbol{a}) \geq 0$, thus its eigenvalues must be nonnegative. A singular value decomposition of **Y** can be written as (Kincaid and Cheney, 1991; Stephenson and Doblas-Reyes, 2000)

$$\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}, \tag{10}$$

where the diagonal matrix $\mathbf{\Sigma}$ contains the singular values of **Y**, which are defined to be the square roots of the (nonnegative) eigenvalues of $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$.

The effective dimensionality of the ensemble is given by the rank of **Y**, which is bounded from above by $\min(n_{\text{ens}}, n_{\text{ng}})$ and typically equal to n_{ens}. If $\text{rank}(\mathbf{Y}) < n_{\text{ens}}$, some members can be removed from the ensemble without any loss of information on the second-order moment (they must be chosen such that $\text{rank}(\mathbf{Y})$ remains constant).

For later use, the eigenvalue spectrum of **B** is shown in Fig. 2. As discussed in Part 1, an initial ensemble was generated at day 60 by perturbing the layer interfaces in the model. Thus, an ensemble of identical biochemical states is generated initially, but the members take different paths during model integration due to the perturbed layers. A consequence is that **B** is singular at day 60 with one eigenvalue accounting for all the variance (not shown in the figure). It can be observed from Fig. 2 that the largest eigenvalues seem to be less dominant for the EnKF analysis ensemble than the forecast ensemble. This is expected, since the members of the former probably are closer. Note also that the largest eigenvalue of the forecast spectrum seems to become more dominant with time.

3.3. Multidimensional scaling

A map marking the relative positions of the ensemble members could be very useful for interpreting ensemble member distances. However, it is problematic to visualize such a map for more than three dimensions, i.e. some kind of projection must be implemented.

One simple method to view an ensemble in a reduced space is to use some kind of area or grid averaging (see Section 3.1). However, such an approach does not necessarily provide very accurate estimates of the relative positions, and more optimal methods should be searched for.

Multidimensional scaling (MDS) provides a very useful approach for projecting an ensemble of n_{ens} points onto a reduced one-, two- or three-dimensional space, which can be drawn on a flat piece of paper (Mardia et al., 1979; Stephenson and Doblas-Reyes, 2000). To be more specific, MDS is concerned with the problem of finding an optimal configuration of n_{ens} points in q-dimensional space, based on information of the distances between the elements in p-dimensional space. Thus, it can be used to monitor

