Ocean Sci., 8, 587–602, 2012
www.ocean-sci.net/8/587/2012/
doi:10.5194/os-8-587-2012

Ocean Science

Estimation of positive sum-to-one constrained zooplankton grazing preferences with the DEnKF: a twin experiment

E. Simon[1,2], **A. Samuelsen**[1,2], **L. Bertino**[1,2], **and D. Dumont**[3]

[1]Nansen Environmental and Remote Sensing Center, Norway
[2]Bjerknes Center for Climate Research, Norway
[3]Institut des sciences de la mer, Université du Québec à Rimouski, Canada

Correspondence to: E. Simon (ehouarn.simon@nersc.no)

Received: 28 February 2012 – Published in Ocean Sci. Discuss.: 19 March 2012
Revised: 18 June 2012 – Accepted: 28 June 2012 – Published: 8 August 2012

Abstract. We consider the estimation of the grazing preferences parameters of zooplankton in ocean ecosystem models with ensemble-based Kalman filters. These parameters are introduced to model the relative diet composition of zooplankton that consists of phytoplankton, small size-classes of zooplankton and detritus. They are positive values and their sum is equal to one. However, the sum-to-one constraint cannot be guaranteed by ensemble-based Kalman filters when parameters are bounded. Therefore, a reformulation of the parameterization is proposed. We investigate two types of variable transformations for the estimation of positive sum-to-one constrained parameters that lead to the estimation of a new set of parameters with normal or bounded distributions. These transformations are illustrated and discussed with twin experiments performed with the 1-D coupled model GOTM-NORWECOM with Gaussian anamorphosis extensions of the deterministic ensemble Kalman filter (DEnKF).

1 Introduction

The development of numerical ocean biogeochemical models over the last two decades has led to more and more complex representations of the interactions between the different trophic levels, notably between different plankton species at the base of the food chain. While the diet of zooplankton is relatively simply represented in the earliest NPZD models – the unique zooplankton group (Z) is feeding only on the unique phytoplankton group (P) (see for example Evans and Parslow (1985)) – the addition of multiple plankton functional types (PFT) for the phyto- and zooplankton aiming at representing different plankton functional groups in the ecosystem (e.g. diatoms, calcifying algae or microzooplankton) leads to more complex diets and grazing preferences must be added. These parameters are always positive, and, although not compulsory, usually add up to one. We refer to the review of Gentleman et al. (2003) for more details concerning the common mathematical formulations of the zooplankton grazing in ocean biological models and their impact on model dynamics.

Grazing preferences specify the direction of the feeding in the space of foods, and so the direction of the transfer from PFTs (the food) to the zooplankton PFTs (the feeder). Therefore, their impact on the distribution of the different PFTs obtained from a model simulation can be significant. For example, Buitenhuis et al. (2010) observed in their global biogeochemical model that "the phytoplankton functional type distributions and the proportions of primary production that are exported or remineralized" were sensitive to the microzooplankton grazing preferences. In the same way, Buitenhuis et al. (2006) conclude their work by suggesting that the representation of mesozooplankton would notably benefit from the improvement of their grazing preferences by taking into account the food quality. For large-scale applications like configurations covering a whole ocean basin, this results in the potential need of a fine spatial tuning of the grazing preferences in order to take into account the adaptation of zooplankton species to their local environments (Gentleman et al., 2003). Direct measurements of grazing preferences for the different zooplankton species would help to optimize the model parameters representing these preferences. However, field data are sparse; the information provided by

the experiments realized in the laboratory does not cover the large spectrum of conditions found in nature (Buitenhuis et al., 2010), and available observations might not be consistent with each other (Buitenhuis et al., 2006).

Multivariate data assimilation methods like ensemble-based Kalman filters make possible the estimation of variables and parameters that are not observed. State variables and parameters can be estimated simultaneously simply by augmenting the state vector with the parameters to estimate (Anderson, 2001; Evensen, 2009). However, the efficient application of ensemble-based data assimilation methods like the ensemble Kalman filter (EnKF; Evensen, 1994, 2003) to ocean ecosystem models is a challenging issue. Beside the nonlinearity of the model, most variables and parameters are strictly positive, producing non-Gaussian state and parameter distributions thereby breaking an important assumption of the linear analysis, and leading to a loss of optimality of Kalman filters. A solution to perform Kalman filter estimation of non-Gaussian variables is the introduction of nonlinear changes of variables – called anamorphosis functions – in order to realize the analysis step with Gaussian distributed transformed variables (Bertino et al., 2003). This approach has proven to be easily applicable in realistic configurations (Simon and Bertino, 2009) and allows the estimation of biased parameters (Doron et al., 2011; Zhou et al., 2011; Simon and Bertino, 2012).

In this study, we focus on the problem of estimating positive sum-to-one constrained parameters. Our aim is to assess the ability of ensemble-based Kalman filters to estimate zooplankton grazing preferences in ocean biogeochemical models. To overcome the issues that ensemble-based Kalman filters cannot guarantee the sum-to-one constraint when a constraint of positiveness applies on the parameters, we investigate two reformulations for which these two constraints are implicit.

The outline of the paper is as follows. We present the different changes of variables for the estimation of positive sum-to-one constrained parameters in Sect. 2. We describe our experimental framework in Sect. 3. Results of the methods are discussed in Sect. 4, and we present our conclusion in Sect. 5.

2 Estimation of positive sum-to-one constrained parameters with ensemble-based Kalman filters

In this section, we describe the general problem of estimating positive sum-to-one constrained parameters with ensemble-based Kalman filters and the issues raised by these constraints. We present a formulation previously suggested by Gelman (1995) to estimate positive sum-to-one constrained parameters in the framework of pharmacokinetics (Gelman et al., 1996). Since the number of food preferences that need to be calibrated can be large in complex ocean biological models (numerous different feeding and fed species), we aim at reducing the number of parameters to estimate. For this reason, we suggest a new formulation that introduces a change of variables based on hyperspherical coordinates.

2.1 Definition of the problem

Let $(\pi_i)_{i=1:N}$ be the N parameters that we wish to estimate. They are positive:

$$\forall i = 1:N \quad \pi_i \geqslant 0, \tag{1}$$

and their sum is equal to one:

$$\sum_{i=1}^{N} \pi_i = 1. \tag{2}$$

They can be estimated with ensemble-based Kalman filter by augmenting the analysis state vector with these parameters. Unfortunately, the conservation of linear properties intrinsic to the ensemble Kalman filter (Evensen, 2003) is not guaranteed for the parameters due to the constraint of positiveness. The truncation of negative values that results from the Kalman analysis can lead to parameter estimates that do not respect the linear sum-to-one property (Eq. 2). Even if the Gaussian anamorphosis extension of ensemble-based Kalman filters makes the estimation possible of positive parameters (Simon and Bertino, 2012), nonlinear transformations do not ensure that they still sum to one.

2.2 Dirichlet distribution and Gelman's formulation

A prior distribution for N positive random parameters with the sum-to-one constraint is the Dirichlet distribution of order N. The $(\pi_i)_{i=1:N}$ can be obtained from N independent gamma distributed random variables $(\phi_i)_{i=1:N}$ as follows:

$$\forall i = 1:N, \pi_i = \frac{\phi_i}{\sum_{k=1}^{N} \phi_k} \quad \text{with} \quad \phi_i \sim \Gamma(\theta_i, 1). \tag{3}$$

Then, the parameters $(\phi_i)_{i=1:N}$ are estimated by assimilating observation with ensemble-based Kalman filters, and the values of the original parameters $(\pi_i)_{i=1:N}$ are obtained from equation (3). Because the parameters $(\phi_i)_{i=1:N}$ are not Gaussian distributed, we suggest to transform them with the Gaussian anamorphosis during the analysis.

Another possibility is to substitute the gamma distribution by the log-normal distribution as suggested by Gelman (1995):

$$\forall i = 1:N, \pi_i = \frac{e^{\phi_i}}{\sum_{k=1}^{N} e^{\phi_k}} \quad \text{with} \quad \phi_i \sim \mathcal{N}(\theta_i, \Sigma_i). \tag{4}$$

In that case, the $(\phi_i)_{i=1:N}$ fulfill the Kalman filtering assumption of Gaussian distributed variables and do not require anamorphosis.

Due to the symmetrical roles played by the parameters $(\phi_i)_{i=1:N}$ in both formulations, the estimation of the parameters $(\pi_i)_{i=1:N}$ is insensitive to the mapping between the parameters $(\phi_i)_{i=1:N}$ and $(\pi_i)_{i=1:N}$ in the change of variables. However, these approaches do not allow for parameters $(\pi_i)_{i=1:N}$ equal to zero, meaning that one food type could not be completely removed by assimilation. This might be undesirable in large-scale configurations for which the diet composition can significantly change from one region to another.

2.3 The hyperspherical coordinate system

The $(\pi_i)_{i=1:N}$ can be seen as a position vector in the Cartesian coordinates of a point π in $\mathbb{R}^N$. A natural idea is to represent this point in another coordinate system. We suggest to introduce $N-1$ angles $(\phi_i)_{i=1:N-1}$ to represent π in the hyperspherical coordinate system that generalizes the spherical coordinate in dimension N. The use of this coordinate system to remove constraints of sum has also been introduced for geometrical applications (Nurmela, 1995). An analogy with a coordinate system describing a point on a sphere shows that 2 angles, longitude and latitude, are required to characterize the position of a point on the surface in 3 dimensions.

$$\begin{cases} \pi_1 = \cos^2(\frac{\pi}{2}\phi_1) \\ \forall i = 2:N-1, \\ \pi_i = \prod_{k=1}^{i-1} \sin^2(\frac{\pi}{2}\phi_k)\cos^2(\frac{\pi}{2}\phi_i) \\ \pi_N = \prod_{k=1}^{N-2} \sin^2(\frac{\pi}{2}\phi_k)\sin^2(\frac{\pi}{2}\phi_{N-1}) \end{cases} \quad (5)$$

with $(\phi_i)_{i=1:N-1}$ $N-1$ random variables distributed on the segment line $[0,1]$. By definition, the $(\pi_i)_{i=1:N}$ are positive and it can be easily shown that their sum is equal to one.

