Available online at www.sciencedirect.com
ScienceDirect
Journal of Marine Systems 68 (2007) 327–348
JOURNAL OF MARINE SYSTEMS
www.elsevier.com/locate/jmarsys

Application of an Ensemble Kalman filter to a 1-D coupled hydrodynamic-ecosystem model of the Ligurian Sea

F. Lenartz [a,*], C. Raick [a], K. Soetaert [b], M. Grégoire [a]

[a] University of Liège, Department Oceanology, Sart-Tilman B6c, B-4000 Liège, Belgium
[b] Netherlands Institute of Ecology, Centre for Estuarine and Marine Ecology, P.O. Box 140, 4400 AC-Yerseke, The Netherlands

Received 15 March 2006; received in revised form 29 November 2006; accepted 4 December 2006
Available online 22 December 2006

Abstract

The Ensemble Kalman filter (EnKF) has been applied to a 1-D complex ecosystem model coupled with a hydrodynamic model of the Ligurian Sea. In order to improve the performance of the EnKF, an ensemble subsampling strategy has been used to better represent the covariance matrices and a pre-analysis step for correcting the non-normality of the members distribution has been implemented. Twin experiments have been realized to assess the performance of the developed tool and a real data assimilation experiment has been conducted to hindcast the ecosystem at the Dyfamed site during the year 2000. Finally the performance of the EnKF has been compared with a Singular Evolutive Extended Kalman (SEEK) filter with a fixed basis. We conclude that, on one hand, there is a benefit in using the subsampling strategy and the lognormal transformation with the EnKF, and on the other hand, this filter presents better performance than the fixed basis version of the SEEK filter. However, it also incurs a large computational cost.

Keywords: Ecosystems; Hydrodynamics; Kalman filters; Ligurian Sea

1. Introduction

Models are ideal tools either to clarify the mechanisms that control concentrations of tracers or the functioning of food webs, or to make predictions about how ecosystems will react to changing environmental conditions. In contrast to well-established hydrodynamic models, ecosystem model structure and parameterization are not universal and can largely differ depending on their purpose. Though biogeochemical models are highly idealized, they are generally strongly non-linear, very sensitive to initial conditions, include many unknown processes and are burdened with a surfeit of parameters, so that the predictability of pelagic systems is limited. In addition, it is impossible to obtain exhaustive knowledge of marine systems through observations, because such surveys would be far too expensive. Consequently, an interesting solution consists merging information from both the model and the biogeochemical measurements; this is called data assimilation (DA).

Data assimilation techniques can be used to improve model performance either by optimizing a reduced set of model parameters or by correcting the state produced by the model, both through a variational or a sequential approach. For operational purposes the combination of parameters optimization and state estimation is most

* Corresponding author.
E-mail address: F.Lenartz@ulg.ac.be (F. Lenartz).

0924-7963/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jmarsys.2006.12.001

promising, nevertheless in this paper we restrain ourselves to state estimation *via* a sequential approach. Among sequential data assimilation techniques, the Kalman filter (KF) introduced by Kalman (1960) and originally devised for linear models, is the most widely used method. This method has the attractive property that it not only propagates the state, but also the model uncertainty. In order to perform DA on non-linear models, several different implementations of the KF have been devised: the Extended version (EKF) proposed by Jaswinski (1970), the Ensemble version (EnKF) proposed by Evensen (1994) and the Singular Evolutive Extended version (SEEK) proposed by Pham et al. (1998). These versions differ notably in the propagation of the model uncertainty and this has important bearing on how far the model can deal with non-linearities and on the computational cost of these methods. We chose to use the EnKF, because the ensemble representation of the probability distribution of the state has the advantage of circumventing the time-consuming propagation of the error covariance matrix, and for highly non-linear models, of not relying on the linearization of the model dynamics. In contrast, the EnKF has the disadvantage of requiring a great number of ensemble members to represent correctly the probability distribution of the state, so that the forecast step is very time-expensive.

In this paper, we test the performance of the EnKF on the 1-D coupled hydrodynamic-ecosystem model proposed by Raick et al. (2005) in order to study the seasonal cycle of the biogeochemical processes in the Ligurian Sea.

The paper is organized as follows: the section Materials and methods includes a brief description of the 1-D coupled hydrodynamic-ecosystem model of the Ligurian Sea in Section 2.1 and a summarized theory of the EnKF in Section 2.2. The ensemble subsampling strategy and the Gaussian anamorphosis are respectively described in Sections 2.2.1 and 2.2.2. Dyfamed data are presented in Section 2.3 and the error measurement tools used in this paper are listed in Section 2.4. Our data assimilation experiments are then presented and analyzed: the twin experiments in Section 3, the real data assimilation experiment in Section 4 and the comparison between the EnKF performance and the SEEK filter with a fixed basis performance in Section 5. Finally, we present the conclusion of this work in Section 6.

2. Materials and methods

2.1. The coupled hydrodynamic-ecosystem model

The model used in this paper is the 1-D coupled hydrodynamic-ecosystem model of the Ligurian Sea developed by Raick et al. (2005) describing the pelagic food web of the Ligurian Sea.

The hydrodynamic model is the primitive equations model, in its 1-D version, developed at the Geo Hydrodynamics and Environmental Laboratory (Gher) of the University of Liège. It is a non-linear, baroclinic

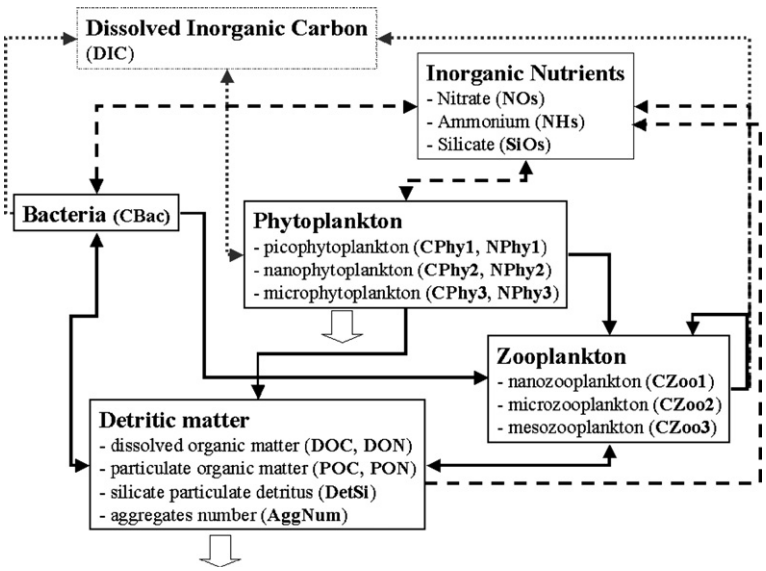

Fig. 1. Representation of the ecosystem model (reprinted from Raick et al., 2005). Each line type represents a kind of flux of matter: solid arrows for organic flows, dashed arrows for inorganic matter flows and dotted arrows for gas flows. Double arrows represent sinking. Dissolved Inorganic Carbon (DIC) is considered as a pool, it is not a state variable of the system.

model which uses a turbulent closure scheme based on the turbulent kinetic energy and on an algebraic mixing length, taking the intensity of both stratification and surface mixing into account (*e.g.* Nihoul and Djenidi, 1987; Delhez et al., 1999). Reduced to its vertical dimension, it contains five state variables: two components of horizontal velocity, the temperature, the salinity and the turbulent kinetic energy. The Gher 1-D hydrodynamic model is described in Lacroix and Grégoire (2002), to simulate the Frontal experiments conducted in the Ligurian Sea from 1984 to 1988. For the Dyfamed experiments of year 2000 (see Raick et al., 2005), the model is forced at the air–sea interface by meteorological data coming from the Côte d'Azur meteorological buoy.

The ecosystem model contains nineteen state variables describing the carbon and nitrogen cycles of the pelagic food web. Phytoplankton and zooplankton are both divided in three size-based compartments and the model includes an explicit representation of the microbial loop including bacteria, dissolved organic matter, nano- and microzooplankton. The internal C:N ratio is assumed variable for phytoplankton and detritus, and constant for zooplankton and bacteria. Silicate is considered as a potential limiting nutrient of phytoplankton's growth. A schematic representation of the ecosystem model is shown in Fig. 1.

The physical and biological models are coupled off-line. Simulations with the hydrodynamic model are performed and then temperature and turbulent diffusion coefficient profiles are stored. After that, the biological model is integrated by the subroutines library Femme, a Flexible Environment for Mathematically Modelling the Environment developed by Soetaert et al. (2002) and designed for implementing, solving and analyzing mathematical models in ecology.

The depth of the vertical domain is 400 m with a zero-flux lower boundary, such that all the organic matter produced in the euphotic layer by primary production is remineralized in the modeled domain. In this way the model is fully conservative. The vertical mesh has an exponential scale in order to take into account the higher variability of the ecosystem in the upper layers. Integration is done using the Euler explicit method with a constant time step of 45 min, except for turbulent mixing which is solved with an implicit

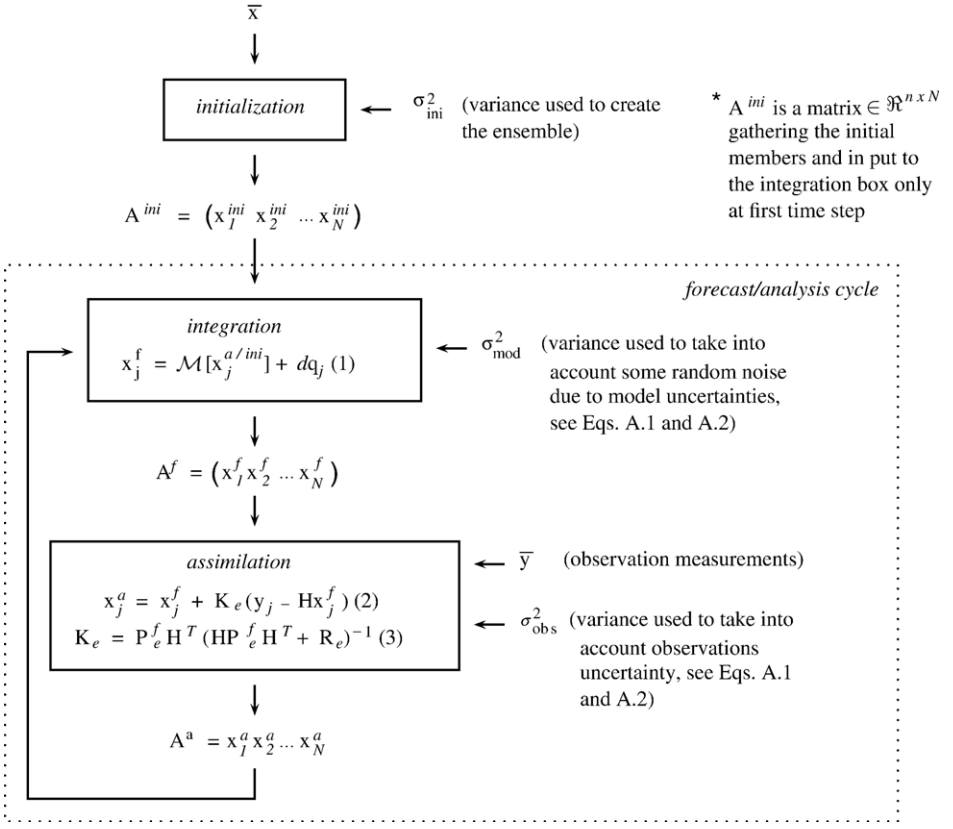

Fig. 2. The EnKF methodology (redrawn from Eknes and Evensen, 2002).

method. The ecosystem is presented in detail in Raick et al. (2005) and validated against the Dyfamed data of year 2000, presented in Section 2.3.

2.2. Ensemble Kalman filter

The EnKF method was formulated by Evensen (1994) in order to resolve some problems accounted with non-linear operators. Instead of calculating the Jacobian of the model operator to propagate the model uncertainty as in the EKF, here the error covariance matrix is derived from an ensemble representation of the state probability distribution. According to us, the two main reasons for using ensemble methods are that (1) they circumvent the time-consuming propagation of the error covariance matrix and the calculation of the model operator and (2) for highly non-linear models, they do not rely on the linearization of the model dynamics and may therefore represent better the model statistics. However, as shown by van Leeuwen and Evensen (1998), the analysis scheme of all Kalman filters relies on the assumed normality of the probability density function used to derive the optimum state, by minimization of a penalty function. For non-linear systems this assumption is of course invalid and we propose a remedy to this problem in Section 2.2.2. Besides, there is another problem inherent to all filters: the non-linearity of the observation operator. Evensen (2003) proposed a solution that consists in augmenting the model state vector with a diagnostic variable which is the model prediction of the measurement. Under these assumptions, it is the statistical noise that dominates the errors in the EnKF, and there are no closure problems or unbounded error variance growth, as have been found with assimilation methods relying on the use of a tangent linear model.