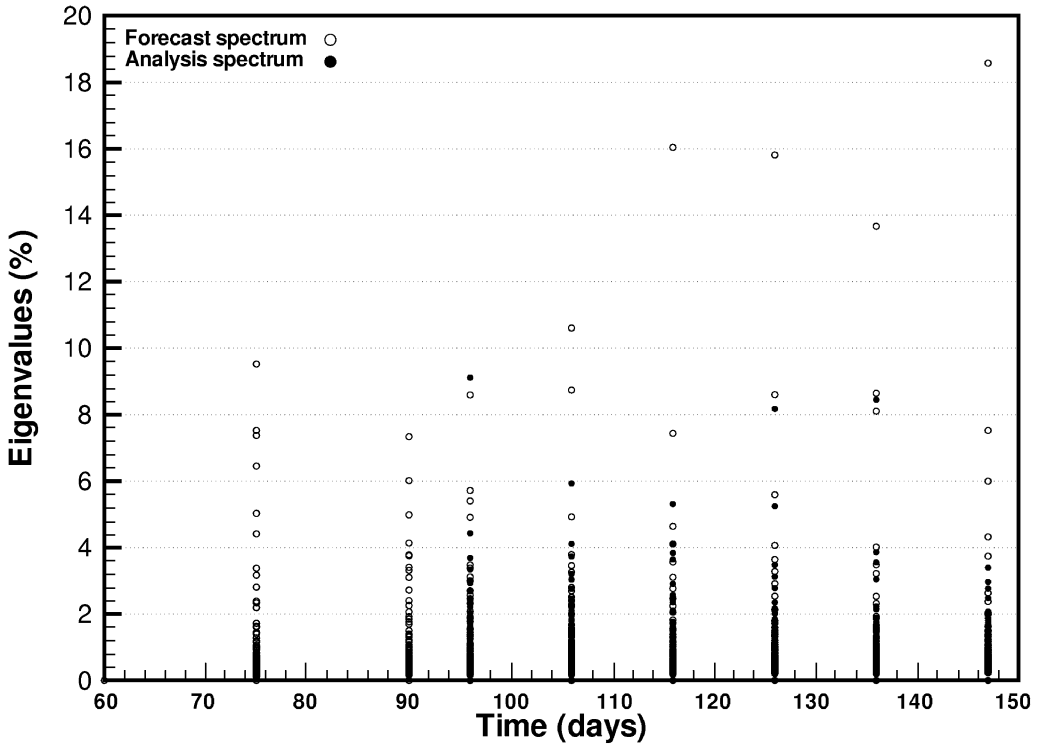

Fig. 2. The evolution of the eigenvalues of **B** (open circles are used for the forecast spectrums and filled circles for the analysis spectrums, respectively). Initially at day 60, all ensemble members are identical, and there is one eigenvalue (not shown) accounting for all the variance, see text for more information. Note that the first assimilation is at day 96.

the evolution of the ensemble member distances in one, two or three dimensions. Note that the solution produced by any MDS method is indeterminate with respect to reflection, rotation and translation. For Euclidean distances, the classical solution in q dimensions is given by the q leading principal coordinates of $\mathbf{X}$ (Mardia et al., 1979). For ensemble member i, they are given by (Stephenson and Doblas-Reyes, 2000)

$$\boldsymbol{p}_q(i) = \begin{pmatrix} \sigma_1 U_{i,1} \\ \vdots \\ \sigma_q U_{i,q} \end{pmatrix}, \tag{11}$$

where σ_i are the singular values of $\mathbf{Y}$ and $U_{i,l}$ is the ith element of the lth leading eigenvector of $\mathbf{B}$. Thus, for plotting the n_{ens} size ensemble in two dimensions ($q=2$), the ith ensemble member have the coordinates $(\sigma_1 U_{i,1}, \sigma_2 U_{i,2})$. If we let $\mathbf{D}$ be a matrix containing the distances between the ensemble members, then $\mathbf{D}$ is Euclidean if and only if the corresponding centered inner product matrix ($\mathbf{B}$) is positive semidefinite (Mardia et al., 1979). For some vector $\boldsymbol{a}$, we have $\boldsymbol{a}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\boldsymbol{a}=(\mathbf{Y}^{\mathrm{T}}\boldsymbol{a})^{\mathrm{T}}(\mathbf{Y}^{\mathrm{T}}\boldsymbol{a})\geq 0$, which means that $\mathbf{B}$ satisfies the positive semidefiniteness property. Note that the classical solution for Euclidean distances is an optimal projection, i.e. it is the best representation of the true distances in state space. Further, for Euclidean distances, the principal coordinates are proportional to the principal components of the state space covariance. However, the solutions are not simply related for non-Euclidean metrics (Stephenson and Doblas-Reyes, 2000). Finally, it should be mentioned that the ensemble mean will always be projected onto the origin in the reduced space. For a general and detailed description of multidimensional scaling, please refer to Mardia et al. (1979).

A two-dimensional map will be based on the two leading singular values of $\mathbf{Y}$ (or equivalently the square root of the eigenvalues of $n_{\text{ng}}\mathbf{B}$). From Fig. 2, it is seen that the two first eigenvalues of $\mathbf{B}$ account for about 13–26% of the total variance for the forecast spectrums and about 6–14% for the analysis spectrums, respectively. This means that a two-dimensional projection will be quite approximative. To improve the precision, one could use the first three principal coordinates in a three-dimensional plot. Also, one could study pairs (or triples) of principal coordinates simultaneously, i.e. in different directions.

The two first principal coordinates for each of the 100 members of the forecast ensemble at day 96 are shown in Fig. 3. By marking each member using a unique number, it is easy to follow the evolution of a specific member during an experiment. This also allows for identification of outliers, which one may want to study in more detail. An outlier should not be removed from the ensemble unless it contains an unacceptable state, e.g. negative biochemical concentrations, quantities diverging to infinity, and so on.

The evolution of the principal coordinates are shown in Figs. 4 and 5; the forecast ensemble is shown in the left plots and the analysis ensemble in the right, respectively. As explained in Part 1, an initial ensemble was generated by perturbing the (physical) layer interfaces at day 60, and the ensemble of initially identical biochemical states spreads out during model integration due to the individual layer distributions.

The evolution of the principal coordinates is as expected. It is seen that the ensemble spreads out during model integration, while it experiences a contraction during each ensemble Kalman filter analysis. As for the grid point averages in Section 3.1, it is seen that the forecast ensemble spread increases during the entire experiment. This must be expected, since our simulation only covers the early part of the spring bloom. Thus, this should be further investigated in a long-term experiment. Another observation is that the ensemble seems essentially Gaussian throughout the experiment, although the analysis at days 106 and 126 indicates some degree of skewness. Remember that only the two largest singular values are included for calculating the two-dimensional projection, and more reliable tests of multivariate skewness and kurtosis should be searched for. This will be further investigated in the next section. Outliers can have a large effect on the ensemble mean, which always is projected onto the origin. For example, in the analysis ensemble at day 136, the two outliers in the top left quadrant must be balanced by a number of other members. An interesting pattern is also seen in the analysis ensemble for day 116.