Again, we suggest to transform the parameters $(\phi_i)_{i=1:N-1}$ with the Gaussian anamorphosis functions during the Kalman filter analysis.

One benefit of this approach is the reduction of the number of parameters to estimate from N to $N-1$, the $(\phi_i)_{i=1:N-1}$ instead of the $(\pi_i)_{i=1:N}$. This is certainly useful for complex systems involving numerous unknown parameters to estimate. However, the estimated values of the $(\pi_i)_{i=1:N}$ can be sensitive to the choice of the mapping to the $(\phi_i)_{i=1:N-1}$ due to the asymmetry of the transformation. For our specific problem of estimating zooplankton grazing preferences, it means that the results might depend on the choice of assigning the types of food to the $(\pi_i)_{i=1:N}$.

2.4 Prior distribution of the $(\phi_i)_{i=1:N-1}$ in the hyperspherical coordinate system

A significant issue lies in the choice of the distributions of the parameters $(\phi_i)_{i=1:N-1}$. When the distributions of the parameters $(\pi_i)_{i=1:N}$ are known or samples are available, the inversion of the hyperspherical coordinate system can provide prior values for the parameters $(\phi_i)_{i=1:N-1}$. This can be done recursively starting from $\phi_1 = \frac{2}{\pi}\arccos(\sqrt{\pi_1})$ and so on. This approach can also be applied to estimate positive sum-to-one constrained state variables evolving in time accordingly to a model dynamics. The variable transformations before and after the analysis using the inverse of Eq. (5) and Eq. (5) guarantee that analyzed variables fulfill both constraints.

Another approach consists in focusing on the direct modeling of the $(\phi_i)_{i=1:N-1}$. This can be an option when too little information on the $(\pi_i)_{i=1:N}$ is available. We suggest to base this choice on the ability to specify prior values and uncertainties for the $(\pi_i)_{i=1:N}$ – their prior expected value $(E[\pi_i])_{i=1:N}$ and variance $(E[(\pi_i - E[\pi_i])^2])_{i=1:N}$ – rather than focusing on their distributions. This leads to the choice of prior values and uncertainties of the $(\pi_i)_{i=1:N}$, for which the parameters of the distributions of the $(\phi_i)_{i=1:N-1}$ will be tuned accordingly. For example, in our particular framework, it would be interesting to start the estimation process with $(\pi_i)_{i=1:N}$ that have the same expected value $\frac{1}{N}$, because this case corresponds to no particular feeding preferences in the diet of the zooplankton species.

We assume that the parameters $(\phi_i)_{i=1:N-1}$ are independent and follow marginal distributions $(\mathcal{D}_i(\Theta_i))_{i=1:N-1}$, which can differ. The prior values for the expectation and variances of the parameters $(\pi_i)_{i=1:N}$ are obtained by an adequate tuning of the $N-1$ parameter sets $(\Theta_i)_{i=1:N-1}$. Let $(m_i)_{i=1:N}$ and $(\sigma_i^2)_{i=1:N}$ be the target means and variances of the $(\pi_i)_{i=1:N}$. Due to the sum-to-one constrain, the expected value and variance of one parameter (π_N without loss of generality) are determined by the choice of the values for the $N-1$ other parameters $(\pi_i)_{i=1:N-1}$.

2.4.1 Specification of the expected values of the $(\pi_i)_{i=1:N}$

The specification of the expected values $(m_i)_{i=1:N-1}$ leads to the resolution of $N-1$ nonlinear equations $(\mathcal{S}_i)_{i=1:N-1}$:

Find $(\Theta_i)_{i=1:N-1}$ such that

$$\begin{cases} \frac{1}{4}(\Phi_{\phi_1}(\pi) + \Phi_{\phi_1}(-\pi)) = m_1 - \frac{1}{2} \\ \forall i = 2:N-1, \\ \frac{1}{4}(\Phi_{\phi_i}(\pi) + \Phi_{\phi_i}(-\pi)) = \frac{m_i}{1 - \sum_{k=1}^{i-1} m_k} - \frac{1}{2} \end{cases} \quad (6)$$

with Φ_{ϕ_i} the characteristic function of the parameter ϕ_i.

The derivation of these equations is detailed in Appendix A.

The existence of solutions to this system of equations $(\mathcal{S}_i)_{i=1:N-1}$ depends on the chosen distributions $(\mathcal{D}_i)_{i=1:N-1}$ and can be found numerically.

2.4.2 Specification of the variances of the $(\pi_i)_{i=1:N}$

The specification of the variances $(\sigma_i^2)_{i=1:N-1}$ leads to the resolution of $N-1$ nonlinear equations (Σ_i):

Find $(\Theta_i)_{i=1:N-1}$ such that

$$\begin{cases} \dfrac{1}{16}(\Phi_{\phi_1}(2\pi)+\Phi_{\phi_1}(-2\pi)) = \\ \qquad -\dfrac{3}{8}+\sigma_1^2+m_1^2-\dfrac{1}{4}(\Phi_{\phi_1}(\pi)+\Phi_{\phi_1}(-\pi)) \\ \forall i = 2:N-1, \\ \dfrac{1}{16}(\Phi_{\phi_i}(2\pi)+\Phi_{\phi_i}(-2\pi)) = -\dfrac{3}{8}-\dfrac{1}{4}(\Phi_{\phi_i}(\pi)+\Phi_{\phi_i}(-\pi)) \\ +\dfrac{\sigma_i^2+m_i^2}{\sum_{k=1}^{i}(-2)^{k-1}(\sigma_{i-k}^2+m_{i-k}^2)\prod_{l=1}^{k-1}\frac{1}{4}(\Phi_{\phi_{i-l}}(\pi)+\Phi_{\phi_{i-l}}(-\pi))} \end{cases} \quad (7)$$

with the conventions $\sigma_0^2+m_0^2=1$ and $\prod_{l=1}^{0}\frac{1}{4}(\Phi_{\phi_{i-l}}(\pi)+\Phi_{\phi_{i-l}}(-\pi))=1$. The values of $(\frac{1}{4}(\Phi_{\phi_i}(\pi)+\Phi_{\phi_i}(-\pi)))_{i=1:N-1}$ depend on the $(m_i)_{i=1:N-1}$ only and are given by Eq. (6). The derivation of these equations is detailed in Appendix B.

Again, the existence of solutions to this system of equations $(\Sigma_i)_{i=1:N-1}$ depends on the chosen distributions $(\mathcal{D}_i)_{i=1:N-1}$ and can be found numerically.

2.4.3 Example with the triangular distribution

In order to illustrate the approaches described above, we specify an equal expected value for the $(\pi_i)_{i=1:N}$ following §2.4.1. It corresponds to the strategy applied for estimating the grazing zooplankton preferences in the numerical experiments shown in §3 and §4. First, we must choose a distribution for the parameters $(\phi_i)_{i=1:N-1}$ and we assume that they follow a triangular distribution:

$$\forall i = 1:N-1, \phi_i \sim \mathcal{T}(0,1,\theta_i). \quad (8)$$

with $\theta_i \in [0,1]$ the mode of the distribution. The probability density function reads

$$\forall i = 1:N-1, f_{\phi_i}(\phi) = \begin{cases} \dfrac{2\phi}{\theta_i}, & \text{for} \quad 0 \leqslant \phi \leqslant \theta_i \\ \dfrac{2(1-\phi)}{1-\theta_i}, & \text{for} \quad \theta_i \leqslant \phi \leqslant 1 \end{cases} \quad (9)$$

The characteristic function Φ_{ϕ_i} is given by

$$\forall i = 1:N-1, \quad \forall t \in \mathbb{R}, \quad \Phi_{\phi_i}(t) = -2\frac{(1-\theta_i)-e^{j\theta_i t}+\theta_i e^{jt}}{\pi^2\theta_i(1-\theta_i)}, \quad (10)$$

with $j^2=-1$.

Then, a prior equal value for the parameters $(\pi_i)_{i=1:N}$ is obtained by an adequate tuning of the $N-1$ modes $(\theta_i)_{i=1:N-1}$. For that particular case, one has

$$\forall i = 1:N, \quad m_i = \frac{1}{N}. \quad (11)$$

And Eq. (6) reads

$$\forall i = 1:N-1, \quad (\mathcal{S}_i) \quad \frac{\cos(\pi\theta_i)+2\theta_i-1}{\pi^2\theta_i(1-\theta_i)}+\frac{N-i-1}{2(N-i+1)}=0. \quad (12)$$

The parameters $(\phi_i)_{i=1:N-1}$ distributed according to $(\mathcal{T}(0,1,\theta_i))_{i=1:N-1}$, with the $(\theta_i)_{i=1:N-1}$ solutions of the system defined by Eq. (12), lead to equal expected values for the $(\pi_i)_{i=1:N}$.

However, it can be shown that solutions exist if and only if

$$N-i < \frac{\pi^2+4}{\pi^2-4} \sim 2.36 \quad (13)$$

It means that only the equations $(\mathcal{S}_{N-1})$ and $(\mathcal{S}_{N-2})$ admit a solution. In practice, it will not be possible to obtain equal prior values of the $(\pi_i)_{i=1:N}$ for $N \geqslant 4$ when using triangular distributed parameters $(\phi_i)_{i=1:N-1}$.

Finally, it is worthy to note that it is not possible to choose the variances of the $(\pi_i)_{i=1:N}$ after having specified their expected values, because the triangular distribution has only one parameter: its mode.