The EnKF integrates an ensemble of model states forward in time using the model equations. Usually the number of members N is of the order of 10^2, whereas the state vector size n is of the order of 10^5 for a 3-D model; in our 1-D model case $n=1900$. During the forecast step, each individual member is integrated using a stochastic differential equation, *i.e.* forced with a random noise component which represents the stochastic model error (Eq. (1) in Fig. 2). It can be shown that such ensemble integration becomes identical to a Markov Chain Monte Carlo method for solving the Fokker–Planck equation for the evolution in time of the probability density of the model state (Evensen, 1994). Since the non-linear dynamics model is used, during the forecast, the only approximation associated with this approach is that a finite number of members are used in the ensemble. Integration in time is performed until measurements are available. At these time instants, an analysis scheme is used to update or correct the model state in a statistically consistent way, *i.e.* by minimizing the error variance of the analyzed estimate in a least square sense, considering the measurements, the model forecast and their respective error statistics (Eqs. (2) and (3) in Fig. 2). At the analysis steps, another approximation exists: the assumption of a Gaussian distribution of the ensemble states. A schematic illustration of the algorithm is given in Fig. 2 and described in the next paragraph. Further details about the EnKF implementation and its applications can be found in Allen et al. (2002), Bertino et al. (2003), Brasseur (2005), Burgers et al. (1998), Eknes and Evensen (2002) and Evensen (2004).

In the initialization phase, a first guess model state is perturbed to create an ensemble of N initial members x_j^{ini} as explained in Appendix A. Then, these perturbed model states are integrated forward in time with the model dynamics, expressed by means of a non-linear model operator $\mathcal{M}$ and a random noise component dq representing the stochastic model error. When a first observation set becomes available, each of the model forecasts x_j^{f} is corrected. During this analysis step, a new ensemble of analyzed model states x_j^{a} is computed (Eq. (2) in Fig. 2) based, (1) on the prior model state x_j^{f}, (2) on the measurements $\bar{y}$, related to the true model state x^{t} through $\bar{y}=Hx^{\mathrm{t}}+\varepsilon^{\mathrm{o}}$, where H denotes the observation operator (which can possibly be non-linear and then denoted $\mathcal{H}$), ε^{o} the observational error and x^{t} the true state, and (3) on the so-called Kalman gain matrix $\boldsymbol{K}_{\mathrm{e}} \in \Re^{n\times m}$ (Eq. (3) in Fig. 2), denoting the degree of correction of the model state at the analysis. This matrix $\boldsymbol{K}_{\mathrm{e}}$ is estimated based on the error covariance matrix of the model forecast state $\boldsymbol{P}_{\mathrm{e}}^{f} \in \Re^{n\times n}$ (Eq. (4)) and the observational error covariance matrix $\boldsymbol{R}_{\mathrm{e}} = \frac{\boldsymbol{E}\boldsymbol{E}^T}{N-1} \in \Re^{m\times m}$ where $\boldsymbol{E} = (\epsilon_1 \epsilon_2 \ldots \epsilon_N) \in \Re^{m\times N}$ is the ensemble of observational errors ε_j combining the instrumental error $\varepsilon_j^{\mathrm{o}}$ and the noise added to generate an ensemble of observations. The Kalman gain can be interpreted as the ratio between the error variance of the forecast and the total error variance projected in the observation space. In the limit of perfect observations ($\boldsymbol{R}_{\mathrm{e}} \sim 0$) of the whole state, the Kalman gain matrix converges to the inverse of the observation operator and the correction will completely follow the data. In contrast, for an extremely accurate model forecast ($\boldsymbol{P}_{\mathrm{e}}^{\mathrm{f}} \sim 0$) the correction is negligible.

After this step, the ensemble of analyzed states is integrated forward until new observations become available and the process is repeated again. At all time step, the EnKF estimate is defined as the ensemble mean.

Let us now introduce two modifications of the EnKF that we test in this work: the Gaussian anamorphosis and the ensemble subsampling strategy. According to our knowledge, these techniques haven't been implemented simultaneously on a complex ecosystem model yet, so this is what we propose to do.

2.2.1. Ensemble subsampling

As said previously, ensemble is used to represent the probability distribution of the state and all statistics derived from it; especially the mean state and the model variability. In order to improve this representation with a restrained number of members, we adopt the ensemble subsampling strategy proposed by Evensen (2004). In the EnKF scheme, the exact error covariance matrix $\boldsymbol{P}$ at the forecast and analysis steps is approximated by its ensemble representation

$$\boldsymbol{P} \sim \boldsymbol{P}_e = \frac{\boldsymbol{A}'\boldsymbol{A}'^T}{N-1} \in \Re^{n\times n}, \tag{4}$$

where $\boldsymbol{A}'$ is the ensemble of model state perturbations, obtained by subtracting the mean ensemble state $\bar{x}$ from each member x_j of the ensemble $\boldsymbol{A}$. If we perform the singular value decomposition of $\boldsymbol{A}'=\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ and consider the eigenvalue decomposition of $\boldsymbol{P}=\boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^T$, then

$$\boldsymbol{P} = \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^T \tag{5}$$

$$\boldsymbol{P}_e = \frac{\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T}{N-1} = \frac{\mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^T}{N-1}, \tag{6}$$

where $\mathbf{U} \in \Re^{n\times n}$ and $\mathbf{V} \in \Re^{N\times N}$ contain respectively the left and the right singular vectors of $\boldsymbol{A}'$, $\boldsymbol{\Sigma} \in \Re^{n\times N}$ its singular values, $\mathbf{Z} \in \Re^{n\times n}$ the eigenvectors of $\boldsymbol{P}$ and $\boldsymbol{\Lambda} \in \Re^{n\times n}$ its eigenvalues. When the number of members N tends toward infinity, the n singular vectors in $\mathbf{U}$ will converge toward the first n eigenvectors in $\mathbf{Z}$, and the square of the singular values $\boldsymbol{\Sigma}^2$ divided by $N-1$ will converge toward the eigenvalues $\boldsymbol{\Lambda}$.

This shows that a good approximation $\boldsymbol{P}_e$ of $\boldsymbol{P}$ is obtained either by increasing the ensemble size N, thus increasing the size of the hyperspace spanned by the members, or by thoroughly choosing the members of the ensemble as the N dominant modes. These N modes are obtained by selecting among βN members, β being an integer larger than 1, those corresponding to the N largest singular values. The latter technique is the ensemble subsampling strategy described in details in Evensen (2004). Hence, given an initial estimate of $\boldsymbol{P}$, *e.g.* from a previous run, we can drastically reduce the size of an ensemble, that provides a good representation $\boldsymbol{P}_e$ of $\boldsymbol{P}$, thanks to this strategy.

In our experiments, the ensemble subsampling has been used to select the N initial members $\boldsymbol{A}^{ini}=(x_1^{ini}x_2^{ini}\ldots x_N^{ini})$. Besides, Burgers et al. (1998) suggested also to create an ensemble of measurements $\boldsymbol{D} = (y_1y_2\ldots y_N) \in \Re^{m\times N}$ in order to avoid underestimating the ensemble variance at the analysis step. Thus, we have subsampled the ensemble of observations $\boldsymbol{D}$ generated according to Eq. (A.2), to get the best possible ensemble representation $\boldsymbol{R}_e$ of the observational error covariance matrix $\boldsymbol{R}$. This improved sampling strategy might be more important for a model with a larger state space; in typical 3-D data assimilation experiments the state vector size is $\mathcal{O}(10^5)$, as for example in Natvik and Evensen (2003), in our 1-D experiments the state vector size is only 1900.

We present in Fig. 3 the normalized singular values of the initial ensemble of normalized perturbations $\hat{A}'^{ini}$ ($N=100$) obtained without subsampling and by using subsampling from starting ensembles of different sizes ($\beta N=300$, 500 and 1000). We clearly see that the conditioning of $\hat{\boldsymbol{A}}'^{ini}$ is improved when a larger starting ensemble is used: the inverse of the condition number (*i.e.* the ratio of the smallest singular value to the largest one) is proportional to β. The presence of a tiny 100th singular value for the curve without resampling comes from the removal of the ensemble mean during the post-treatment, making the matrix columns linearly dependent. Because we have only plotted the first 100 singular values of the spectrum for the larger ensembles, the very small last singular value of each of them does not appear in Fig. 3.

2.2.2. Gaussian anamorphosis

Applying Kalman filters to biogeochemical models must be done carefully: on one hand, the state vector has to be physically consistent to be further propagated by

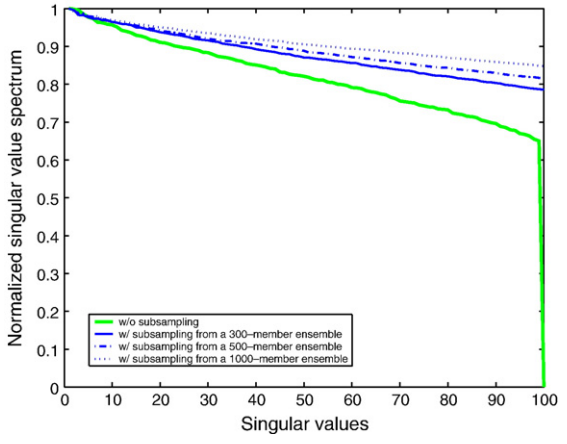

Fig. 3. Spectrum of the normalized singular values of the initial ensemble.

the model (*e.g.* negative concentrations are totally unrealistic) and on the other hand, state and observation vectors have to be multivariate normally distributed variables for optimal use of the linear statistical analysis. The latter requirement is rarely encountered when the model used is non-linear. One solution is the Gaussian anamorphosis.

Let us consider two distinct but related variables, the original ones x being physically consistent and the transformed ones $\tilde{\mathbf{x}}$ suitable for linear estimation, linked to each other through $\tilde{\mathbf{x}}=\psi(x)$, where ψ is the anamorphosis function, which converts the original variables x into normally distributed ones $\tilde{\mathbf{x}}$. This is obviously less constraining than expecting the same variable to fulfill both physical and statistical requirements. Similarly, we also introduce an anamorphosis function χ for the observations $\tilde{\mathbf{y}}=\chi(y)$. In this way, the observation operator H becomes

$$\tilde{H} = \chi \circ H \circ \psi^{-1}, \tag{7}$$

where $\circ$ denotes a Hadamard product (component by component product). If we group the transformed state vectors $\tilde{\mathbf{x}}$ and the transformed measurement vectors $\tilde{\mathbf{y}}$ in matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{D}}$ respectively, we can write $\tilde{\mathbf{A}}^{a}=\tilde{\mathbf{A}}^{f}+\tilde{\mathbf{K}}_{e}(\tilde{\mathbf{D}}-\tilde{\mathbf{H}}\tilde{\mathbf{A}}^{f})$, with $\tilde{\mathbf{K}}_{e}=\tilde{\mathbf{P}}_{e}\tilde{\mathbf{H}}^{T}(\tilde{\mathbf{H}}\tilde{\mathbf{P}}_{e}\tilde{\mathbf{H}}^{T}+\tilde{\mathbf{R}}_{e})^{-1}$. This is the exact copy of the EnKF analysis step, except that all matrices are tilded, meaning that they have been computed with the multivariate normally distributed variables (as opposed to the physical ones). Once the correction is achieved, we transform back the variables according to $x=\psi^{-1}(\tilde{\mathbf{x}})$. The physically consistent variables obtained by this process are the samples of a probability density function from which we can directly draw unbiased statistics, so that no extra computation is required to take the mean, the standard deviation or other statistics of the ensemble. In order to simplify the procedure, we make the assumption that the variables at different locations are identically distributed, so that the anamorphosis function used is the same all over the spatial domain.

2.2.2.1. Distribution of the forecast state variables. The distribution of the forecast state variables is exemplified in Fig. 4 at different times (*i.e.* at days 100, 150 and 200) and depths (*i.e.* at 0.5 m, 10 m and 95 m) through the histograms of the forecast members (*i.e.* CPhy1, CPhy3, NPhy3, Czoo3 and SiOs) during an assimilation run. The number of ensemble members for this run is 150.