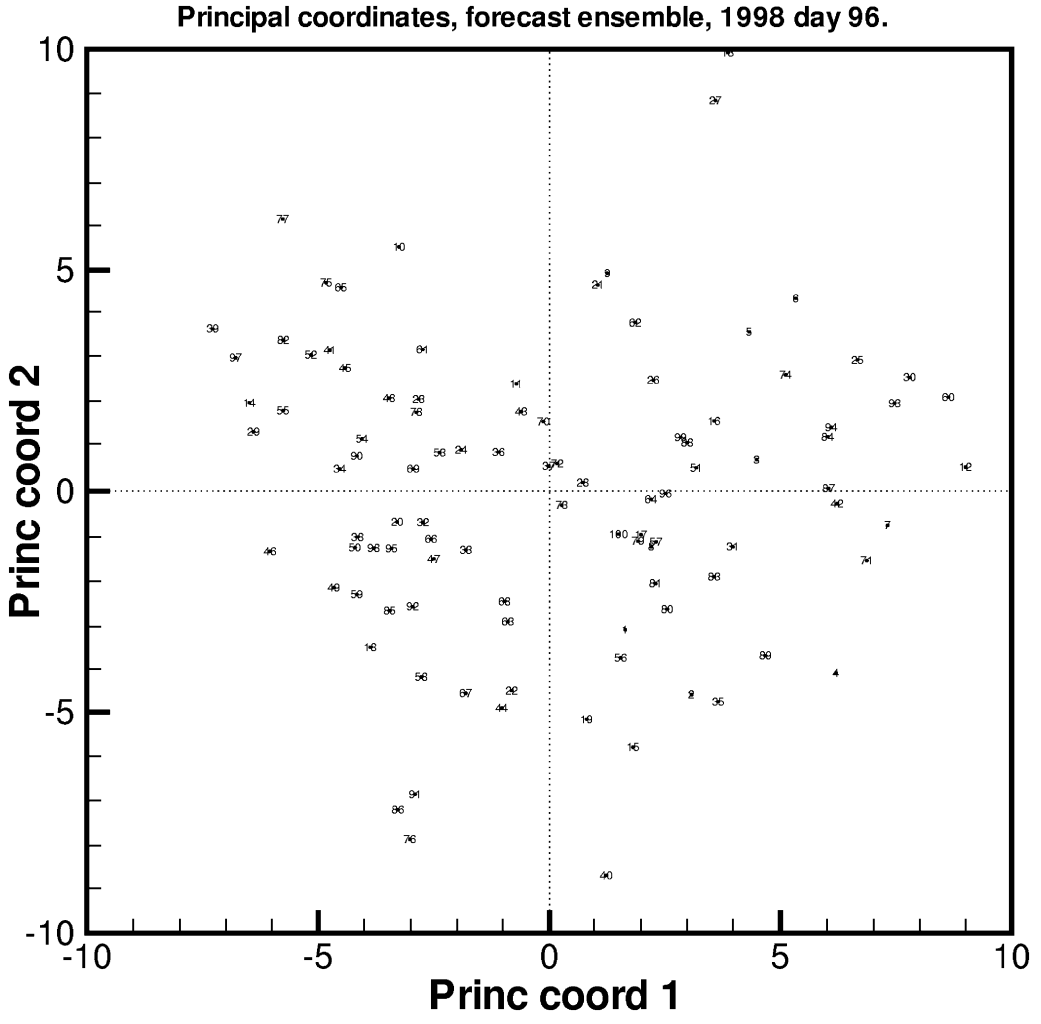

Fig. 3. The principal coordinates of the forecast ensemble members at day 96. Each member is marked by a unique number from 1 to 100.

3.4. Higher-order statistical moments

Even though the ensemble may be described by nonnormal statistics, only the Gaussian part is used in the EnKF analysis. Thus, if the true errors are dominated by statistical moments of order three or higher, the analysis estimate will not be optimal. In an ensemble-based method, the full probability density is sampled by a finite ensemble, which contains information about any statistical quantity. As shown below, this allows us to find estimates for the degree of nonnormality of an ensemble, and thus monitor the validity of performing the analysis based on only the Gaussian part of it.