2.5 Truncated Gaussian distribution and linear inequality constraints

Another strategy to reduce the number of parameters to estimate consists in formulating the problem simply using linear inequality constraints. Without loss of generality, we choose to only estimate the first $N-1$ parameters $(\pi_i)_{i=1:N-1}$ and the last parameter is given by the sum-to-one constraint: $\pi_N = 1 - \sum_{i=1}^{N-1} \pi_i$. The problem amounts to estimating the $(\pi_i)_{i=1:N-1}$ under the N inequality constraints $\forall i = 1:N, \pi_i \geqslant 0$.

This can be done with the truncated Gaussian filter suggested by Lauvernet et al. (2009) under the assumption that the $(\pi_i)_{i=1:N-1}$ have a truncated Gaussian distribution. However, the application of this filter is not as easy in practice than the simple variable transformations suggested in the previous sections and can be computationally expensive for large systems due to the use of a Gibbs sampler to sample the truncated-Gaussian distribution and to estimate its location vector and scale matrix.

3 Experimental framework

3.1 The 1-D ocean ecosystem model

The experiments were performed in a 1-D vertical configuration of the coupled model GOTM-NORWECOM representative of the station Mike (66° N, 2° W) in the North Sea.

The 1-D ocean water column model is the General Ocean Turbulence Model (GOTM; Burchard et al., 1999, 2005; Umlauf and Burchard, 2005) that transports physical quantities with hydrodynamic primitive equations and turbulence schemes. A relaxation towards temperature, salinity and horizontal velocity profiles from the TOPAZ[1] system (Bertino and Lisæter, 2008) is used with a relaxation time of 14 days. The vertical advection velocity is specified to zero. The depth is 2034 m, and the model uses a Cartesian grid of 55 vertical levels with a minimum thickness of 1 m at the top level, increasing exponentially towards the bottom.

The NORWegian ECOlogical Model system (NORWECOM; Aksnes et al., 1995; Skogen and Søiland, 1998) is coupled to GOTM. The current version of this model includes two classes of phytoplankton (diatom and flagellates), two classes of zooplankton (meso- and microzooplankton) derived with the same formulation from the model ECOHAM4 (Pätsch et al., 2009), three types of nutrients (inorganic nitrogen, phosphorus and silicon) and detritus (nitrogen, phosphorus), biogenic silica, and oxygen, so that the ecosystem state vector is made of 11 variables. The chlorophyll a concentration (CHLA) is computed from the model diatoms and flagellates concentrations (DIA and FLA) by Eq. (14):

$$\mathrm{CHLA} = \frac{\mathrm{DIA} + \mathrm{FLA}}{0.8} \tag{14}$$

The constant conversion factor $0.8\,\mathrm{mmol\,N\,mg^{-1}\,Chl}\ a$ is added to obtain the chlorophyll concentration in $\mathrm{mg\,m^{-3}}$, the standard unit of data produced from satellite, from the phytoplankton concentration in $\mathrm{mmol\,N\,m^{-3}}$. The mesozooplankton (MES) feed on diatoms (one assumes that the flagellates are too small to be fed on by mesozooplankton), detritus (DEN) and microzooplankton (MIC). The microzooplankton feed on both classes of phytoplankton (flagellates and diatoms) and on detritus. Both classes of zooplankton have the choice of their food among three variables of the model and compete against each other for feeding on detritus and diatoms. For both classes of zooplankton, the formulation of the grazing $G_{i=1:3}$ on the variable $i = 1:3$ reads

$$\forall i = 1:3, G_i = g \frac{\pi_i X_i^2}{\sum_{k=1}^{N} \pi_k X_k (X_k + K_{1/2})} Z \tag{15}$$

with Z the concentration of meso- or microzooplankton feeder, $(X_k)_{k=1:3}$ the concentration of the different variables they feed on, $(\pi_k)_{k=1:3}$ the grazing preferences, $K_{1/2}$ the half-saturation constant for ingestion by zooplankton and g the zooplankton maximum growth rate. The second order modified Patankar-Runge-Kutta scheme is used for the source and sinks dynamics.

The dynamics of phytoplankton blooms in the first 100 m in the reference solution is illustrated in Fig. 1.

3.2 Data assimilation experiments

In order to assess the performances of the two formulations, twin experiments have been conducted: the true state and the observations are produced by a deterministic simulation of the model involving meso- and microzooplankton grazing preferences that differ from equal preferences. The values of preferences used to build the reference solution will be called "true" values in the following. These values have been arbitrary chosen and are summarized in Table 1.

The observations are the chlorophyll in the two first layers of the model and are defined as follows:

$$\mathbf{y}_n = \mathbf{H}_n \mathbf{x}_n^t \times E, \quad \text{with} \quad E \sim \Gamma(\frac{1}{\sigma_o^2}, \sigma_o^2), \quad \sigma_o = 0.3 \tag{16}$$

We construct the observations by multiplying the true surface chlorophyll with a gamma distributed observation error with a standard deviation around 30 % (average should be 1).

The observation are assimilated with a Gaussian anamorphosis extension of the deterministic ensemble Kalman filter (DEnKF). This method is based on the DEnKF (Sakov and Oke, 2008) and consists in introducing Gaussian anamorphosis functions in order to realize the analysis step with Gaussian distributed transformed variables. More details can be

[1] http://topaz.nersc.no

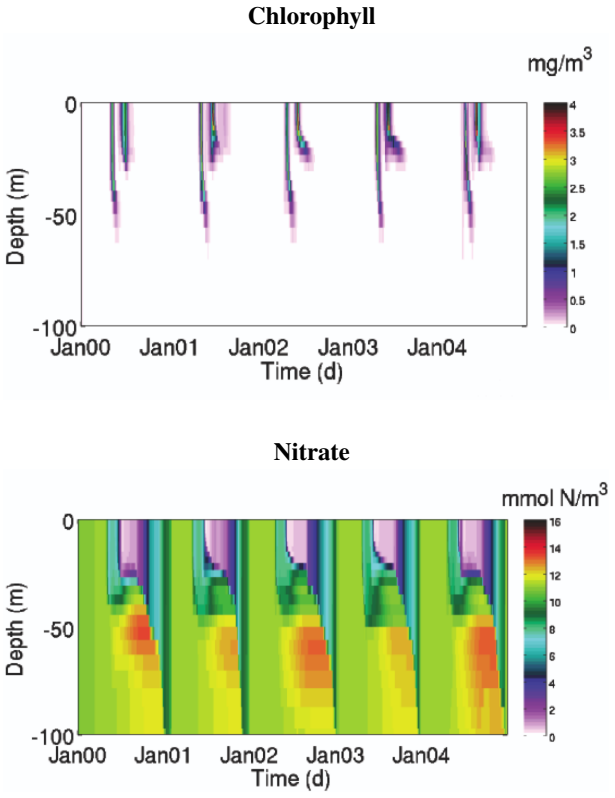

Fig. 1. Reference solution: time evolution of chlorophyll and nitrate in the upper 100m from 1 January 2000 to 31 December 2004.

found in Simon and Bertino (2012). The state and parameter estimations are conducted jointly by augmenting the state vector with the parameters that are estimated. In this study, the state vector is made up of all the vertical components of the ten state variables (the oxygen is not corrected during the analysis) and the parameters $(\phi_i)_{i=1:n}$, n depending on the formulation that is chosen. In the Gelman formulation, we take six parameters (three parameters controlling the preferences times two zooplankton types). In the spherical formulation, we take four parameters (two parameters controlling the preferences times two zooplankton types).

The ensemble contains 100 members. The background state ensemble is generated by adding a truncated-Gaussian perturbation to the solution $\mathbf{x}(t=0)$:

$$\forall i = 1:100, \quad \mathbf{x}_b^i = \max(0, \mathbf{x}(t=0) \times (1+b_i)) \tag{17}$$

with $b \sim \mathcal{N}(0, \sigma_b^2)$. σ_b is chosen to be equal to 0.3 for all the state variables. In the Gelman formulation, the parameter ensemble is initialized by assuming that the parameters $(\phi_i)_{i=1:3}$ are normally distributed according to $\mathcal{N}(0, \sigma = 2)$. In the spherical formulation, we assume that the parameters $(\phi_i)_{i=1:2}$ follow a triangular distribution:

$$\forall i = 1:2, \phi_i \sim \mathcal{T}(0, 1, \theta_i). \tag{18}$$

with $\theta_i \in [0,1]$ the mode of the distribution. The triangular distribution is simulated from the uniform distribution thanks to the MINMAX method suggested by Stein and Keblis (2009). The prior values for the parameters $(\pi_i)_{i=1:3}$ are obtained by an adequate tuning of the 2 modes $(\theta_i)_{i=1:2}$. Equal preferences are obtained by solving the two nonlinear equations:

$$\begin{aligned} (\mathcal{S}_1) \quad & \frac{\cos(\pi\theta_1) + 2\theta_1 - 1}{\pi^2\theta_1(1-\theta_1)} + \frac{1}{6} = 0. \\ (\mathcal{S}_2) \quad & \frac{\cos(\pi\theta_2) + 2\theta_2 - 1}{\pi^2\theta_2(1-\theta_2)} = 0. \end{aligned} \tag{19}$$