We can observe that CPhy1 is approximately normally distributed in the whole water column at day 100, and that the contents in carbon and nitrogen of the third group of phytoplankton at day 150 have very similar distributions, more lognormal than normal near the surface and at 95 m and left-tailed at 10 m depth. We see an analogous tendency for CZoo3, whereas the SiOs is normally distributed in the three chosen layers. These

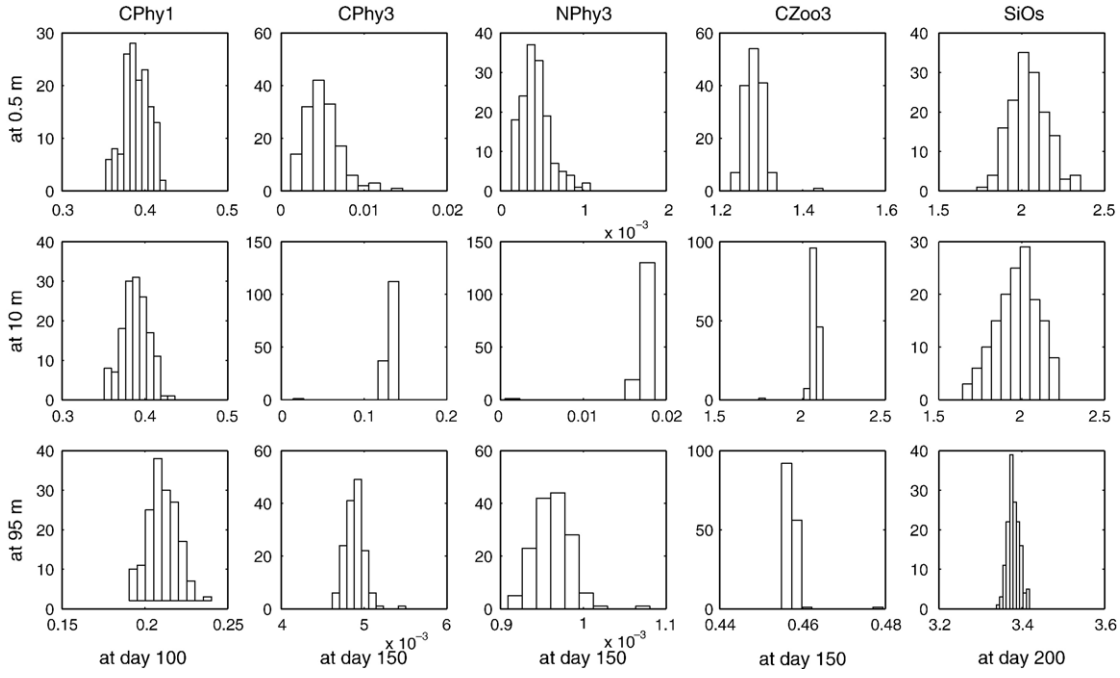

Fig. 4. Distribution of the state variables obtained by an ensemble assimilation run (see Fig. 1 for the definition of the state variables).

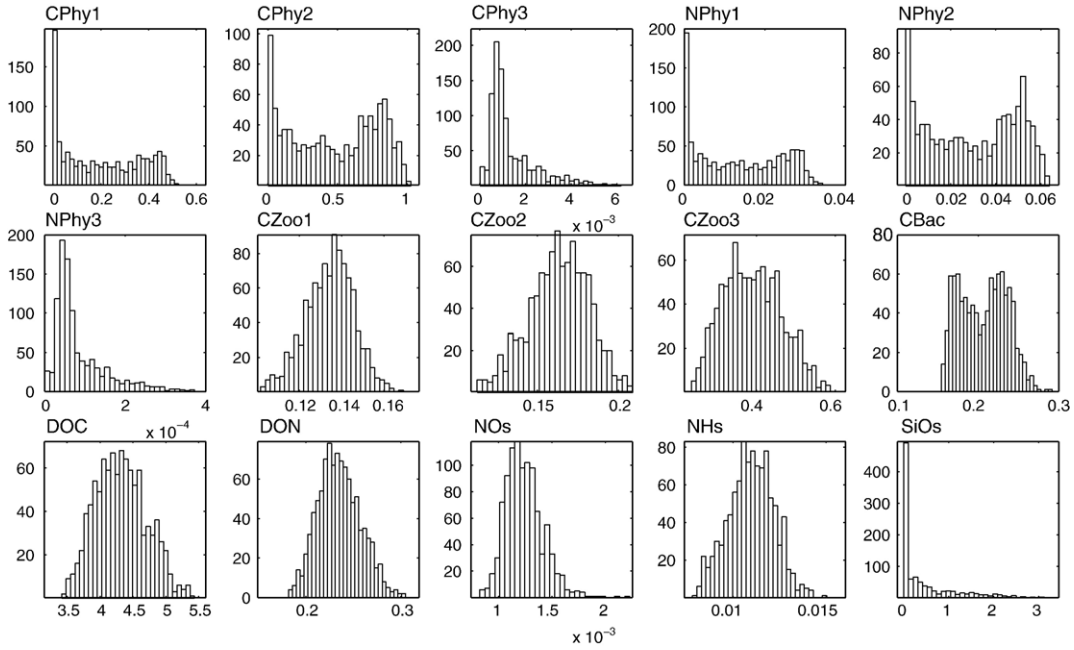

Fig. 5. Assumed distribution of some variables measurements obtained by a Monte Carlo simulation.

few examples illustrate that, in practice, the spatial homogeneity hypothesis is rarely respected.

2.2.2.2. Distribution of the observations. As we cannot use multiple measurements for estimating the probability density function of the measurements, we can use a Monte Carlo simulation. Though this method is not perfect, it can be useful when one has no idea at all about the measurements distribution. We propose in Fig. 5 the histograms of each variable, taken at observation points,

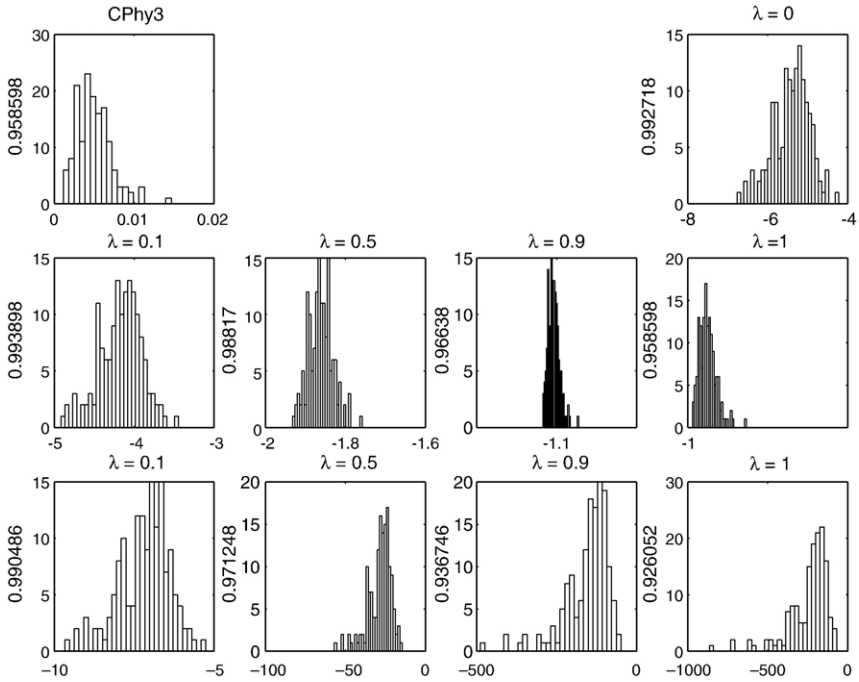

Fig. 6. Distributions of the Box–Cox transformed CPhy3 at 0.5 m depth variable.

resulting from a total of 1000 runs realized by perturbing the most sensitive parameters identified in Raick et al. (2005) (±10% around the calibrated value). We see that phytoplankton observations (especially the third group, which represents the diatoms) as well as nitrate and silicate observations are more lognormally than normally distributed, whereas totally different distributions characterize the other measurements.

2.2.2.3. Box–Cox transformations. We now propose an automatic and objective procedure to check whether or not one should proceed to a transformation of the forecast variables or of the observations, before the analysis step. First, we compute, for each layer, the correlation coefficient $R_{x \text{ or } y}$ between the original variables x or observations y and a normal distribution $\mathcal{N}_{(0,1)}$, all sorted in ascending order (cumulative density function); secondly we estimate the correlation coefficient $R_{\tilde{x}}$ or $\tilde{y}$ between the transformed variables $\tilde{\mathbf{x}}$ or observations $\tilde{\mathbf{y}}$ and the same normal distribution. Finally, we compare space-averaged values and adopt the transformation giving the highest correlation. Interesting trial functions are the Box–Cox transformations

$$\tilde{\mathbf{x}}_\lambda = \frac{x^\lambda - 1}{\lambda} \text{ for } \lambda \neq 0 \tag{8}$$

$$= \ln x \text{ for } \lambda = 0. \tag{9}$$

These functions are conservative (no loss or gain of matter when variables are back-transformed) but restricted to only positive values. We present in Fig. 6 the distribution of the ensemble forecast CPhy3 variable and the histograms corresponding to its transformations according to various values of parameter λ. The y-axis label is the correlation coefficient between the considered original or transformed variable and a normal distribution. We see that, at this depth, a lognormal transformation already improves the normality of the distribution, although the difference between the correlation coefficient values is not large (about 2%). Note that for the assimilation runs realized for this work, we restrain ourselves to the case $\lambda = 0$, *i.e.* the lognormal transformation. However, by examining some forecast ensembles (see Fig. 6), we see that the extension of the algorithm to other cases ($\lambda = s \in \Re$) could be beneficial.

2.2.2.4. Lognormal distribution. We propose to generate the initial members x_j^{ini} and the perturbed measurements y_j of the lognormal-labeled variables and observations in a way that differs from the usual addition of a pseudo random normal field to the best guess estimate, as presented in Appendix A. Hence, the jth member of the sampled ensemble is given by

$$v(z)_{\log,j} = \exp(\mu_{\log}(z) + \sigma_{\log}(z)\mathcal{N}_{(0,1)j}). \tag{10}$$

Consider you know the best guess estimate of the distribution (*i.e.* a prescribed mean μ) and the uncertainty on this best guess estimate (*i.e.* a prescribed standard deviation σ). In order to build a lognormal distribution, from which we can derive the statistics parameters μ and σ, by computing respectively its mean and its standard deviation, we have to define $\mu_{\log}(z)$ and $\sigma_{\log}(z)$ of Eq. (10) as follows: $\mu_{\log}(z) = \ln\mu(z) - \frac{1}{2}\ln\left(1 + \frac{\sigma^2}{\mu^2}\right)$ and $\sigma_{\log}(z) = \sqrt{\ln\left(1 + \frac{\sigma^2}{\mu^2}\right)}$. Under the assumption of a lognormal distribution of both state variables and observations, the transformed observation operator is given by $\tilde{\mathbf{H}} = \ln(\exp(\mathbf{H})) = \mathbf{H}$.

2.3. Data

A large data base including biological, physical, chemical and meteorological data is available for the Ligurian Sea. Since 1991, the time-series program Dyfamed (DYnamics of atmospheric Fluxes in the MEDiterranean sea) records measurements in a selected site in the central part of the Ligurian Sea, shown in Fig. 7, in order to study the response of the ecosystem to climate variability and anthropogenic inputs. The Dyfamed program has been organized in the scope of the French-JGOFS (Joint Global Ocean Flux Studies) program (Marty, 2002). The existence of this large data base and the particular hydro-dynamic conditions with moderate horizontal advection make the Dyfamed site an ideal test area for performing 1-D modeling studies.

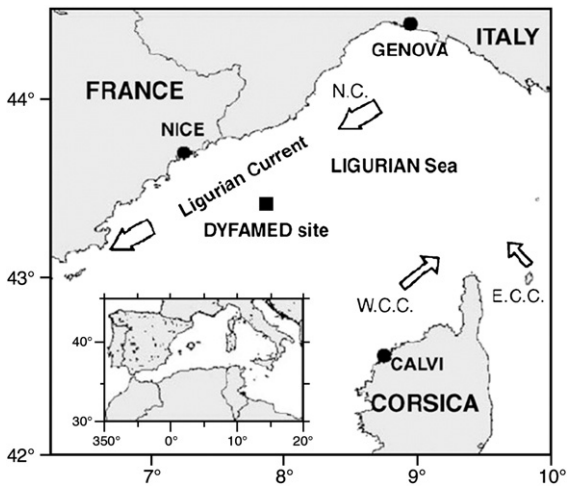

Fig. 7. Location of the Dyfamed site (43°25′N, 7°52′E), reprinted from Marty and Chiaverini (2002).