Some simple measures of multivariate skewness and kurtosis (flatness) can be written as (Mardia, 1970; Mardia et al., 1979)

$$b_{\text{skew}} = \frac{1}{n_{\text{ens}}^2} \sum_{k=1}^{n_{\text{ens}}} \sum_{l=1}^{n_{\text{ens}}} g_{kl}^3 \tag{12}$$

$$b_{\text{kurt}} = \frac{1}{n_{\text{ens}}} \sum_{k=1}^{n_{\text{ens}}} g_{kk}^2, \tag{13}$$

where

$$g_{kl} = (\mathbf{x}_k - \bar{\mathbf{x}})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x}_l - \bar{\mathbf{x}}) = \mathbf{y}_k^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{y}_l. \tag{14}$$

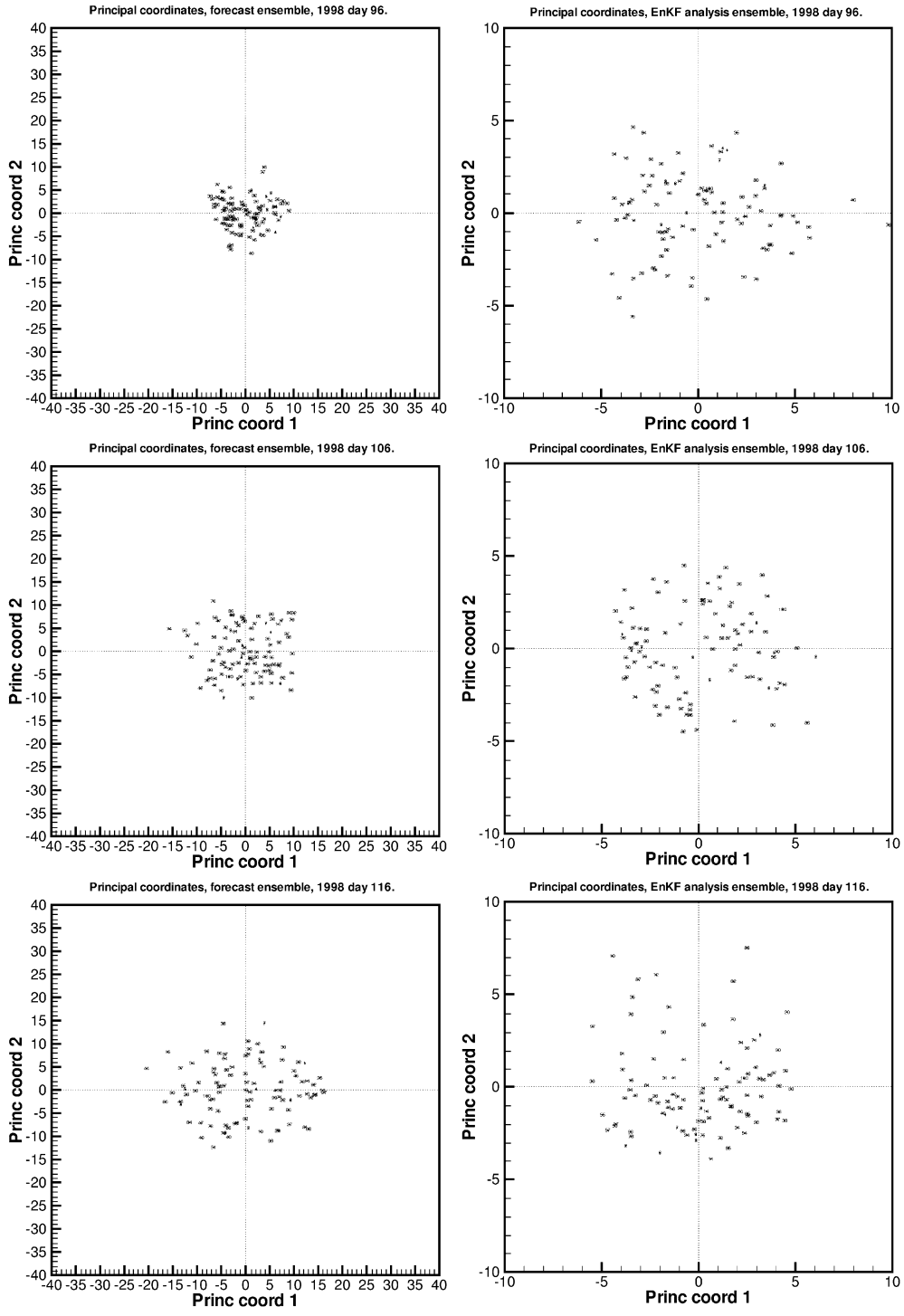

Fig. 4. Principal coordinates for the forecast ensemble (left) and for the analysis ensemble (right) at day 96 (top), day 106 (middle) and day 116 (bottom). Note the different scales on the left and right plots.