A solution to the equations $(\mathcal{S}_i)_{i=1:2}$ exists and can be found numerically: $\theta_1 = 0.8905$ and $\theta_2 = 0.5$. The mapping of the preferences in Eq. (5) is as follows:

$$\begin{array}{cc} \text{Mesozooplankton} & \text{Microzooplankton} \\ \left\{ \begin{array}{l} \pi_{DIA} = \cos^2(\frac{\pi}{2}\phi_1) \\ \pi_{MIC} = \sin^2(\frac{\pi}{2}\phi_1)\cos^2(\frac{\pi}{2}\phi_2) \\ \pi_{DET} = \sin^2(\frac{\pi}{2}\phi_1)\sin^2(\frac{\pi}{2}\phi_2) \end{array} \right. & \left\{ \begin{array}{l} \pi_{DET} = \cos^2(\frac{\pi}{2}\phi_1) \\ \pi_{\mathrm{FLA}} = \sin^2(\frac{\pi}{2}\phi_1)\cos^2(\frac{\pi}{2}\phi_2) \\ \pi_{DIA} = \sin^2(\frac{\pi}{2}\phi_1)\sin^2(\frac{\pi}{2}\phi_2) \end{array} \right. \end{array} \tag{20}$$

This choice is motivated by our wish to respect the symmetry between the two classes of phytoplankton in the definition of the observed variable (same weight for the FLA and DIA variables when computing the chlorophyll concentration). Since mesozooplankton only eat one type of phytoplankton (diatoms), π_1 is associated with π_{DIA}. The microzooplankton feed on the two classes of phytoplankton respectively; we associate π_2 and π_3 with π_{DIA} and π_{FLA}, respectively. This asymmetry, which appears only in the spherical formulation, may create a dependence of the results on the parameter assignment. However, experiments with different assignments between the microzooplankton grazing preferences and the $(\pi_i)_{i=1:3}$ led to similar results (not shown). We noted a slight decrease (increase) in the RMS errors in the estimates of the microzooplankton (mesozooplankton) grazing preferences compared to the results shown in the following. Nevertheless, the observed robustness of the estimation to the asymmetry of the transformation could be application-dependent and further experiments might be required in a different framework.

Gaussian anamorphosis functions are applied to state variables and parameters except for parameters transformed by the Gelman formulation. In the latter case, the $(\phi_i)_{i=1:3}$ are already normal-distributed and Gaussian anamorphosis is not necessary (see Sect. 2). The strategy to build the anamorphosis functions differs between the chlorophyll and the other state variables and parameters (if necessary) and is a variation of the hybrid approach described in Simon and Bertino (2012). Since the chlorophyll concentration in the ocean is usually assumed to have a log-normal distribution

(Campbell, 1995), its anamorphosis function is the logarithmic function. For other state variables and the parameters, anamorphosis functions are built from the empirical marginal distributions of the variables. The empirical anamorphosis functions are computed from a sample of the forecast ensemble and are then piecewise linearly interpolated to obtain the Gaussian anamorphosis functions. Their tails are linear and their last segments extrapolated towards specified biological minimum and maximum values. The spherical formulation introduces parameters that are bounded on both sides and for which the odds to reach the bounds during the assimilation are not null. The succession of analysis steps can build-up discontinuities ("atoms") of the distribution at the bounds which are not handled by the piecewise linear anamorphosis function – zero slopes are not invertible(Simon and Bertino, 2012). Extending the first and last segments until they include the first values outside of the atoms seems to resolve the issue. The observation error ϵ^o is assumed to have a log-normal distribution: $\log(\epsilon^o) \sim \mathcal{N}(0, \sigma_o^2)$ with $\sigma_o = 0.3$. It results in a normal-distributed observation error for the transformed observations with a standard deviation equal to 0.3.

The model includes perturbations on the phyto- and zooplankton components of the state variables. Similarly to the generation of the background state variables, truncated-Gaussian random variables are added every twelve hours. The standard deviation of these perturbations decreases linearly towards zero in the eight deepest layers in order to obtain a smooth transition between the deep layers and the bottom layer where no perturbations are applied. Furthermore, no perturbations are added to the parameters during the model integration, so they remain constant between two analysis steps.

Starting from the background state variables and parameters, a one-year ensemble simulation is performed without assimilation. Assimilation cycles are then performed over four years with a frequency of one analysis step every seven days. This frequency for observing the system is relatively low considering the short time scales of the bloom phenomenon. Figure 2 represents the time evolution of the chlorophyll concentration in the two first top layers in the reference solution and in the assimilated observations. We note that during the blooms the 7-day sampling of the reference run leads to only one observation during the first peak (diatom bloom) and only one or two observations during the second peak (flagellate bloom). Furthermore, the maximum values reached by the concentrations in the reference solution during these two peaks are generally not captured by the observations. Blooms are mostly represented in the observations as two Dirac pulses with highly uncertain amplitude and timing. This usually results in difficulties for the ensemble-based Kalman filter methods to correctly estimate the state of the system, and notably to estimate some parameters. This is a real issue for ocean ecosystem models: the weak production (apart from the bloom periods) results in low innovations and spread of the chlorophyll concentration in the ensemble, and weak corrections by the filters during most of the year. An increase of the sampling frequency to four days would be enough to obtain a good representation of blooms in this simple 1-D configuration, more specifically the transition phases, and potentially improve the quality of the estimation. Nevertheless, assimilating observations more frequently might not be affordable in realistic 3-D configurations due to the computational costs that it implies. The use of an asynchronous version of the EnKF (Sakov et al., 2010) would be a solution to tackle these issues but is out of the scope of this study.

In order to check the robustness of the estimation against random initial conditions and observation errors, we repeated the experiment 20 times. That is, 20 initial ensembles (combined state-parameter background) and 20 sets of observations were generated. Nevertheless, the different assimilation systems used the same state component of the background ensemble and observations for each of the 20 realizations. The diagnostics shown in Sect. 4 are averaged over these 20 experiments.

4 Data assimilation results

4.1 Overall error evolution

We are interested in the time evolution of the relative root mean square error (RMS) and the relative ensemble standard deviations (STD) of the solution of the two different formulations. These diagnostics are averaged over 20 experiments. The expression at time t_n of these two quantities is as follows:

$$\mathrm{RMS}(t_n) = \frac{\frac{1}{N_{\mathrm{exp}}} \sum_{i=1}^{N_{\mathrm{exp}}} \sqrt{\sum_{\mathbf{k}\in\Omega} (\mathbf{x}^t(t_n,\mathbf{k}) - \bar{\mathbf{x}}(t_n,\mathbf{k},i))^2}}{\sqrt{\sum_{\mathbf{k}\in\Omega} \mathbf{x}^t(t_n,\mathbf{k})^2}} \tag{21}$$

$$\mathrm{STD}(t_n) = \frac{\frac{1}{N_{\mathrm{exp}}} \sum_{i=1}^{N_{\mathrm{exp}}} \sqrt{\frac{1}{N-1} \sum_{\mathbf{k}\in\Omega} \sum_{m=1}^{N} (\mathbf{x}^m(t_n,\mathbf{k},i) - \bar{\mathbf{x}}(t_n,\mathbf{k},i))^2}}{\sqrt{\sum_{\mathbf{k}\in\Omega} \mathbf{x}^t(t_n,\mathbf{k})^2}} \tag{22}$$

where Ω is the domain of computation, N is the number of members, $\mathbf{x}^m$ is the forecast member m, N_{exp} is the number of experiments, $\mathbf{x}^t$ is the true state, and $\bar{\mathbf{x}}$ is the mean of the forecast ensemble.

Figure 3 represents the evolution of the relative RMS and standard deviation over five years for the diatoms, flagellates and the micro- and mesozooplankton. These diagnostics are averaged over the whole water column, and Ω represents the 55 vertical layers. The evolution of the spatial average of the true state is plotted (green dashed line) in order to provide

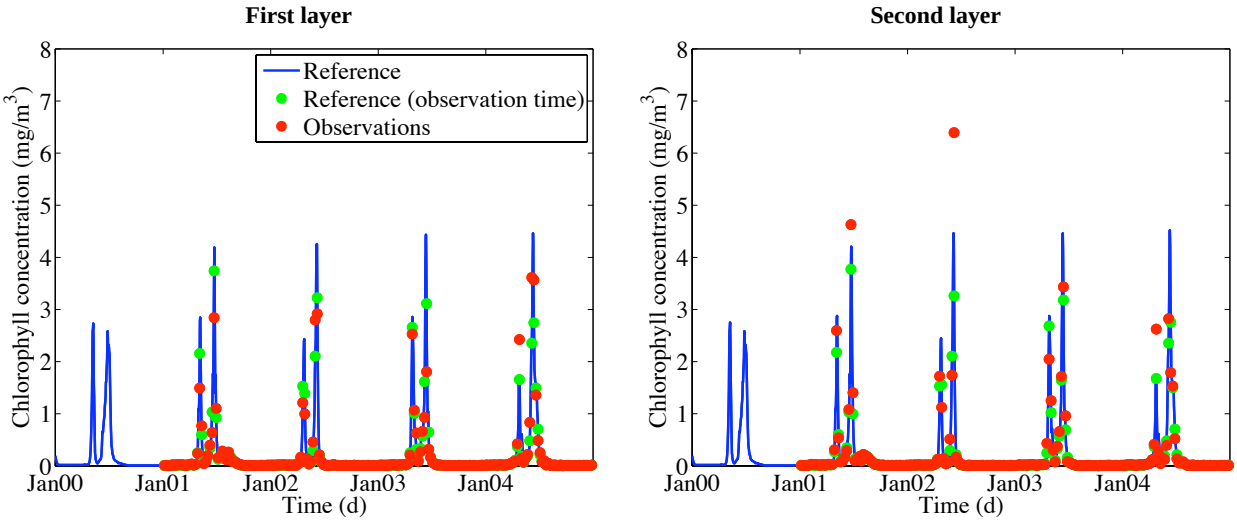

Fig. 2. Time evolution of the chlorophyll concentration at the surface (two first layers) in the reference solution (blue line), in the reference solution at observation times (green circles) and in the observations (red circles) for one experiment.

information on the yearly dynamics of the variables. No assimilation is performed during the first year. First, we note that both formulations lead to a reduction of the RMS error and the standard deviation for flagellates and their grazers, the microzooplankton. The peaks in the error for flagellates occur at the end of the flagellates blooms, which are too short in the assimilated solution, notably around 25 m depth. The evolutions of the standard deviation and RMS error are in agreement during the last bloom both for microzooplankton and flagellates, which highlights a good representation of the error by the ensemble during that period. An improvement of the detritus component of the solutions is also observed (not shown).