Nutrients (nitrite, nitrate, silicate and phosphate) profiles are described in detail in Bethoux et al. (1998, 2002). Temperature and salinity data are presented in Marty et al. (2002). Abundance and biomass of free-living bacteria, heterotrophic nanoflagellates and ciliates are described in Tanaka and Rassoulzadegan (2002) and Tamburini et al. (2002). A range of pigments have been detected, in order to characterize different phytoplankton groups (*e.g.* Vidussi et al., 2000, 2001; Marty et al., 2002; Marty and Chiaverini, 2002). Particulate organic matter, in carbon and nitrogen have also been measured. We have chosen particularly year 2002 for the calibration and the validation of the coupled model because of the data availability during this year. Data have been monthly recorded (11 dates) at 12 water depths between 0 and 200 m depth (5 m, 10 m, 20 m, 30 m, 40 m, 50 m, 70 m, 90 m, 110 m, 130 m, 150 m and 200 m): pigments of chlorophyll which provides information about the three phytoplankton groups, nutrients and detritic matter. Nanoflagellates (Zoo1 group), ciliates (Zoo2 group) and bacteria have been measured every month from May 1999 to March 2000: three profiles of these variables are available for the year 2000. No information about the errors on data measurements are available. This is actually the reason why we propose to determine the observations distribution by a Monte Carlo simulation in the previous subsection. All data are available through the Dyfamed Observatory data base : http://www.obs-vlfr.fr/jgofs2/sodyf/home.htm.

2.4. Assessing the model error

When the correspondence between two fields has to be quantified, Taylor (2001) has shown that some different but complementary statistical information (the standard deviation, Root-Mean-Square difference and correlation coefficient) can be considered and summarized in a single plot, the Taylor diagram. We briefly explain this error measurement tool in Appendix B.

In addition we also compute the evolution in time of the space-averaged RMS error and the time-averaged RMS error profile. If we denote by $f_{i,j}$ the forecast and by $r_{i,j}$ the reference fields defined at a given point of the vertical domain (the ith layer) and at a given time (the jth day), the previous quantities are respectively given by

$$\mathrm{RMS}_j^s = \sqrt{\sum_{i=1}^{N_z} \frac{dz_i}{D}\left(f_{i,j}-r_{i,j}\right)^2}, j = 1,2,\ldots,N_t \qquad (11)$$

$$\mathrm{RMS}_i^t = \sqrt{\sum_{j=1}^{N_t} \frac{dt_j}{T}\left(f_{i,j}-r_{i,j}\right)^2}, i = 1,2,\ldots,N_z. \qquad (12)$$

where N_z is the number of discrete points in space, N_t the number of discrete points in time, dz_i the thickness of the ith layer, dt_j the jth time interval, $D=\Sigma_{i=1}^{Nz} dz_i$ the size of the spatial domain and $T=\Sigma_{j=1}^{Nt} dt_i$ the duration of the simulation. We also compute the time evolution of the $\mathrm{RMS}_j^{\mathrm{ens}}$, which provides an estimate of the error that the filter makes on the state at forecast and analysis steps:

$$\mathrm{RMS}_j^{\mathrm{ens}} = \sqrt{\sum_{i=1}^{N_z} \frac{dz_i}{D}\sigma_i^2}, \; j = 1,2,\ldots,N_t, \qquad (13)$$

where the σ_i^2 are the diagonal elements of the ensemble representation of the error covariance matrix. If the analysis is done properly the predicted $\mathrm{RMS}^{\mathrm{ens}}$ and the actual $\mathrm{RMS}^{\mathrm{s}}$ should be of the same order of magnitude.

3. Twin experiments

3.1. Twin experiments strategy

Traditionally one performs twin experiments in order to determine the experimental DA protocol (*e.g.* the assimilation time frequency atf, the ensemble size $\mathrm{ens}_{\mathrm{size}}$, the nature of the observed and corrected variables, respectively $\mathrm{nat}_{\mathrm{obs}}$ and $\mathrm{nat}_{\mathrm{cor}}$) that should be used to assimilate real observations data. For this purpose, we run a reference simulation, considered as the true solution of the system and which is obtained by perturbing the final state of the spinup solution of the model. Perturbation is done according to Eq. (A.2) with a relative standard deviation of 10%. From this run we extract pseudo-observations that are assimilated in an EnKF run. The EnKF is initialized with the annual mean profile of the reference simulation as the best guess estimate used to create the members of the ensemble A^{ini}. This is done in the same way as the initial state of the reference run was created from the spinup final state. We take into account the model uncertainty by perturbing everyday the state variables according to Eq. (A.1), but with a standard deviation corresponding, for each variable and at each depth, to 3% of the annual mean profile of the reference run. In parallel to the EnKF run, we also run an uncorrected simulation, which propagates the same ensemble members without DA. This simulation gives an estimate of the uncorrected state and allows to measure the benefit of data assimilation. As precised previously, in order to avoid underestimating the ensemble variance at the analysis step, we also generate an ensemble of observations D according to Eq. (A.2) and with a relative standard deviation $\sigma_{\mathrm{obs}}=30\%$. Thus, our twin experiments are pretty close to real DA experiments, in which the uncertainty on the measurements is sometimes quite important.

We perform several test cases by changing some of the twin experiment parameters and for each of them we measure the error in the different ways presented in Section 2.4. In Table 1 are the different test cases and their respective parameters values. Because we have a complete reference field (*i.e.* a value at every time step and in every layer), we can use it and the complete fields resulting from the assimilation runs to compute the error statistics; obviously this can not be done in real data assimilation experiments.

3.2. Twin experiments results

3.2.1. Nature of the observed variables impact — T1

The nature of the variables that we should observe to get the best state estimate is a decisive issue. As said in Section 2.2, one of the fundamental hypotheses of the original KF theory is the linearity of the model and observation operators. The EnKF can be applied to non-linear models but does not solve the problem of using a non-linear observation operator $\mathcal{H}$, unless the technique proposed by Evensen (2003) and mentioned in Section 2.2, though not used in this work, is implemented in your filter. As shown in Raick et al. (2005), in this model we don't use fixed Chl:C and Chl:N ratios, so that if we want to assimilate phytoplankton in chlorophyll units such a non-linear operator should be used; therefore we will restrain our observations to measurements that are linearly linked to the state variables, avoiding a source of errors for the assimilation process. In these test cases, we correct nine state variables (the carbon and nitrogen components of the three groups of phytoplankton and the carbon component of the three groups of zooplankton) with a composition of the observation vector that varies (test cases T1A, T1B, T1C, T1D and T1E). However, in real data assimilation experiments, we are limited by the diversity of the variables observed, thus we do not often have the choice of which variables to observe.

Table 1
Twin experiments combination of parameters for each test case

Twin experiments: test cases						
Study type	Test ID	nat_{obs}	nat_{cor}	atf	asf	ens_{size}
nat_{obs}	T1A	P_C	P_C, P_N, Z_C	30	10	100
	T1B	P_N	P_C, P_N, Z_C	30	10	100
	T1C	Z_C	P_C, P_N, Z_C	30	10	100
	T1D	P_C, P_N	P_C, P_N, Z_C	30	10	100
	T1E	P_C, P_N, Z_C	P_C, P_N, Z_C	30	10	100
nat_{cor}	T2A	P_C	P_C	30	10	100
	T2B	P_N	P_N	30	10	100
	T2C	Z_C	Z_C	30	10	100
atf	T3A	P_C, P_N, Z_C	P_C, P_N, Z_C	100	10	100
	T3B	P_C, P_N, Z_C	P_C, P_N, Z_C	50	10	100
	T3C	P_C, P_N, Z_C	P_C, P_N, Z_C	5	10	100
asf	T4A	P_C, P_N, Z_C	P_C, P_N, Z_C	30	100	100
	T4B	P_C, P_N, Z_C	P_C, P_N, Z_C	30	25	100
ens_{size}	T5A	P_C, P_N, Z_C	P_C, P_N, Z_C	30	10	150
	T5B	P_C, P_N, Z_C	P_C, P_N, Z_C	30	10	300

with P_C=CPhy1, CPhy2, CPhy3,
P_N=NPhy1, NPhy2, NPhy3,
Z_C=CZoo1, CZoo2, CZoo3,
nat_{obs} is the nature of the observed variables,
nat_{cor} is the nature of the corrected variables,atfis the assimilation time frequency and is expressed in days,
asf is the assimilation space frequency and is expressed in boxes,
ens_{size} is the number of members constituting the ensemble.

We have plotted in Fig. 8 the time-averaged RMS^t errors profiles (Eq. (12)) of some integrated variables, derived from state variables as follows: $CPhy = \sum_{i=1}^{3} CPhy_i$, $CZoo = \sum_{i=1}^{3} CZoo_i$, $Nut = NO_3 + NH_4 + SiO_2$, $DOM = DOC + DON$ and $POM = POC + PON + SiPOM$ (see Fig. 1). As we can see, the most positive impact of the assimilation is obtained in the test case T1E, where the composition of the observation and estimation vectors are identical. We also see that the zooplankton and the detritic organic matter are not very sensitive neither to the assimilation process, nor to the nature of the observed variables. For phytoplankton, the observation of zooplankton alone (T1C) does not improve significantly the results, whereas the observation of carbon or nitrogen content of phytoplankton (respectively T1A and T1B) reduces the errors very similarly, so that the curves are nearly superimposed. Unfortunately for the real data assimilation experiment, we do not have measurements for zooplankton concentrations all along the year, so that we choose the test case T1D (observation of CPhy1,2,3 and NPhy1,2,3), as the reference twin experiment.

3.2.2. Nature of the corrected variables impact — T2

We now focus on determining which variable, by its correction, has the most useful impact on system dynamics. As opposed to experiments T1, in T2 test cases we only correct the variables that we observe. We present in Fig. 9 the space-averaged RMS^s error (Eq. (11)) for the integrated variables CPhy, CZoo, Nut, DOM and POM of the model in T2A, T2B, T2C and T1D experiments. We note that the variables that mostly benefit of data assimilation are the carbon content of phytoplankton and the nutrient pool. Before we take a look at the T2 curves, it is worth noting that the T1D experiment (nat_{obs}=P_C, and P_N, nat_{cor}=P_C, P_N and Z_C) gives the smallest RMS^s error for all variables represented in Fig. 9. The correction of only P_C (T2A) is the best choice of assimilation protocol among the T2

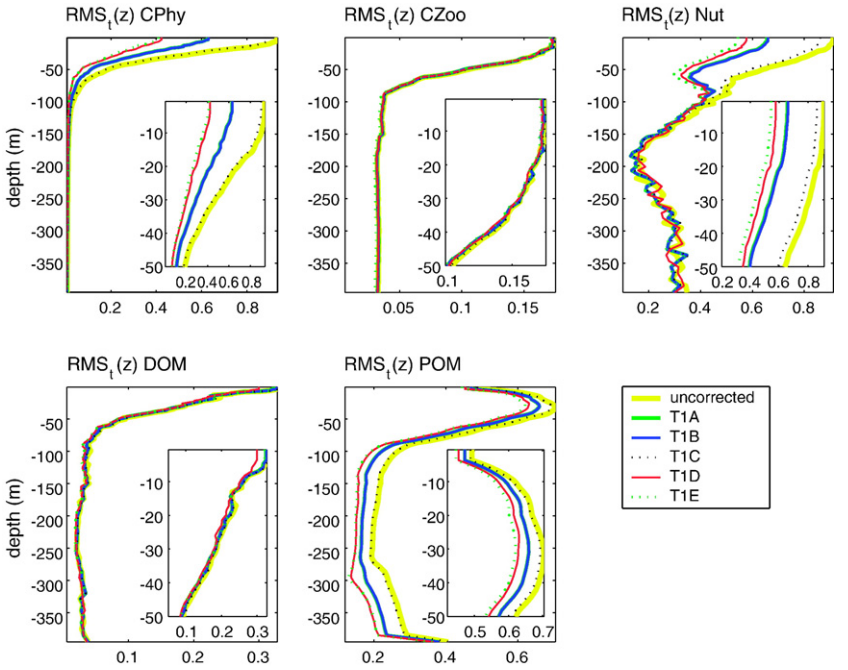

Fig. 8. Twin experiments T1: influence of the composition of the observation vector — Time-averaged RMS^t errors in the whole water column for the variables CPhy, CZoo, Nut, DOM and POM.

experiments. It especially improves the results for the phytoplankton and the nutrients. Observation and correction of P_N or Z_C just slightly improve the results for phytoplankton but actually do not differ that much from the uncorrected run for the other variables. These results exhibit the importance of a correct representation of phytoplankton in carbon in the system; actually carbon is the main element of this ecosystem model. Phytoplankton primary production and nutrient uptake are decoupled and nutrients can be internally stored by phytoplankton. Because of these storages, the nutrient content of the phytoplankton cell does not reflect the instantaneous performance of the cell, in contrast to the internal content in carbon that can not be stored. This is why assimilating only phytoplankton biomass in nitrogen and letting the system dynamics update the other variables does not allow to converge to the reference solution (see Raick et al., 2005).