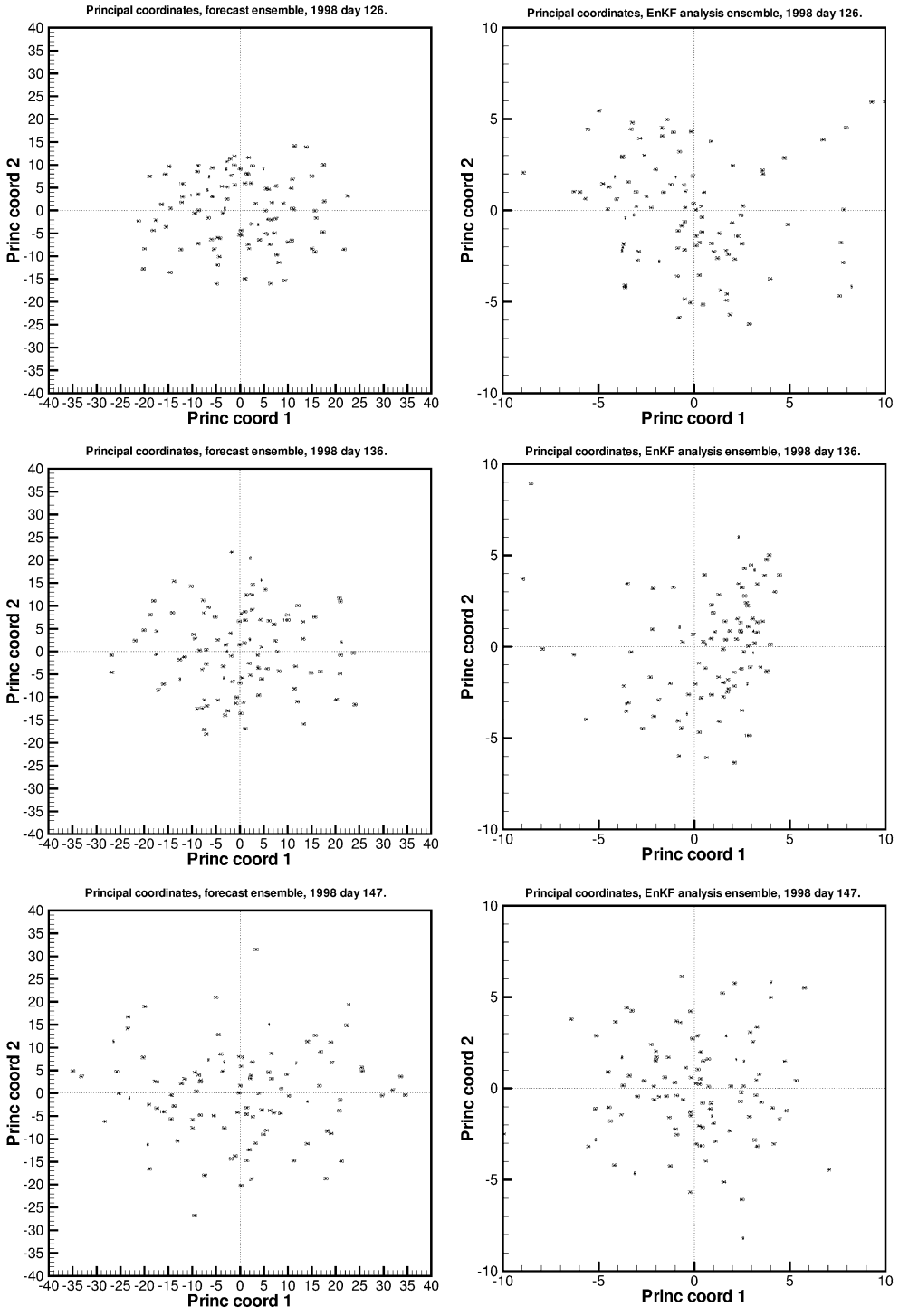

Fig. 5. Principal coordinates for the forecast ensemble (left) and for the analysis ensemble (right) at day 126 (top), day 136 (middle) and day 147 (bottom). Note the different scales on the left and right plots.

These measures are consistent with standard measures of univariate skewness and kurtosis, and they are invariant under linear transformations. Statistical moments up to order three are included in b_{skew}, while moments up to order four, excluding the order three moments, are used in the estimate b_{kurt} (Mardia et al., 1979).