The impact of data assimilation on the diatoms is mixed. We note a large increase of the standard deviation after all the diatoms blooms and a large peak in the RMS error during the first year with assimilation associated with a too long bloom. The RMS error decreases year after year for both formulations and reaches its lowest values during the fourth year. However, a large peak is still present in the error during the final bloom for the spherical formulation. This is due to the presence of a strong subsurface chlorophyll maximum at a 70 m depth. Because the silica cycle depends only on the diatom concentration, these large peaks in the error result in a low increase of the RMS error for both silicate components during the bloom every year leading to final error around 10 % (not shown). In the same way, data assimilation cannot significantly reduce the RMS errors for the mesozooplankton. On average, the solutions obtained with the Gelman formulation present a lower error than the ones obtained with the spherical formulation. Finally, nitrate and phosphate are not significantly impacted during the assimilation (not shown). On average, their RMS errors are low (less than 5 %) and exhibit low oscillations during the blooms.

4.2 Evolution of the parameters

Figure 4 represents the time evolution of the mean and standard deviation of the ensemble for the meso- and microzooplankton grazing preferences. First, we note that both formulations lead on average to reasonably good final estimates of the microzooplankton grazing preferences. The largest corrections occurring during the first two blooms result in a convergence of the estimation towards the true values of the preferences in less than two years for both formulations. However, we note larger corrections during the last bloom with the Gelman formulation that can be explained by a larger spread for the preferences in the ensemble inherited from the initial ensembles. The Gelman formulation introduces a distribution with two parameters – the mean and the variance of the normal distribution (see Eq. 4) – which makes it possible to choose the mean and the standard deviation of the prior preferences. Our use of a distribution with one parameter – the mode of the triangular distribution (see Eq. 18) – allows only for the choice of the mean for the prior preferences. In these experiments, the prior variances chosen for the normal-distributed parameters in the Gelman formulation lead to initial variances for the prior preferences that are larger than the ones obtained with the spherical formulation.

The mean and standard deviation of the 20 means of the preferences in the ensemble obtained at the end of the experiments are specified in Table 1 and the RMS error in Table 2. On average, the Gelman formulation produces slightly better estimates of the preferences for diatoms and detritus, while

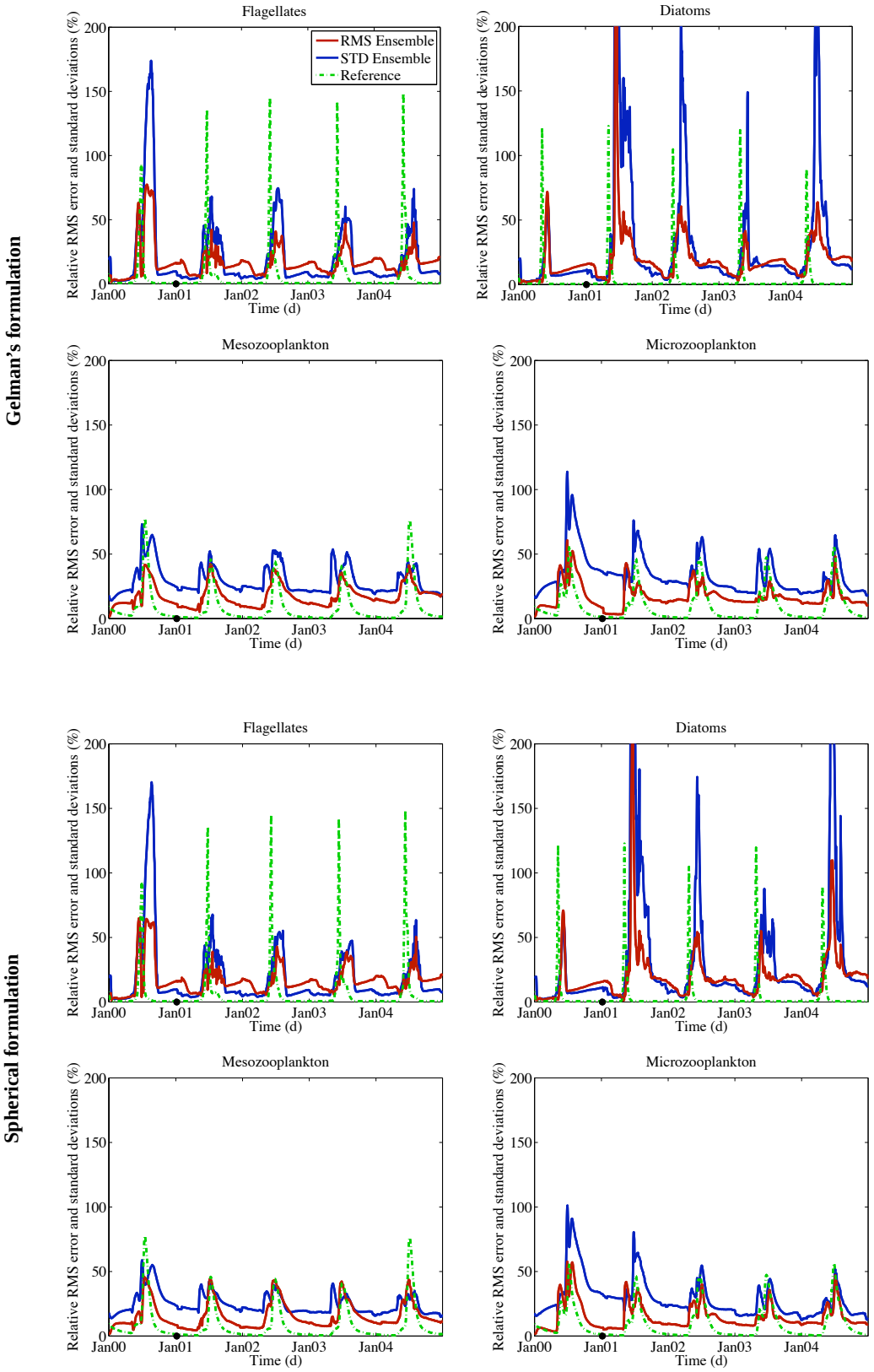

Fig. 3. Evolution with time of the relative RMS error and standard deviation computed over the water column and averaged over the 20 experiments. The spatial mean of the reference solution is plotted to highlight the seasonal dynamics (green dashed curve). The black dot highlights the date of the first analysis.

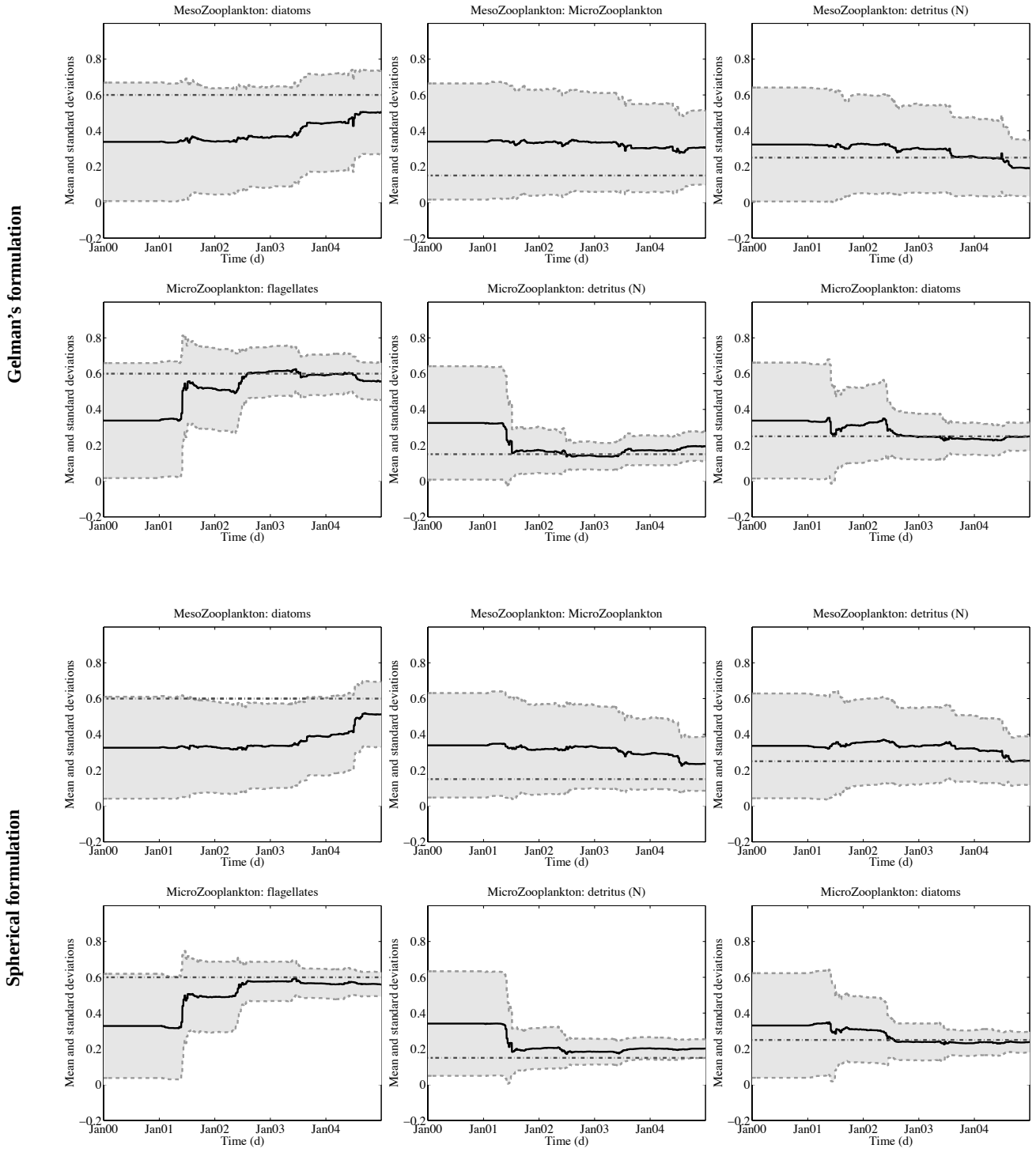

Fig. 4. Evolution with time of the averaged mean (black line) and averaged mean plus/minus the standard deviation (shaded area) of the grazing preferences. The true value is highlighted with a dark dashed-dotted line.