3.2.3. Assimilation time frequency impact — T3

We clearly see the key role played by the atf in Fig. 10, where we present the RMS^s and RMS^t differences (Eqs. (11), (12)) averaged on the main state variables of the model (CPhy1,2,3, NPhy1,2,3, CZoo1,2,3, CBac, NOs, NHs, DOC, DON, CPOM and NPOM). The more often we assimilate data, the better the results. The assimilation time frequency is obviously limiting in a real biogeochemical data experiment, because biological surveys are generally not automatic, except for surface chlorophyll concentration that can be estimated through the use of satellite data. Nevertheless, we note that assimilating data every 30 days (T1D), as can be done in a real DA experiment, already yields a significant reduction of the RMS errors. The difference with T3C (every 5 days) is hardly visible on the RMS^t plot, confirming that an every 30 days DA protocol is a good choice for operational use.

3.2.4. Assimilation space frequency impact — T4

The importance of the asf can be seen in Fig. 11, where we present the mean RMS^s and RMS^t differences (Eqs. (11), (12)) (see Section 3.2.3), for the test cases T4A, T4B and T1D. The more data we assimilate, the better the results; nevertheless we see that the assimilation of surface data (T4A), as satellites can provide, already significantly improves the model state estimation. In a real DA experiment, as well as the atf, the asf is also limiting, because of the necessity of human intervention to collect the biological measurements. We see that assimilating data every 10 boxes (T1D), which approximately corresponds to the spatial repartition of the observations at the Dyfamed site for the year 2000 (Section 2.3), reduces significantly the RMS errors.

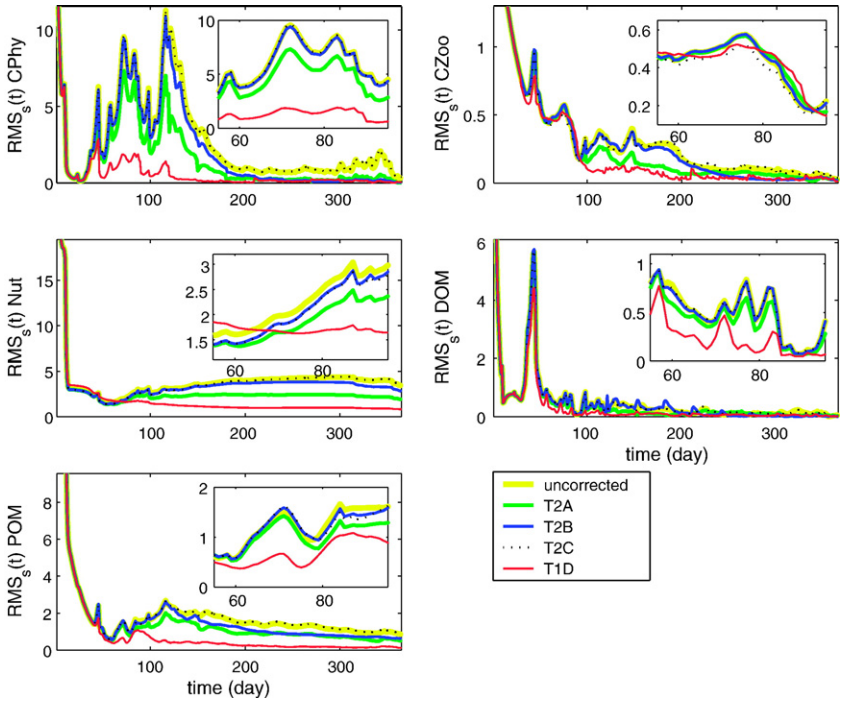

Fig. 9. Twin experiments T2: influence of the composition of the estimation vector — Space-averaged RMS^s errors along the year for the variables CPhy, CZoo, Nut, DOM and POM.

3.2.5. Ensemble size impact — T5

As we suppose intuitively, the use of a larger ensemble improves its representation and thus the statistics that we can derive from it (*e.g.* the mean, the (co)variance). Unfortunately, representation improvement is proportional to the square root of the number of members, whereas the computational load is directly proportional to the number of members, so we have to make the best possible compromise. For these one-year simulations, we choose a 100-member ensemble

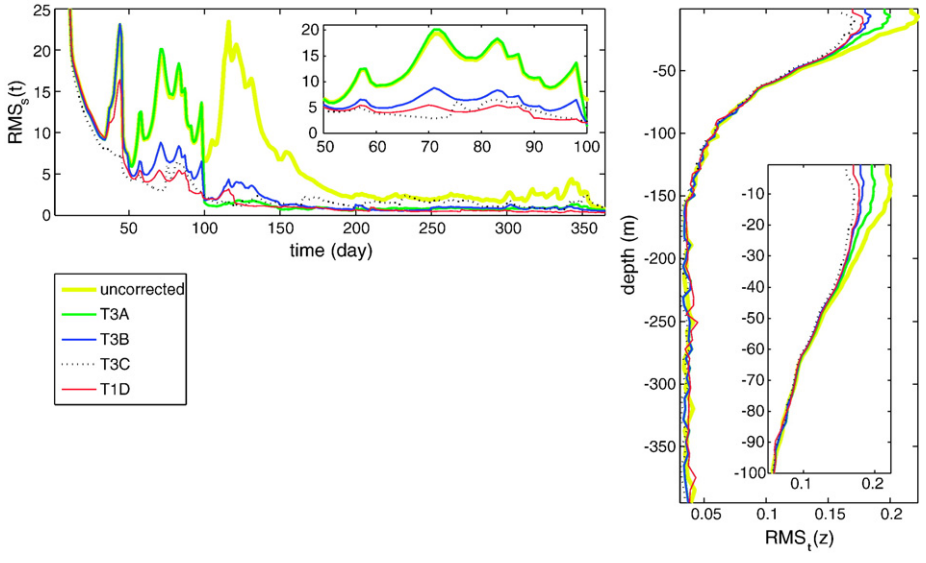

Fig. 10. Twin experiments T3: influence of the assimilation time frequency — Space-averaged RMS^s (left) and time-averaged RMS^t (right) errors for the main state variables.

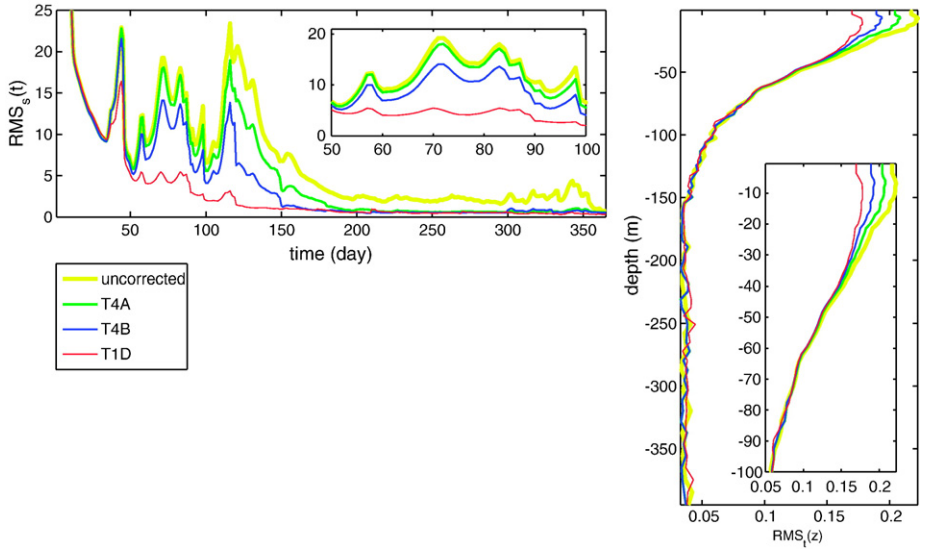

Fig. 11. Twin experiments T4: influence of the assimilation spatial frequency — Space-averaged RMS^s (left) and time-averaged RMS^t (right) errors for the main state variables.

because of its good performance in relation to its relatively affordable cost. Actually, the decrease of the RMS difference for the whole run with a larger ensemble was minor to the square root of the ensemble size; this confirms our choice of a 100-member ensemble.

We present in Fig. 12 on one hand, the space-averaged RMS^s (Eq. (11)) error and on the other hand, the absolute difference between the predicted RMS_j^{ens} (Eq. (13)) and the actual RMS^s errors for the variable CPhy2, the dominant variable of the model. We see that the larger the ensemble, the lower both errors, and also the nearer to the actual error is the forecast error. The large difference between the predicted and the actual errors at first time steps comes from the fact that the standard deviation prescribed to initialize the ensemble around the annual mean profile is lower than the difference between this profile and the reference run.

3.2.6. Comparison between the classical EnKF and our implementation

In Sections 2.2.1 and 2.2.2, we have presented two tools used to improve the classical version of the EnKF, Evensen's ensemble subsampling strategy and the lognormal transformation. Here we show comparative plots of the classical and the enhanced EnKF

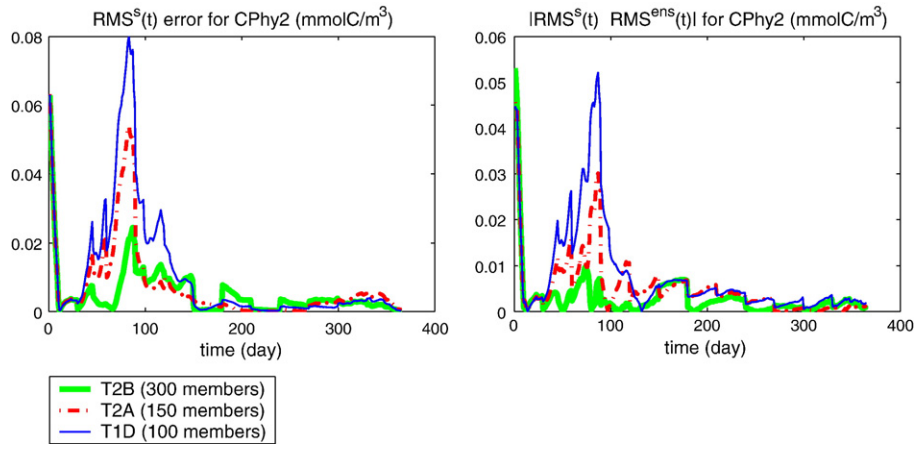

Fig. 12. Twin experiments T5: influence of the ensemble size — Space-averaged RMS^s error (left) and absolute difference between the RMS^s and the RMS^{ens} errors (right) for the dominant variable CPhy2.

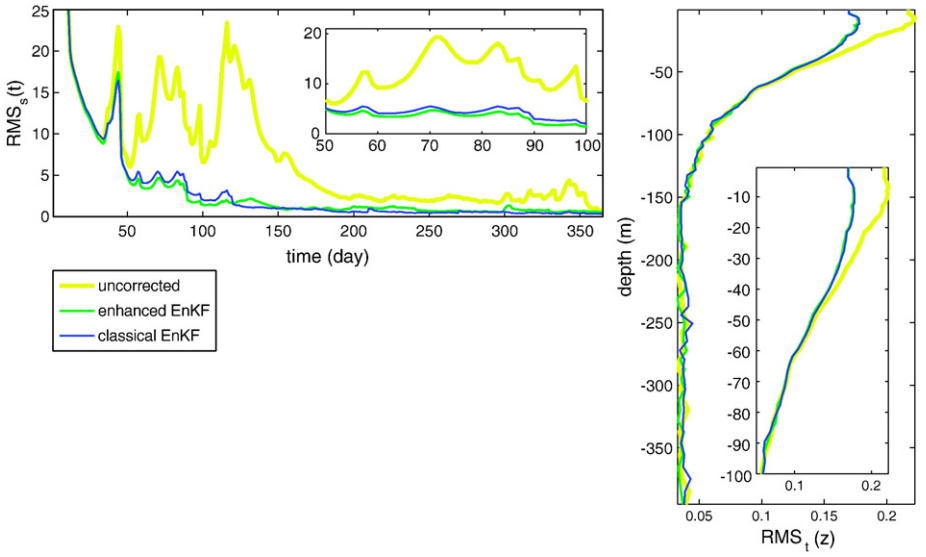

Fig. 13. Comparison between the classical EnKF and our implementation — Space-averaged RMS^s (left) and time-averaged RMS^t (right) errors for the main state variables.

performances when applying the T1D data assimilation protocol. Though the difference is not huge, the results obtained by the EnKF with the two adapted tools are better than the ones obtained with the original implementation of the EnKF, especially during the spring bloom period (see Fig. 13). As already mentioned in Section 2.2.1, the subsampling strategy would have a more positive impact for a model with a larger state vector.

So, if the *a priori* task of determining whether the variable is lognormally distributed has already been performed, or if an automatic procedure to assess the probability distribution is implemented, there is no reason not using the Gaussian anamorphosis. The ensemble subsampling strategy can be recommended in all cases, as long as the computational cost of the singular value decomposition is affordable. It improves the state estimate as well as the error variance estimate.