Mardia (1970) showed that for a multinormal distribution (i.e. assuming infinitely many ensemble

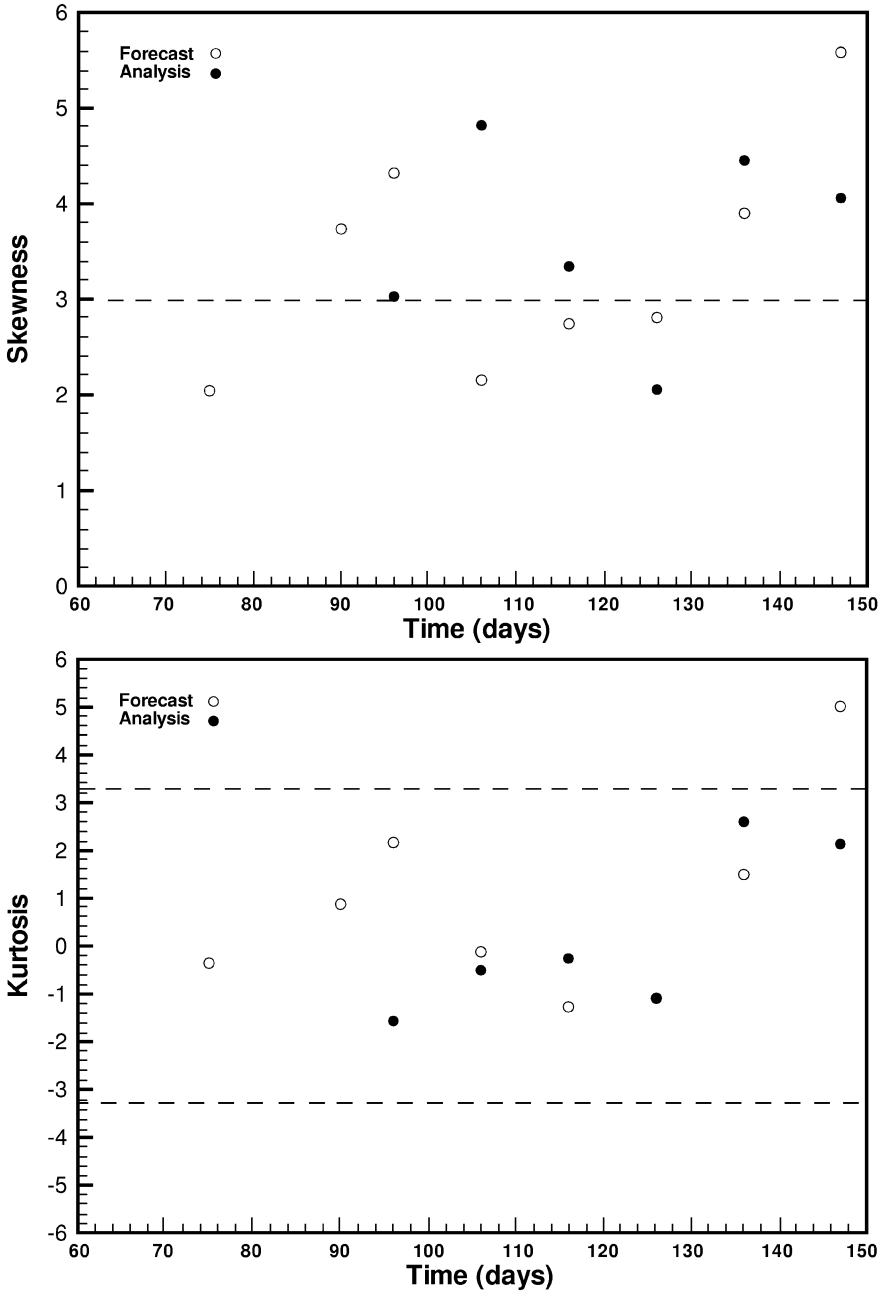

Fig. 6. Measures of multivariate skewness $b_{\mathrm{skew}}(q)$ and kurtosis $b_{\mathrm{kurt}}(q)-q(q+2)$, including the five largest eigenvalues of the spectrum ($q=5$). Asymptotic 95% confidence limits are shown as dashed lines.

members), population skewness and kurtosis can be written as

$$\beta_{\mathrm{skew},p} = 0, \tag{15}$$

$$\beta_{\mathrm{kurt},p} = p(p+1), \tag{16}$$

where p is the number of control variables. Further, assuming that n_{ens} ensemble members are extracted from a multinormal distribution, Mardia (1970) found that as $n_{\mathrm{ens}} \rightarrow \infty$,

$$b_{\mathrm{skew}} \sim \frac{6}{n_{\mathrm{ens}}}\chi_f^2, \tag{17}$$

$$b_{\mathrm{kurt}} - p(p+2) \sim \sqrt{\frac{8p(p+2)}{n_{\mathrm{ens}}}}N(0,1), \tag{18}$$

where χ_f^2 is the chi-squared distribution with $f = p(p+1)(p+2)/6$ degrees of freedom and $N(0,1)$ is a standard normal distribution with mean 0 and variance equal to 1.