Table 1. Zooplankton grazing preferences $(\pi_i)_{i=1:3}$: mean and standard deviation (computed over the 20 experiments) of the means of preferences obtained at the final time.

Mesozooplankton			
Diet	DIA	MIC	DET
True value	0.6	0.15	0.25
Gelman	0.50 ± 0.19	0.31 ± 0.16	0.19 ± 0.09
Spherical	0.51 ± 0.19	0.24 ± 0.11	0.25 ± 0.13
Microzooplankton			
Diet	DET	FLA	DIA
True value	0.15	0.6	0.25
Gelman	0.19 ± 0.1	0.56 ± 0.09	0.25 ± 0.05
Spherical	0.20 ± 0.09	0.56 ± 0.10	0.24 ± 0.05

Table 2. Zooplankton grazing preferences $(\pi_i)_{i=1:3}$: relative RMS error (computed over the 20 experiments) of the means of preferences obtained at the final time.

Mesozooplankton			
Diet	DIA	MIC	DET
Prior (%)	45	120	32
Gelman (%)	35	146.6	44
Spherical (%)	33.3	86.7	48
Microzooplankton			
Diet	DET	FLA	DIA
Prior (%)	120	45	32
Gelman (%)	66.7	16.7	20
Spherical (%)	66.7	16.7	20

both formulations lead to the same estimate of the preferences for flagellates. Both formulations lead to a similar decrease of the RMS error in the estimate of the three preferences. However, the ternary plots of the final estimates of the preferences for the 20 experiments in Fig. 5 show that the number of experiments, for which the assimilation provides corrections in the direction of the true value for the three preferences, is larger with the spherical formulation than with Gelman's: only two points do not belong to the shaded area representing the subspace of preferences defined by $0 \leqslant \pi_{\mathrm{DET}} \leqslant 1/3$, $1/3 \leqslant \pi_{\mathrm{FLA}} \leqslant 1$ and $0 \leqslant \pi_{\mathrm{DIA}} \leqslant 1/3$ (decrease of the preferences for the diatoms and detritus and increase of the preference for the flagellates) with the spherical formulation compared to four points with the Gelman formulation.

The estimation of the mesozooplankton grazing preferences is less successful. On average, we note in Fig. 4 that the corrections are very weak during the first two years of assimilation. The reduction of the standard deviation of the three preferences is very low for both formulations suggesting a weaker sensitivity of the surface chlorophyll to the mesozooplankton grazing preferences compared to the microzooplankton grazing preferences. This is highlighted in Fig. 6 by the Pearson correlation coefficients between the surface chlorophyll and the microzooplankton that are much larger than the ones between the surface chlorophyll and the mesozooplankton.

On average, the spherical formulation leads to slightly better final estimates of the preferences than the Gelman formulation (see Table 1). The assimilation tends to strongly correct the preference for the detritus to the detriment of microzooplankton. It results in larger RMS errors in the final estimates of these two preferences compared to the prior values (see Table 2). The ternary plots in Fig. 5 show that in 45 % of the experiments the estimation with the Gelman formulation does not jointly improve the three preferences. For most of these experiments, this is due to an erroneous increase of the preference for the microzooplankton. The rate of failure decreases to 30 % of the experiments with the spherical formulation. For most cases, this is due to an erroneous increase of the preference for the detritus to the detriment of diatoms. This is also highlighted by the increase of the RMS error in the final estimate of the preference for the detritus (see Table 2). However, experiments done with different assignments of the microzooplankton grazing preferences in the transformation led to higher RMS errors, notably in the preference for microzooplankton, and a rate of failure equal to 45 % (not shown). This suggests that the performances of both approaches do not significantly differ. Furthermore, we think that these difficulties faced by the DEnKF to correctly estimate the mesozooplankton grazing preferences are related to the configuration of the experiments rather than the variable transformations. As stated earlier, the surface chlorophyll seems to be more sensitive to the microzooplankton than to the mesozooplankton in the model. Furthermore, improvements could be obtained by changing the experimental framework, for example the observation frequency, the specified observation error, etc.

5 Conclusions

In this study, we investigated the problem of estimating N positive sum-to-one constraint parameters with ensemble-based Kalman filters with the purpose of estimating zooplankton grazing preferences that are commonly used in ocean ecosystem models.

We have suggested a new formulation of the grazing preferences introducing a change of variables based on the hyperspherical coordinate system. This formulation results in the estimation of a reduced number $(N-1)$ of bounded parameters. Issues raised by estimating non-Gaussian distributed parameters with Kalman filters can be tackled by using the Gaussian anamorphosis. Furthermore, the two systems of

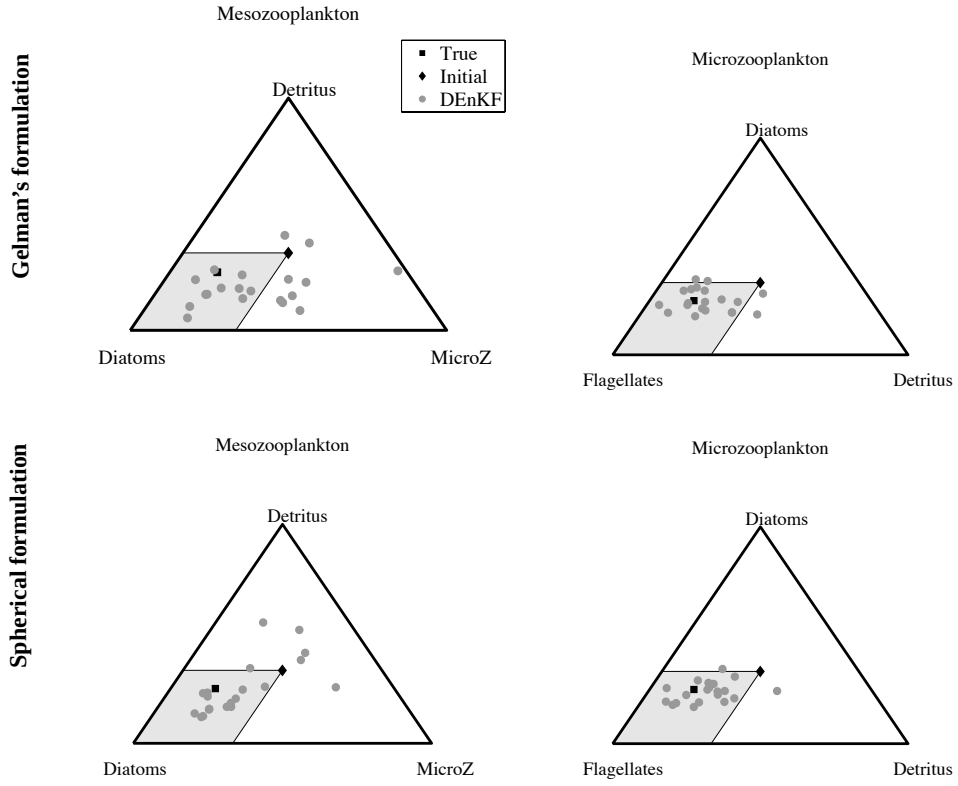

Fig. 5. Ternary plots of the final estimate (mean of the ensemble) of the grazing preferences parameters for the 20 experiments. The estimates obtained after assimilation are plotted with grey circles, the true set of parameters with a black square and the mean of the background set of parameters with a black diamond.

$N-1$ nonlinear equations to be solved, in order to obtain target prior values and variances of the preferences, are also exhibited.

The performances of this approach and the one suggested by Gelman (1995) based on the Dirichlet distribution have been assessed in the framework of twin experiments realized in a 1-D configuration of the coupled model GOTM-NORWECOM. Both approaches lead to improved estimates of the microzooplankton grazing preferences. They present the same difficulties to estimate the mesozooplankton grazing preferences that can be explained by the configuration of the experiments: the observed variable, the chlorophyll, constitutes only one type of food (diatoms) for the mesozooplankton diet compared to two (diatoms and flagellates) for the microzooplankton diet. Furthermore, the results obtained with the spherical formulation for the mesozooplankton are not significantly better and cannot be guaranteed for more complex realistic configurations.

Both approaches present theoretical and practical advantages. The Gelman formulation leads to the estimation of Gaussian distributed parameters, a property that presents theoretical advantages in the context of Kalman filtering. Furthermore, this formulation is naturally symmetric with regards to the mapping of parameters. This formulation is straightforward to apply for any number of preferences. However, it can require to estimate a large number of parameters in complex systems. The spherical formulation reduces the number of parameters to estimate but can require a choice of their prior distribution and to solve nonlinear systems of equations accordingly if the inversion of the hyperspherical coordinate system cannot be applied.