3.2.7. Conclusions of the twin experiments

The twin experiments have been performed to determine a suitable data assimilation scheme in order to

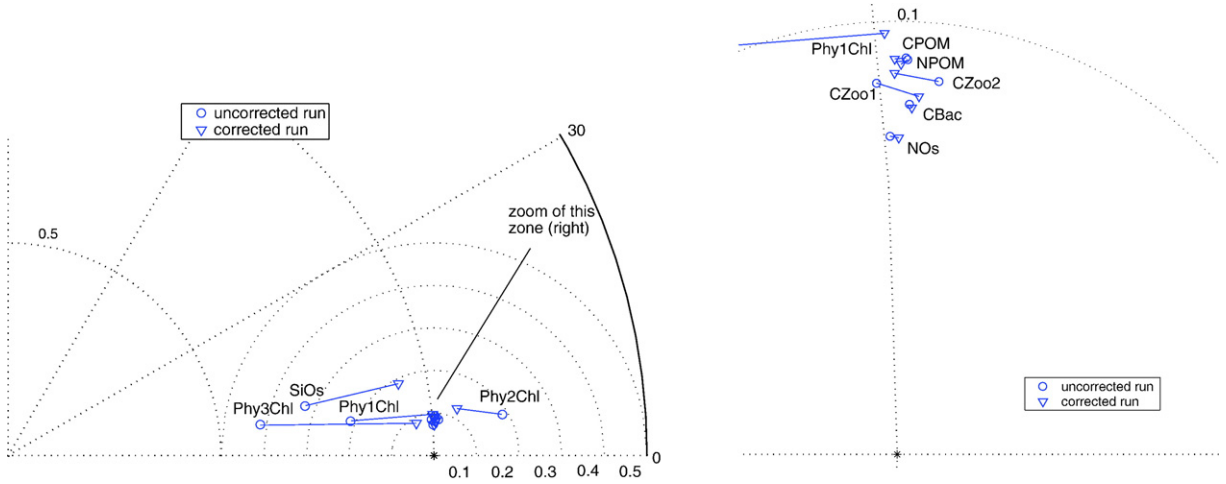

Fig. 14. Adimensionalized Taylor diagram: comparison between the uncorrected run and the assimilation protocol T1D run for the 3 groups of phytoplankton variables (Phy1,2,3Chl) in chlorophyll, the bacteria (CBac), the two smaller groups of zooplankton (CZoo1,2) and the particulate organic matter in carbon and nitrogen (CPOM and NPOM), nitrate (NOs) and silicate (SiOs).

assimilate real data observations. The selected protocol is the T1D: we observe the contents in carbon and nitrogen of the 3 groups of phytoplankton ($nat_{obs}=P_C$ and P_N) and correct, in addition to these variables, the content in carbon of the 3 groups of zooplankton ($nat_{cor}=P_C$, P_N and Z_C), with an assimilation time

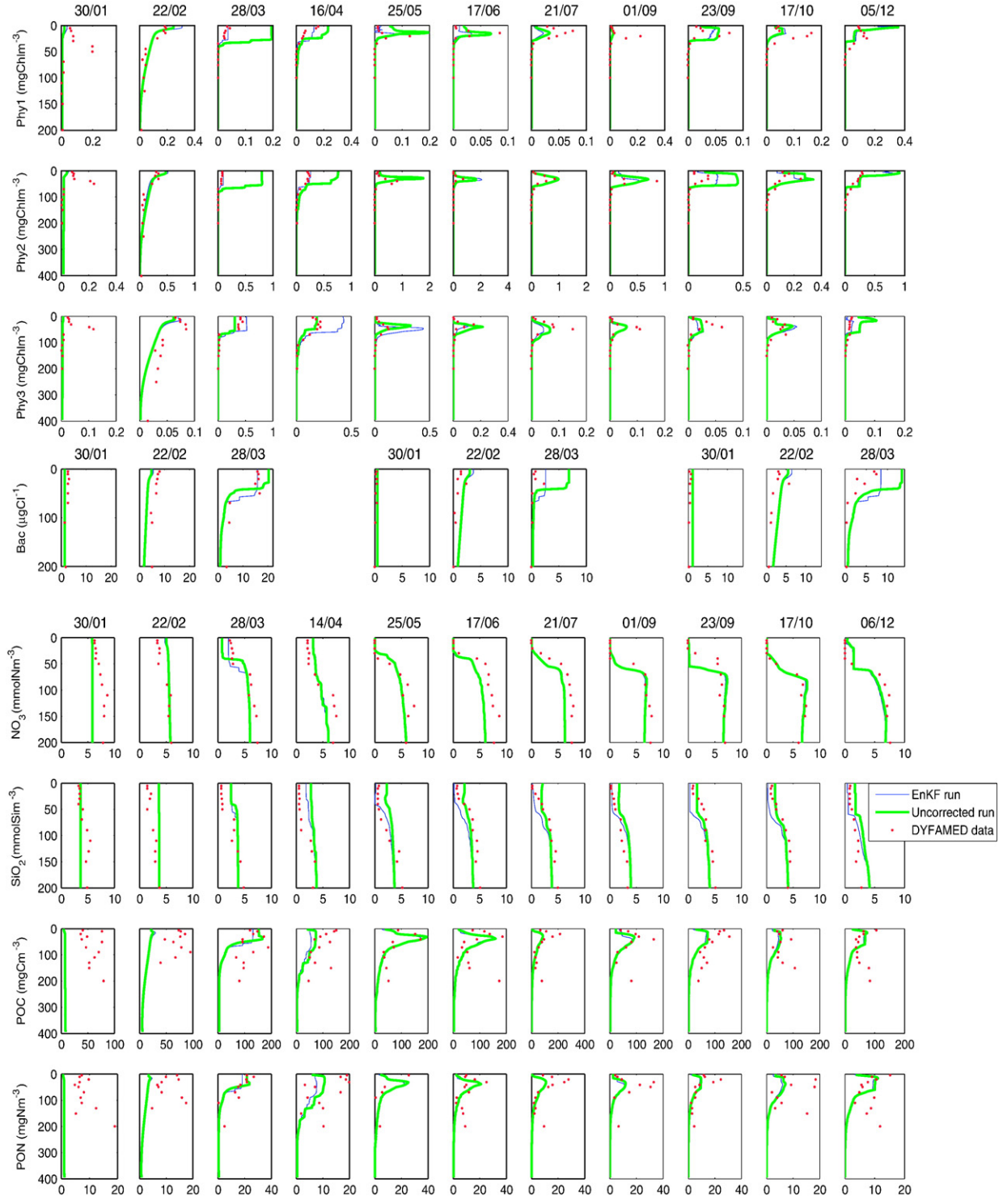

Fig. 15. Real data assimilation experiment results: comparison between the profiles of the free run, the EnKF run (analysis) and the Dyfamed data.

frequency of 30 days (atf=30) and an assimilation space frequency of 10 boxes (asf=10), which correspond approximately to the time and space frequencies available for the data at the Dyfamed site. We use a 100-member ensemble ($ens_{size}=100$) and a relative standard deviation of the observational error of 30% ($\sigma_{obs}=30\%$). We present in Fig. 14 the adimensionalized Taylor diagram obtained by comparing both the uncorrected run and the test case T1D to the reference run. The Taylor diagram, defined in Sections 2.4 and Fig. B.1, is used to visualize the improvements in terms of RMS errors and correlation coefficients. Each point on the diagram represents, for one particular variable, the statistical analysis averaged over the entire one-year simulation. The largest improvements are made for the phytoplankton groups and the silicate. Neither nitrate, nor bacteria changed a lot with the data assimilation.

4. Real data assimilation experiments

The year 2000 was chosen for the real data assimilation experiment ($nat_{obs}=P_C$ and P_N, $nat_{cor}=P_C$, P_N and Z_C and $ens_{size}=100$) because of the availability of the data and the possible comparison with the work of Raick et al. (2005). We have used the Dyfamed data base presented in Section 2.3. Phytoplankton pigments of each phytoplankton group have been converted in carbon biomass using the Chl:C(z,t) ratio extracted from a converged model simulation (Raick et al., 2005) in order to avoid errors and non-linearities due to the units conversion. Because of the lack of information about the observational errors, we consider for each variable and at each depth, a relative standard deviation of 30% of the recorded value. Note that for the corrected and the uncorrected runs, we start from the same set of initial conditions, which is the state of the system at the end of the one year converged simulation.

In Fig. 15 the profiles for the free run and the EnKF run of the chlorophyll content for the 3 groups of phytoplankton, nitrate, silicate and the particulate organic matter contents in carbon and nitrogen, at assimilation dates all along the year, as well as the profiles of the carbon content of bacteria and the 2 groups of zooplankton for which we have observations, at assimilation dates during the first 3 months of the year.

An inspection of the profiles shows that the first real improvement of the EnKF run in comparison to the free run happens in late March. At this moment, the data assimilation allows estimating accurately the chlorophyll concentration of the phytoplankton groups Phy1 and Phy2, as well as the carbon content of the bacteria and the two groups of zooplankton Zoo1 and Zoo2, and nitrate. The EnKF run slightly overestimates Phy3, whereas the free run underestimates it. In April and May, we still have a good estimation of the first 2 groups of phytoplankton, whereas the Phy3 group is largely overestimated; we also note a small improvement of the results for SiO_2 during these months. During the rest of the year, the free run and EnKF run phytoplankton profiles hardly differ, except for Phy1 and Phy2 in September and Phy3 in December, which are improved by the observation and correction through the filter. From April to December the EnKF run and the free run nitrate profiles are not distinguishable. The SiO_2 is mainly underestimated in the EnKF from the beginning of summer until the end of the year due to the increase of CPhy3 in March. The particulate organic matter contents in carbon and nitrogen from the EnKF run do not sensitively differ from the ones obtained during the uncorrected run.

In Fig. 16, we show the RMS^s and RMS^t corresponding to the uncorrected and to the corrected runs. For phytoplankton groups, we see that the forecast does not largely improve the results, in opposition to the analysis, which has generally a smaller RMS^s than the uncorrected run. The contribution of the data assimilation experiment to the estimation of nitrate and particulate organic matter contents in carbon and in nitrogen are not significant. Silicate is better estimated during winter and spring but worse during the rest of the year. For bacteria and zooplankton groups, results are improved by the EnKF run in March but we only have data for the first 3 months of year 2000. By inspection of the RMS^t profiles we see that the assimilation of the biological measurements is quite beneficial for Phy1 and Phy2, whereas the error on Phy3 is larger than the one obtained by an uncorrected run. For the other variables, we note a small general improvement of the results given by the EnKF, in comparison to the free run results.

5. Comparison with the SEEK filter

We finally compare the real data assimilation experiment performed by the EnKF and by a SEEK filter with a fixed basis (Raick et al., in press). For each filter, we choose its best assimilation protocol, *i.e.* the T1D test case for the EnKF (observation of CPhy1,2,3 and NPhy1,2,3; correction of CPhy1,2,3, NPhy1,2,3 and CZoo1,2,3; ensemble of 100 members and relative standard deviation of the observational error of 30%) and for the SEEK filter we observe and correct CPhy1,2,3, letting the dynamics update all the other variables, we use 10 EOFs to represent the error subspace and a forgetting factor of 0.75 (Raick et al., in press).