Stephenson and Doblas-Reyes (2000) estimated $b_{\mathrm{skew}}(q)$ and $b_{\mathrm{kurt}}(q)$ for the leading q principal components by introducing the rank q pseudo-inverse $\mathbf{C}^{-q}$ of the ensemble covariance $\mathbf{C}$, i.e.

$$g_{kl} \approx \mathbf{y}_k^{\mathrm{T}}\mathbf{C}^{-q}\mathbf{y}_l = n_{\mathrm{ens}}\sum_{i=1}^{q} U_{i,k}U_{i,l}. \tag{19}$$

To conclude, large deviations of the sample measures $b_{\mathrm{skew}}(q)$ and $b_{\mathrm{kurt}}(q) - q(q+2)$ from zero can be used as an indication of nonnormality, and confidence limits for the null hypothesis of normality can be found from Eqs. (17) and (18) with $p = q$.

Fig. 6 shows the evolution of $b_{\mathrm{skew}}(5)$ and b_{kurt} (5)–35. Note that for $q = 5$, we have $f = 35$, and the 95% confidence limit of χ_{35}^2 is 49.802. Using Eq. (17), this gives an asymptotic confidence limit of 2.988 for $b_{\mathrm{skew}}(5)$. Similarly, the standard normal distribution $N(0,1)$ has a 95% confidence limit of 1.960, resulting in an asymptotic confidence limit of 3.280 for $b_{\mathrm{kurt}}(5)-35$.

The measure of skewness includes variations outside the asymptotic 95% confidence limit expected for normally distributed data (only 5% should be out of bound for a normal distribution). However, the deviations are moderate, indicating a weakly skewed ensemble. The skewness may be a result of nonlinear evolution, or it may simply be due to sampling. Also, we only allow for positive biochemical concentrations, e.g. negative values are set to zero after the analysis, and the lower end of the distribution is therefore truncated. Note that the duration of our experiment is relatively short, and it would be very interesting to observe possible systematic trends in a long-term experiment.

The estimate of kurtosis is consistent with a normal distribution at 95% confidence for most of the time; only for the forecast ensemble at day 147, $b_{\mathrm{kurt}}(q) - q(q+2)$ is significant, indicating a weakly flat ensemble. Again, it would be interesting to investigate the evolution of the kurtosis in a long-term experiment.

From Fig. 2, it is seen that the first five eigenvalues, which are included in the calculations for the moment measures, account for about 26–40% of the total variance for the forecast spectra, and about 14–23% for the analysis spectra, respectively. To investigate the sensitivity with respect to the cutoff of the eigenvalue spectrum, we also estimated the skewness and kurtosis for $q = 10$. The results were very similar to those in Fig. 6.

4. Summary

In the Part 1 paper, we investigated an advanced data assimilation method, the ensemble Kalman filter, with a three-dimensional biochemical model of the North Atlantic, utilizing chlorophyll data from the SeaWiFS ocean colour sensor. A simple experiment for April and May 1998 with data assimilation every 10th day was presented, with some additional sensitivity tests at the first assimilation time at day 96. It was shown that the ensemble Kalman filter was able to provide estimates of phytoplankton consistent with the data, and that the multivariate analysis also affected the other model variables. Further, it was seen that the variance of each variable decreased during the assimilation.

The extension of the traditional Kalman filter for linear dynamics to methods appropriate for nonlinear systems is nontrivial. While an initially Gaussian field remains Gaussian during linear evolution, higher-order moments may develop for nonlinear processes.

An important property of ensemble-based methods is that one can monitor various statistical properties. Any statistical information can be extracted from the ensemble, which means that one can test the validity of any hypothesis about the error statistics (see below).