In this study, we have used the triangular distribution for its simplicity and its applicability in our ecosystem model. But this distribution is not suitable to obtain equal prior values for more than three preferences and does not allow the tuning of their variances. From five preferences onwards – N equal four can be solved via the introduction of the Hopf coordinate system – the questions of the choice of the distribution and the resolution of the systems remain open. However, the inversion of the hyperspherical coordinate system could provide a prior ensemble for the $(\phi_i)_{i=1:N-1}$ if an ensemble for the preferences is available. This suggests that the Gelman formulation is more suitable in the framework of few zooplankton species with a diet involving numerous types of food, while the spherical formulation could be more suitable in the framework of numerous zooplankton species with a diet involving few types of food.

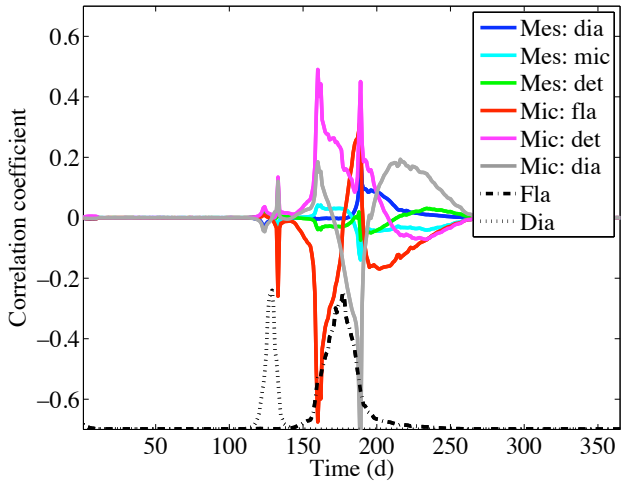

Fig. 6. Time evolution of the Pearson correlation coefficients between the surface chlorophyll and the meso- and microzooplankton grazing preferences during the first year (no assimilation) and averaged over the 20 experiments, for the spherical formulation. The evolution of the averaged surface chlorophyll concentration due to diatoms (resp. flagellates) is plotted with a dark dotted line (resp. with a dark dash-dotted line).

Appendix A

Derivation of the system of equations $(\mathcal{S}_i)_{i=1:N-1}$ to choose the mean of the preferences $(\pi_i)_{i=1:N}$

Let $(\phi_i)_{i=1:N-1}$ be $N-1$ independent random variables following marginal distributions $(\mathcal{D}_i)_{i=1:N-1}$ involving a set of parameters $(\Theta_i)_{i=1:N-1}$ and with a support equal to the segment line $[0,1]$. We note f_{ϕ_i} the probability density function of the parameter ϕ_i for all $i = 1:N-1$:

$$\forall i = 1:N-1, \quad \begin{array}{rcl} f_{\phi_i} : [0,1] & \to & \mathbb{R}_+ \\ \phi & \mapsto & f_{\phi_i}(\phi) \end{array} \tag{A1}$$

Let $(\pi_i)_{i=1:N}$ N random variables defined by Eq. (5). We aim to choose the values of the set of parameters $(\Theta_i)_{i=1:N-1}$ to obtain the expected values $(m_i)_{i=1:N}$, with $\sum_{k=1}^{N} m_k = 1$, for the variables $(\pi_i)_{i=1:N}$:

$$\forall i = 1:N, \quad E[\pi_i] = m_i \tag{A2}$$

We start with a preliminary calculus. By introducing the relation $\cos^2(a) = \frac{1+\cos(2a)}{2}$ and using the property of a probability density function f: $\int_{\mathbb{R}} f(\phi)d\phi = 1$, one has

$$\int_0^1 \cos^2(\frac{\pi}{2}\phi) f_{\phi_i}(\phi)d\phi = \frac{1}{2} + \frac{1}{2}\int_0^1 \cos(\pi\phi) f_{\phi_i}(\phi)d\phi \tag{A3}$$

By introducing $\cos(a) = \frac{e^{ja}+e^{-ja}}{2}$ with $j^2 = -1$, it leads to

$$\begin{aligned} \int_0^1 \cos^2(\frac{\pi}{2}\phi) f_{\phi_i}(\phi)d\phi &= \frac{1}{2} + \frac{1}{4}\int_0^1 e^{j\pi\phi} f_{\phi_i}(\phi)d\phi \\ &\quad + \frac{1}{4}\int_0^1 e^{-j\pi\phi} f_{\phi_i}(\phi)d\phi \\ &= \frac{1}{2} + \frac{1}{4}(\Phi_{\phi_i}(\pi) + \Phi_{\phi_i}(-\pi)) \end{aligned} \tag{A4}$$

with Φ_{ϕ_i} the characteristic function of the parameter ϕ_i. By defining

$$\forall i = 1:N-1, \quad h(\Theta_\mathbf{i}) = \frac{1}{4}(\Phi_{\phi_i}(\pi) + \Phi_{\phi_i}(-\pi)) \tag{A5}$$

The last equation reads

$$\int_0^1 \cos^2(\frac{\pi}{2}\phi) f_{\phi_i}(\phi)d\phi = \frac{1}{2} + h(\Theta_\mathbf{i})\,. \tag{A6}$$

In the same way, one has

$$\int_0^1 \sin^2(\frac{\pi}{2}\phi) f_{\phi_i}(\phi)d\phi = \frac{1}{2} - h(\Theta_\mathbf{i})\,. \tag{A7}$$

Now, for $i = 1$, one has

$$E[\pi_1] = \int_0^1 \cos^2(\frac{\pi}{2}\phi) f_{\phi_1}(\phi)d\phi = \frac{1}{2} + h(\Theta_\mathbf{1}), \tag{A8}$$

and the Eq. (A2) reads

$$h(\Theta_\mathbf{1}) = m_1 - \frac{1}{2}, \tag{A9}$$

which is equivalent to the equation $(\mathcal{S}_1)$ defined by Eq. (6).

Now, let i be an integer between 2 and $N-1$. By definition of the variables $(\pi_i)_{i=1:N}$, one has

$$\forall i = 2:N-1, \quad E[\pi_i] = \int_{[0,1]^{N-1}} \prod_{k=1}^{i-1} \sin^2(\frac{\pi}{2}\phi_k)\cos^2(\frac{\pi}{2}\phi_i) f_{(\phi_l)_{l=1:N-1}}(\phi)\mathbf{d}\phi \tag{A10}$$

Because the variables $(\phi_l)_{l=1:N-1}$ are independent, it leads to

$$\forall i = 2:N-1,\quad E[\pi_i] = \prod_{k=1}^{i-1}\int_0^1 \sin^2(\frac{\pi}{2}\phi_k) f_{\phi_k}(\phi)d\phi \int_0^1 \cos^2(\frac{\pi}{2}\phi_i) f_{\phi_i}(\phi)d\phi \tag{A11}$$

By introducing Eq. (A6) and Eq. (A7), one obtains

$$\forall i = 2:N-1,\quad E[\pi_i] = \prod_{k=1}^{i-1}(\frac{1}{2} - h(\Theta_k))(h(\Theta_i)+\frac{1}{2}) \tag{A12}$$

It leads to $\forall i = 2:N-1$:

$$\begin{aligned}
&\frac{E[\pi_i]}{E[\pi_{i-1}]} = \frac{m_i}{m_{i-1}} \\
\Leftrightarrow\ &\frac{\prod_{k=1}^{i-1}(\frac{1}{2} - h(\Theta_k))(h(\Theta_i)+\frac{1}{2})}{\prod_{k=1}^{i-2}(\frac{1}{2} - h(\Theta_k))(h(\Theta_{i-1})+\frac{1}{2})} = \frac{m_i}{m_{i-1}} \\
\Leftrightarrow\ &\frac{(\frac{1}{2} - h(\Theta_{i-1}))(h(\Theta_i)+\frac{1}{2})}{(h(\Theta_{i-1})+\frac{1}{2})} = \frac{m_i}{m_{i-1}} \\
\Leftrightarrow\ &h(\Theta_i) = -\frac{1}{2} + \frac{m_i}{m_{i-1}}\frac{1+2h(\Theta_{i-1})}{1-2h(\Theta_{i-1})}
\end{aligned} \tag{A13}$$

Finally, we obtain a recurrence between the variables $(h(\Theta_i))_{i=1:N-1}$:

$$\begin{cases} h(\Theta_1) = m_1 - \frac{1}{2} \\ \forall i = 2:N-1, \\ \qquad h(\Theta_i) = -\frac{1}{2} + \frac{m_i}{m_{i-1}}\frac{1+2h(\Theta_{i-1})}{1-2h(\Theta_{i-1})} \end{cases} \tag{A14}$$

The solution of Eq. (A14) is given by

$$\begin{cases} h(\Theta_1) = m_1 - \frac{1}{2} \\ \forall i = 2:N-1, \\ \qquad h(\Theta_i) = \frac{m_i}{1-\sum_{k=1}^{i-1} m_k} - \frac{1}{2} \end{cases}, \tag{A15}$$

which corresponds to the system of equations $(\mathcal{S}_i)_{i=1:N-1}$.

We must now check that the relation $E[\pi_N] = m_N = 1 - \sum_{k=1}^{N-1} m_k$ is satisfied.

$$\begin{aligned}
\frac{E[\pi_N]}{E[\pi_{N-1}]} &= \frac{\int_{\mathbb{R}} \sin^2(\frac{\pi}{2}\phi_{N-1}) f_{\phi_{N-1}}(\phi)d\phi}{\int_{\mathbb{R}} \cos^2(\frac{\pi}{2}\phi_{N-1}) f_{\phi_{N-1}}(\phi)d\phi} \\
&= \frac{1-2h(\Theta_{N-1})}{1+2h(\Theta_{N-1})} \\
&= \frac{1-\sum_{k=1}^{N-2} m_k - m_{N-1}}{m_{N-1}}
\end{aligned} \tag{A16}$$

Because of $E[\pi_{N-1}] = m_{N-1}$, one does have $E[\pi_N] = m_N = 1 - \sum_{k=1}^{N-1} m_k$.