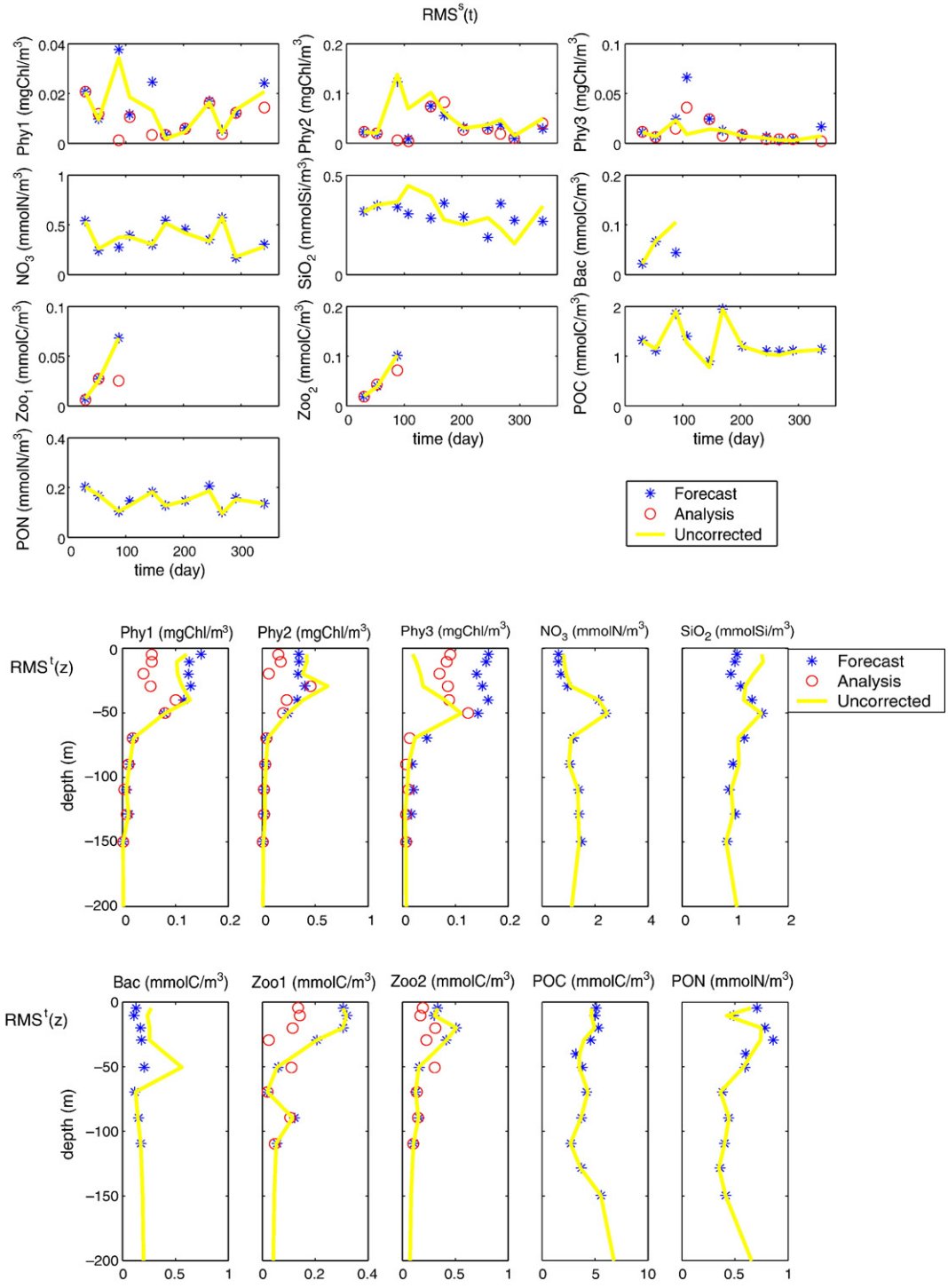

Fig. 16. Real data assimilation experiment results: comparison between the RMS^s (above) and RMS^t (below) of the free run and the EnKF run (forecast and analysis).

Curves being very near from each other, we do not present the concentrations profiles of the different variables; hence we prefer just use the error measurement tools. We show in Figs. 17 and 18 the RMS^s and RMS^t errors for the EnKF and the SEEK runs. Time evolution of Phy1,2 and Zoo1,2 errors are globally

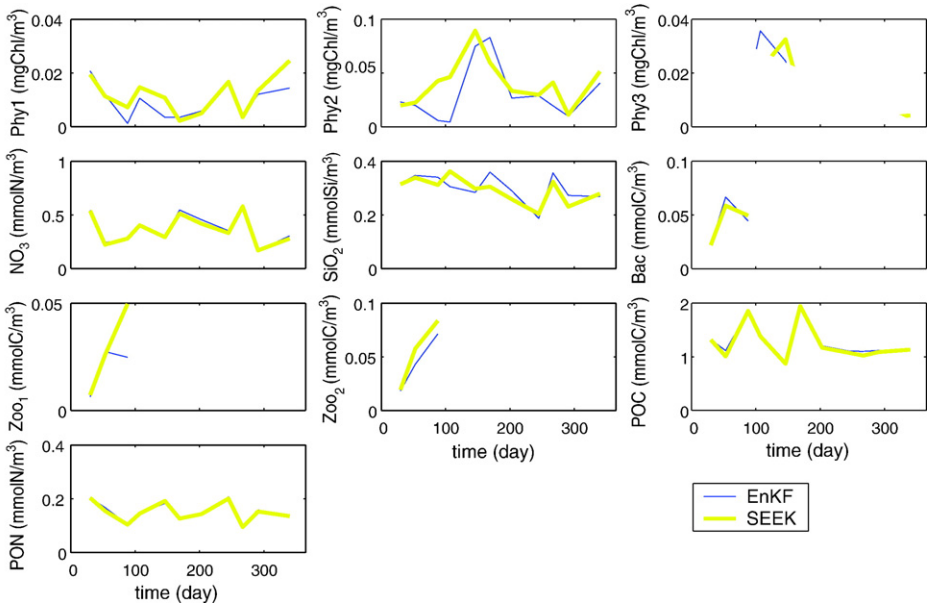

Fig. 17. Real data assimilation experiment results: comparison between the RMS^s of the EnKF run and the SEEK run.

smaller for the EnKF than for the SEEK filter. Besides, the Phy3, NO_3, SiO_2, Bac, POC and PON RMS^s differences are equivalent for both filters; for instance, the concentration of microphytoplankton in spring is best estimated by the SEEK but the tendency is reversed in summer. By inspection of the RMS^t profiles, we see that Phy1 and Phy2 results are better in the whole water column for the EnKF than for the SEEK, whereas for Phy3, we can not draw general conclusion; the estimation of the concentrations in microphytoplankton

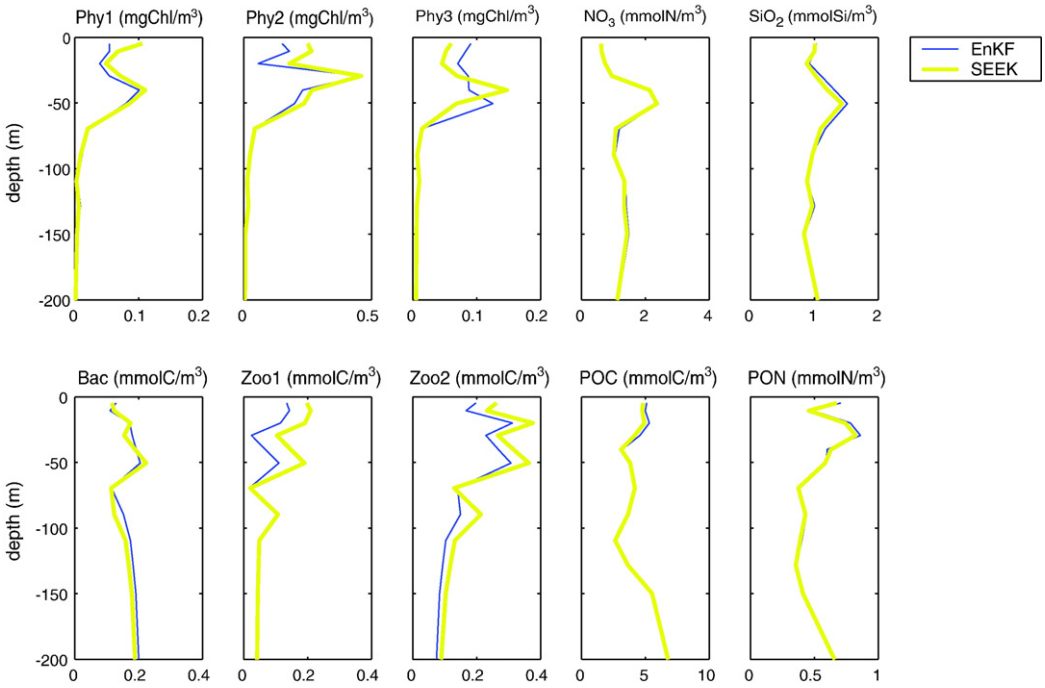

Fig. 18. Real data assimilation experiment results: comparison between the RMS^t of the EnKF run and the SEEK run.

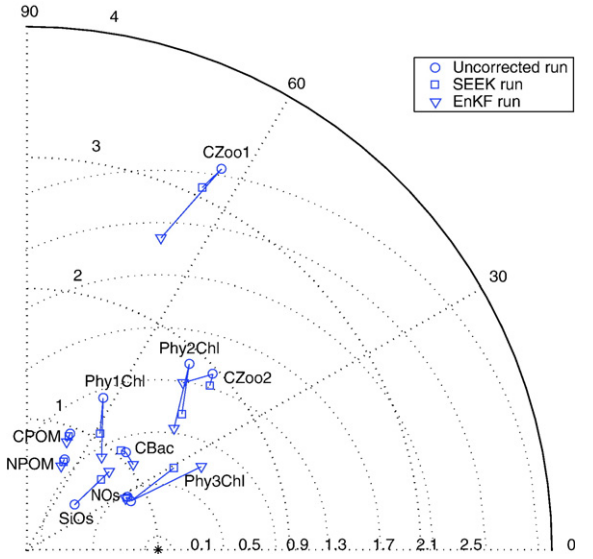

Fig. 19. Real data assimilation experiment results: adimensionalized Taylor diagram comparing the EnKF to the SEEK filter.

in the first 30 m is better for the SEEK and between 30 m and 50 m better for the EnKF. Nitrate, silicate, particulate organic matter contents in carbon and nitrogen estimates are very similar for both the data assimilation techniques. Again, we see the advantage of the EnKF over the SEEK in estimating the concentrations in carbon of the 2 groups of zooplankton.

Finally, we present a comparative adimensionalized Taylor diagram for the real data experiment performed with each filter. We note that the two assimilation protocols yield a general improvement for the Phy1Chl and Phy2Chl estimates, by significantly increasing the correlation R, reducing the normalized centered pattern RMS difference $\hat{E}'$ and approaching the value of 1 for the normalized standard deviation $\hat{\sigma}_f$. Phy3Chl estimates of the EnKF and the SEEK filters are both worse than the free run solution. For CZoo1, the similarity remains unchanged but both the variability and the amplitude of the error are improved in comparison to the uncorrected run, especially for the EnKF; for CZoo2, R is slightly increased for the SEEK and slightly decreased for the EnKF, whereas both $\hat{E}'$ and $\hat{\sigma}_f$ are improved for the experiments. For POC, PON and NOs we do not perceive a real difference in the results, neither between the filters, nor between each filter and the free run. For SiOs, we just note an improvement of the variability representation. The CBac variable is globally (R, $\hat{E}'$ and $\hat{\sigma}_f$) improved by the EnKF run and hardly changed by the SEEK run.

At last, we can conclude that, regarding only the results, the EnKF works slightly better than the SEEK filter (Fig. 19). However, the computation of the solution with the EnKF is far more expensive than with the SEEK; typically, a one-year run with 11 data assimilation dates, takes about 2 min to be performed with the SEEK and 1 h with the EnKF, because of the propagation of 100 members instead of 1.

6. Conclusion

An Ensemble Kalman filter (EnKF) has been applied on a 1-D coupled hydrodynamic-ecosystem model of the Ligurian Sea. In order to improve its performance, it has been equipped with a subsampling strategy to better represent the statistics of the initial ensemble, given a limited number of members, and a pre-analysis step, consisting in applying an appropriate transformation to fulfill as well as possible the Gaussian assumption, required to reach the optimality of the filter. Twin experiments were performed in order to test the performance of the method and to determine a suitable data assimilation (DA) protocol with a view to perform real *in-situ* DA experiments. The ideal DA protocol is quite intuitive, the more data we assimilate and the more precise they are, the better the results. In the case we don't have measurements of many different variables (*i.e.* phytoplankton contents in carbon or nitrogen, nitrates, silicates, zooplankton contents in carbon and nitrogen, *etc.*) regularly spaced in time, the only observation and correction of phytoplankton contents in carbon, letting the dynamics of the model updating the other variables, the results were already positively affected.

In addition, the comparison of two versions of the filter has shown us that the ensemble subsampling strategy is slightly beneficial, as well as the lognormal transformation, which requires however an *a priori* work. However, the improved sampling strategy might be more beneficial for a model with a larger state space.

Then, we assimilate data of year 2000 at station Dyfamed in order to improve the model results. We observed that the larger improvements were happening during the late bloom, which is actually the time period of greatest variations of phytoplankton concentrations. The comparison of this hindcast of the Ligurian ecosystem for this year with the one realized with a SEEK filter, has highlighted a slightly better ability of the EnKF to catch the non-linearity of the system dynamics, but with a very higher computational cost.

Acknowledgements

The research was conducted thanks to the EU project MERG-CT-2004-012756 SIMPLIC and was supported

by the Belgian Foundation for Scientific Research to which the last author is affiliated. We would like to thank J.-C. Marty for the hydrodynamic and biogeochemical data coming from the Dyfamed station, and METEO France Côte d'Azur for the meteorological data. This is MARE publication 100 and NIOO-KNAW publication 3964.