One simple approach to monitor the evolution of the ensemble, is to use some kind of state vector average. For example, we calculated grid point averages for each variable in the surface layer, and plots of phytoplankton and nitrate were shown. As expected, the ensemble of averages spreads out during model integration, while it converges during an ensemble Kalman filter analysis (see Fig. 1). Further, the phytoplankton ensemble experiences a higher degree of convergence than the nitrate ensemble, which is expected since phytoplankton is the measured variable of the multivariate system. To conclude, some kind of state variable average may be a useful first approach for monitoring an ensemble. However, important information is probably lost by taking the mean, and more optimal methods should be used for a thorough analysis.

Multidimensional scaling (MDS) is a powerful method to interpret ensemble member distances in some reduced space. For Euclidean distances, the classical solution, which is given by plotting the q leading principal coordinates of the data matrix $\mathbf{X}$, is optimal. The quality of the projection will be dependent on the dominance of the first q eigenvalues. In our two-dimensional projections, the two first eigenvalues accounted for about 13–26% of the total variance for the forecast spectrums and about 6–14% for the analysis spectrums, which means that the projected distances are quite approximative. (However, one could increase the precicion by considering three-dimensional projections, or by studying pairs or triples of principal coordinates.) As for the state variable averages, the principal coordinates experienced a divergence during model integration and a convergence during each ensemble Kalman filter analysis. In fact, the spread of the forecast ensemble increased throughout the experiment, which could indicate an unstable model. However, the experiment only included the early part of the spring bloom, which is the most unstable period of the year for the biochemical system. It should also be mentioned that by marking each member by a unique number, one can follow it throughout the simulation. Also, outliers can be identified and studied in detail.

Ensemble-based methods can provide information about any statistical moment. For example, multivariate estimates of skewness and kurtosis (flatness) can be found, and simple criteria exist to determine the relevance of these. Thus, one can evaluate the hypothesis of normally distributed data. In our experiment, the evolution of the skewness included variations outside the 95% confidence limit expected for a Gaussian ensemble. However, the deviations were not large, indicating weakly skewed data. The kurtosis was consistent with a Gaussian distribution at 95% confidence for most of the time, only the last forecast ensemble seemed to be weakly flat. To conclude, measures of ensemble skewness and kurtosis provide information about the third- and fourth-order moments of the ensemble, indicating the degree of nonnormality. Minimum variance estimators (or estimators assuming normally distributed errors) are commonly used to assimilate data into nonlinear ocean models, and monitoring the long-term evolution of skewness and kurtosis should be done routinely in order to verify that the hypothesis of Gaussian error fields is valid at all analysis times. For our biochemical model, where the moment measures indicated a weakly skew ensemble with flatness consistent with a normal distribution (but weakly flat at the final time), it should be appropriate to use the ensemble Kalman filter for the assimilation.

Acknowledgements

This work has been supported by the EC-MAST-III DIADEM project (MAS3-CT98-0167), by the Research Council of Norway through project 125793/410 and through a grant of computing time from the Norwegian Supercomputing Committee (TRU). The SeaWiFS data have been processed at the Joint Research Centre in Ispra, Italy.

References

Anton, H., 1991. Elementary Linear Algebra, 6th ed. Wiley, New York.

Bennett, A.F., 1992. Inverse Methods in Physical Oceanography. Cambridge Univ. Press, Cambridge. ISBN 0-521-38568-7.

Bleck, R., Rooth, C., Hu, D., Smith, L.T., 1992. Salinity-driven thermocline transients in a wind- and thermohaline-forced isopycnic coordinate model of the North Atlantic. J. Phys. Oceanogr. 22, 1486–1505.

Drange, H., 1994. An isopycnic coordinate carbon cycle model for the North Atlantic: and the possibility of disposing of fossil fuel CO_2 in the ocean, PhD thesis, Department of mathematics, University of Bergen/Nansen Environmental and Remote Sensing Center, Edv. Griegsvei 3A, N-5037 Solheimsviken, Norway.

Drange, H., 1996. A 3-dimensional isopycnic coordinated model of the seasonal cycling of carbon and nitrogen in the Atlantic Ocean. Phys. Chem. Earth 21 (5–6), 503–509.

Jazwinski, A.H., 1970. Stochastic Processes and Filtering Theory. Academic Press, New York.

Kincaid, D., Cheney, W., 1991. Numerical Analysis, Brooks/Cole, ISBN 0-534-13014-3.

Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika 57 (3), 519–530.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, London.

Natvik, L.-J., Evensen, G., 2003. Assimilation of ocean colour data into a biochemical model of the North Atlantic: Part 1. Data assimilation experiments. J. Mar. Syst. (this issue).

Stephenson, D.B., Doblas-Reyes, F.J., 2000. Statistical methods for interpreting Monte Carlo ensemble forecasts. Tellus 52A, 300–322.