Appendix A. Member perturbation

For a given state variable or measurement, in order to generate the initial and observation ensemble members, or to perturb the state vector by model errors, as in Eknes and Evensen (2002), we add to the best guess estimate a pseudo random field drawn from a normal distribution, with a zero mean and a prescribed absolute or relative standard deviation. Hence the *j*th ensemble element is given either by

$$v(z)_j = \mu(z) + \sigma(z)\mathcal{N}_{(0,1)j} \tag{A.1}$$

in the case of a prescribed absolute standard deviation, or by

$$v(z)_j = \mu(z)(1 + \sigma(z)_{\mathrm{rel}}\mathcal{N}_{(0,1)j}), \tag{A.2}$$

in the other case, with σ having the same units as μ and σ_{rel} being dimensionless. The dependence of the standard deviation with z allows us to take into account a smoothing of the observational error with depth. Here we use a decrease of $\sigma(z)$ and $\sigma_{\mathrm{rel}}(z)$ with z according to $\exp\left(-\left(\frac{z}{\mathcal{L}}\right)^2\right)$ where $\mathcal{L}$ denotes a decorrelation length appropriate to the considered variable; typically it corresponds to the nutricline for the nutrients, the euphotic layer for the plankton variables, *etc.*

Appendix B. Taylor diagram

Let us denote by $f_{i,j}$ the forecast and by $r_{i,j}$ the reference fields defined at a given point of the vertical domain (the *i*th layer) and at a given time (the *j*th day) and consider the following statistical quantities:

1. their space and time variability can be assessed by their respective standard deviations σ_{f} and σ_{r}

$$\sigma_{\mathrm{f}} = \sqrt{\sum_{i=1}^{N_t}\sum_{i=1}^{N_z}\frac{dt_j}{T}\frac{dz_i}{D}(f_{i,j}-\overline{f})^2}, \tag{B.1}$$

$$\sigma_{\mathrm{r}} = \sqrt{\sum_{i=1}^{N_t}\sum_{i=1}^{N_z}\frac{dt_j}{T}\frac{dz_i}{D}(r_{i,j}-\overline{r})^2}, \tag{B.2}$$

where N_t is the number of discrete points in time, N_z the number of discrete points in space, dt_j the *j*th time interval, dz_i the thickness of the *i*th layer, $D=\Sigma_{i=1}^{Nz}\, dz_i$ the size of the spatial domain, $T=\Sigma_{j=1}^{Nt}\, dt_j$ the duration of the simulation, $\overline{f}$ and $\overline{r}$ the time and space-averaged

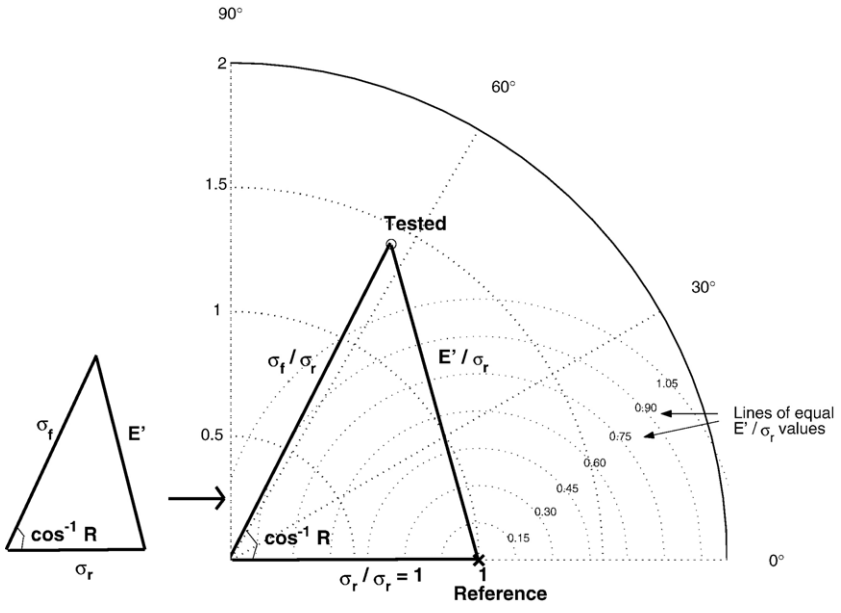

Fig. B.1. Normalized Taylor diagram reprinted from Raick et al. (2005) and originally designed by Taylor (2001).

forecast and reference fields; 2. the amplitude of the errors between the two fields can be compared by computing the Root-Mean-Square (RMS) difference

$$\mathrm{RMS} = \sqrt{\sum_{i=1}^{N_t}\sum_{i=1}^{N_z}\frac{dt_j}{T}\frac{dz_i}{D}\left(f_{i,j}-r_{i,j}\right)^2}; \tag{B.3}$$

3. their similarity is given by the correlation coefficient R

$$R = \frac{\mathrm{cov}(f,r)}{\sigma_f\sigma_r} = \frac{\sum_{j=1}^{N_t}\sum_{i=1}^{N_z}\frac{dt_j}{T}\frac{dz_i}{D}\left(f_{i,j}-\overline{f}\right)\left(r_{i,j}-\overline{r}\right)}{\sigma_f\sigma_r}. \tag{B.4}$$

All of the above statistics are actually related if the RMS error is decomposed into a mean component $\bar{E}$ and a centered pattern RMS difference E' according to

$$\mathrm{RMS}^2 = \overline{E}^2 + E'^2, \tag{B.5}$$

where $\bar{E}=\overline{f}-\overline{r}$ and

$$E' = \sqrt{\sum_{i=1}^{N_t}\sum_{i=1}^{N_z}\frac{dt_j}{T}\frac{dz_i}{D}\left(\left(f_{i,j}-\overline{f}\right)-\left(r_{i,j}-\overline{r}\right)\right)^2} \tag{B.6}$$

With these definitions we can write the relationship between these quantities in a form common to the law of cosines $E'^2=\sigma_f^2+\sigma_r^2-2\sigma_f\sigma_r R$, that can lead to a geometrical representation of the correspondence between these fields. These statistics can then be summarized in a single plot, the Taylor diagram (Taylor, 2001). The correlation coefficient R and the centered pattern RMS difference E' between the two fields, along with the standard deviation are all indicated by a single point on a 2D plot (see Fig. B.1 — left). Because of the different units of measure, the statistics of the different variables have to be non- dimensionalized before appearing on the same diagram. The centered pattern RMS and the two standard deviations are normalized by σ_r:

$$\hat{E}' = \frac{E'}{\sigma_r} \quad ; \quad \hat{\sigma}_f = \frac{\sigma_f}{\sigma_r} \quad ; \quad \hat{\sigma}_r = 1 \tag{B.7}$$

This leaves the correlation coefficient unchanged and yields a normalized diagram as we can see in Fig. B.1 (right). Because normalized by itself, $\hat{\sigma}_r$ is always plotted at unit distance from the origin along the abscissa. The correspondence of a particular variable to the data can be made by inspecting the position of the corresponding point on the diagram. Its azimuthal position arccos R indicates the correlation: the smaller is the angle, the better is the correlation. The distance between the model output point and the point representing the reference provides information about the centered pattern RMS difference E'. Its distance from origin indicates its normalized standard deviation $\hat{\sigma}_f$, a value of 1 indicates that the forecast and reference fields have similar standard deviations.

References

Allen, J.I., Eknes, M., Evensen, G., 2002. An Ensemble Kalman filter with a complex marine ecosystem model: hindcasting phytoplankton in the Cretan Sea. Annales Geophysicae 20, 1–13.

Bertino, L., Evensen, G., Wackernagel, H., 2003. Sequential data assimilation techniques in oceanography. International Statistical Review 71, 223–241.

Bethoux, J.P., Morin, P., Chaumery, C., Connan, O., Gentili, B., Ruiz-Pino, D.P., 1998. Nutrients in the Mediterranean Sea, mass balance and statistical analysis of concentrations with respect to environmental change. Marine Chemistry 63 (1–2), 155–169.

Bethoux, J.P., Morin, P., Ruiz-Pino, D.P., 2002. Temporal trends in nutrients ratios: chemical evidence of Mediterranean ecosystem changes driven by human activity. Deep-Sea Research. Part 2. Topical Studies in Oceanography 49, 2007–2016.

Burgers, G., van Leeuwen, P.J., Evensen, G., 1998. Analysis scheme in the Ensemble Kalman filters. Monthly Weather Review 126, 1719–1724.

Brasseur, P., 2005. Ocean data assimilation using sequential methods based on the Kalman Filter. In: Verron, J., Chassignet, E. (Eds.), GODAE, an Integrated View of Oceanography: Ocean Weather Forecasting in the 21st Century. Kluwer Academic Press.

Delhez, E.J.M., Grégoire, M., Nihoul, J.C.J., Beckers, J.-M., 1999. Dissection of the GHER turbulence closure scheme. Journal of Marine Systems 21, 379–397.

Eknes, M., Evensen, G., 2002. An Ensemble Kalman filter with a 1-D marine ecosystem model. Journal of Marine Systems 36, 75–100.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research 99, 10143–10162.

Evensen, G., 2003. The Ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dynamics 53, 343–367.

Evensen, G., 2004. Sampling strategies and square root analysis schemes for the EnKF. Ocean Dynamics 54, 539–560.

Jaswinski, A.H., 1970. Stochastic Processes and Filtering Theory. Academic Press.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Transactions of the ASME—Journal of Basic Engineering 82, 13–45.

Lacroix, G., Grégoire, M., 2002. Revisited ecosystem model (MODECOGeL) of the Ligurian Sea: seasonal and interannual variability due to atmospheric forcing. Journal of Marine Systems 37 (4), 229–258.

Marty, J.-C., 2002. The Dyfamed time-series program (French-JGOFS). Deep-Sea Research. Part 2. Topical Studies in Oceanography 49, 1963–1964.

Marty, J.-C., Chiaverini, J., 2002. Seasonal and interannual variations in phytoplankton production at Dyfamed time-series station, northwestern Mediterranean Sea. Deep-Sea Research. Part 2. Topical Studies in Oceanography 49, 2017–2030.

Marty, J.-C., Chiaverini, J., Pizay, M.-D., Avril, B., 2002. Seasonal and interannual dynamics of nutrients and phytoplankton pigments in the western Mediterranean Sea at the Dyfamed time-series station (1991–1999). Deep-Sea Research. Part 2. Topical Studies in Oceanography 49, 1965–1985.

Natvik, L.J., Evensen, Geir, 2003. Assimilation of ocean colour data into a biochemical model of the North Atlantic. Part I Data assimilation experiments. Journal of Marine Systems 40–41, 127–153.

Nihoul, J.C.J., Djenidi, S., 1987. Perspectives in Three-dimensional Modelling of the Marine System. Elsevier, Amsterdam.

Pham, D.T., Verron, J., Roubaud, M.-C., 1998. A singular Evolutive Kalman filter for data assimilation in oceanography. Journal of Marine Systems 16 (3–4), 323–340.

Raick, C., Delhez, E.J.M., Soetaert, K., Grégoire, M., 2005. Study of the seasonal cycle of the biogeochemical processes in the Ligurian Sea using a 1D interdisciplinary model. Journal of Marine Systems 55, 177–203.

Raick, C., Alvera-Azcarate, A., Barth, A., Brankart, J.-M., Soetaert, K., Grégoire, M., in press. Application of a SEEK filter to a 1D biogeochemical model of the Ligurian Sea: twin experiments and real in-situ data assimilation. Journal of Marine Systems, doi:10.1016/j.jmarsys.2005.06.006.

Soetaert, K., deClippere, V., Herman, P.M.J., 2002. Femme, a flexible environment for mathematically modelling the environment. Ecological Modelling 151, 177–193.

Tamburini, C., Garcin, J., Ragot, M., Bianchi, A., 2002. Biopolymer hydrolysis and bacterial production under ambient hydrostatic pressure through a 2000 m water column in the NW Mediterranean. Deep-Sea Research. Part 2. Topical Studies in Oceanography 49, 2109–2123.

Tanaka, T., Rassoulzadegan, F., 2002. Full-depth profile (0–2000 m) of bacteria, heterotrophic nanoflagellates and ciliates in the NW Mediterranean Sea: vertical partitioning of microbial trophic structures. Deep-Sea Research. Part 2. Topical Studies in Oceanography 49, 2093–2107.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research 106, 7186–7192.

van Leeuwen, P.J., Evensen, G., 1998. Data assimilation and inverse methods in terms of a probabilistic formulation. Monthly Weather Review 124, 2898–2913.

Vidussi, F., Marty, J.-C., Chiaverini, J., 2000. Phytoplankton pigment variations during the transition from spring bloom to oligotrophy in the northwestern Mediterranean Sea. Deep-Sea Research. Part 1. Oceanographic Research Papers 47, 423–445.

Vidussi, F., Claustre, H., Manca, B.B., Luchetta, A., Marty, J.-C., Chiaverini, J., 2001. Phytoplankton pigment distribution in relation to upper thermocline circulation in the eastern Mediterranean Sea during winter. Journal of Geophysical Research 106 (C9), 19,939–19,956.