Appendix B

Derivation of the system of equations $(\Sigma_i)_{i=1:N-1}$ to choose the variance of the preferences $(\pi_i)_{i=1:N}$

Let $(\pi_i)_{i=1:N}$ N random variables defined by Eq. (5). We aim to choose the values of the set of parameters $(\Theta_i)_{i=1:N-1}$ to obtain the variances $(\sigma_i^2)_{i=1:N}$ of the variables $(\pi_i)_{i=1:N}$ assuming that their expected values are equal to $(m_i)_{i=1:N}$:

$$\forall i = 1:N,\quad E[\pi_i^2] - m_i^2 = \sigma_i^2 \tag{B1}$$

We start with preliminary calculus. As previously, one obtains the following by using trigonometric formulas:

$$\begin{aligned}
\int_0^1 \cos^4(\frac{\pi}{2}\phi) f_{\phi_i}(\phi)d\phi &= \frac{3}{8} + \frac{1}{2}\int_0^1 \cos(\pi\phi) f_{\phi_i}(\phi)d\phi \\
&\quad + \frac{1}{8}\int_0^1 \cos(2\pi\phi) f_{\phi_i}(\phi)d\phi \\
&= \frac{3}{8} + h(\Theta_i) + g(\Theta_i)
\end{aligned} \tag{B2}$$

with h defined in Eq. (A5) and g as

$$\forall i = 1:N-1,\quad g(\Theta_{\mathbf{i}}) = \frac{1}{16}(\Phi_{\phi_i}(2\pi) + \Phi_{\phi_i}(-2\pi)). \tag{B3}$$

In the same way, one has

$$\int_0^1 \sin^4(\frac{\pi}{2}\phi) f_{\phi_i}(\phi)d\phi = \frac{3}{8} - h(\Theta_i) + g(\Theta_i) \tag{B4}$$

For $i = 1$ one has

$$\begin{aligned} E[\pi_1^2] &= \int_0^1 \cos^4(\frac{\pi}{2}\phi) f_{\phi_1}(\phi) d\phi \\ &= \frac{3}{8} + h(\Theta_1) + g(\Theta_1) \\ &= \sigma_1^2 + m_1^2 \end{aligned} \tag{B5}$$

It leads to

$$g(\Theta_{\mathbf{1}}) = -\frac{3}{8} - h(\Theta_1) + \sigma_1^2 + m_1^2 \tag{B6}$$

where $h(\Theta_1)$ is given by Eq. (A9). Now, let i be an integer between 2 and $N-1$. One has

$$E[\pi_i^2] = \int_{[0,1]^{N-1}} \prod_{k=1}^{i-1} \sin^4(\frac{\pi}{2}\phi_k) \cos^4(\frac{\pi}{2}\phi_i) f_{(\phi_l)_{l=1:N-1}}(\phi) \mathbf{d}\phi \tag{B7}$$

Following the same strategy as in Appendix A, it leads to

$$\frac{E[\pi_i^2]}{E[\pi_{i-1}^2]} = \frac{\sigma_i^2 + m_i^2}{\sigma_{i-1}^2 + m_{i-1}^2}$$

$$\Leftrightarrow \frac{\prod_{k=1}^{i-1}(\frac{3}{8} - h(\Theta_k) + g(\Theta_k))(\frac{3}{8} + h(\Theta_i) + g(\Theta_i))}{\prod_{k=1}^{i-2}(\frac{3}{8} - h(\Theta_k) + g(\Theta_k))(\frac{3}{8} + h(\Theta_{i-1}) + g(\Theta_{i-1}))} = \frac{\sigma_i^2 + m_i^2}{\sigma_{i-1}^2 + m_{i-1}^2} \tag{B8}$$

$$\Leftrightarrow \frac{(\frac{3}{8} - h(\Theta_{i-1}) + g(\Theta_{i-1}))(\frac{3}{8} + h(\Theta_i) + g(\Theta_i))}{(\frac{3}{8} + h(\Theta_{i-1}) + g(\Theta_{i-1}))} = \frac{\sigma_i^2 + m_i^2}{\sigma_{i-1}^2 + m_{i-1}^2}$$

$$\Leftrightarrow g(\Theta_i) = -\frac{3}{8} - h(\Theta_i) + \frac{\sigma_i^2 + m_i^2}{\sigma_{i-1}^2 + m_{i-1}^2} \frac{3 + 8h(\Theta_{i-1}) + 8g(\Theta_{i-1})}{3 - 8h(\Theta_{i-1}) + 8g(\Theta_{i-1})}$$

where $h(\Theta_{i-1})$ and $h(\Theta_i)$ are given by Eq. A15.

Finally, we obtain a recurrence between the variables $(g(\Theta_i))_{i=1:N-1}$:

$$\begin{cases} g(\Theta_{\mathbf{1}}) = -\frac{3}{8} - h(\Theta_1) + \sigma_1^2 + m_1^2 \\ \forall i = 2 : N-1, \\ g(\Theta_i) = -\frac{3}{8} - h(\Theta_i) + \frac{\sigma_i^2 + m_i^2}{\sigma_{i-1}^2 + m_{i-1}^2} \frac{3 + 8h(\Theta_{i-1}) + 8g(\Theta_{i-1})}{3 - 8h(\Theta_{i-1}) + 8g(\Theta_{i-1})} \end{cases} \tag{B9}$$

The solution of Eq. (B9) is given by the system of equations $(\Sigma_i)_{i=1:N-1}$:

$$\begin{cases} g(\Theta_1) = -\frac{3}{8} - h(\Theta_1) + \sigma_1^2 + m_1^2 \\ \\ \forall i = 2 : N-1, \\ g(\Theta_i) = -\frac{3}{8} - h(\Theta_i) \\ \qquad + \frac{\sigma_i^2 + m_i^2}{\sum_{k=1}^{i} (-2)^{k-1} (\sigma_{i-k}^2 + m_{i-k}^2) \prod_{l=1}^{k-1} h(\Theta_{i-l})} \end{cases} \tag{B10}$$

with the conventions $\sigma_0^2 + m_0^2 = 1$ and $\prod_{l=1}^{0} h(\Theta_l) = 1$.

The variance σ_N^2 of the preference π_N cannot be chosen and depends on the values specified for the preference π_{N-1}. It is given by

$$\begin{aligned} \sigma_N^2 = &-(1 - \sum_{i=1}^{N-1} m_i)^2 \\ &(\sigma_{N-1}^2 + m_{N-1}^2) \frac{3 - 8h(\Theta_{\mathbf{N-1}}) + 8g(\Theta_{\mathbf{N-1}})}{3 + 8h(\Theta_{\mathbf{N-1}}) + 8g(\Theta_{\mathbf{N-1}})} \end{aligned} \tag{B11}$$

Acknowledgements. The authors wish to thank the two anonymous referees for their helpful and constructive comments. This study has been funded by the eVITA-EnKF project from the Research Council of Norway and the MyOcean and GreenSeas IP from the European Commission's 7th FP. A grant of CPU time from the Norwegian Supercomputing Project (NOTUR2) has been used.

Edited by: P. Brasseur

References

Aksnes, D., Ulvestad, K., Baliño, B., Berntsen, J., and Svendsen, E. : Ecological modelling in coastal waters: towards predictive physical-chemical-biological simulation models, Ophelia, 41, 5–36, 1995.

Anderson, J. L.: An ensemble adjustment Kalman filter for data assimilation, Month. Weath. Rev., 129, 2884–2903, 2001.

Bertino, L., Evensen, G., and Wackernagel, H.: Sequential Data Assimilation Techniques in Oceanography, International Statistical Review, 71, 223–241, 2003.

Bertino, L. and Lisæter, K. A.: The TOPAZ monitoring and prediction system for the Atlantic and Arctic Oceans, J. Operat. Oceanogr., 1, 15–19, 2008.

Buitenhuis, E., Le Quéré, C., Aumont, O., Beaugrand, G., Bunker, A., Hirst, A., Ikeda, T., O'Brien, T., Piontkovski, S., and Straile, D.: Biogeochemical fluxes through mesozooplankton, Global Biogeochem. Cy., 20, GB2003, doi:10.1029/2005GB002511, 2006.

Buitenhuis, E. T., Rivkin, R. B., Sailley, S., and Le Quéré, C.: Biogeochemical fluxes through microzooplankton, Global Biogeochem. Cy., 24, GB4015, doi:10.1029/2009GB003601, 2010.

Burchard, H., Bolding, K., and Villareal, M. R.: GOTM, a general ocean turbulence model: theory, implementation and test cases, European Commission Technical Report, Brussels, 1999.

Burchard, H., Deleersnijder, E., and Meister, A.: Application of modified Patankar schemes to stiff biogeochemical models for the water column, Ocean Dynam., 55, 326–337, 2005.

Campbell, J. W.: The lognormal distribution as a model for bio-optical variability in the sea, J. Geophys. Res., 100, 13237–13254, 1995.

Doron, M., Brasseur, P., and Brankart, J.-M.: Stochastic estimation of biogeochemical parameters of a 3D ocean coupled physical-biogeochemical model: twin experiments, J. Mar. Syst., 87, 3–4, 2011.

Evans, G. T. and Parslow, J. S.: A model of annual plankton cycles, Biol. Oceanogr., 3, 327–347, 1985.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99, 10143–10182, 1994.

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

Evensen, G.: The Ensemble Kalman filter for combined state and parameter estimation, IEEE Control Systems Magazine, 29, 83–104, 2009.

Gelman A.: Method of Moments Using Monte Carlo Simulation, J. Comput. Graph. Stat., 4, 36–54, 1995.

Gelman, A., Bois, F., and Jiang, J.: Physiological Pharmacokinetic Analysis Using Population Modeling and Informative Prior Distributions, J. Am. Stat. Assoc., 91, 1400–1412, 1996.

Gentleman, W., Leising, A., Frost, B., Strom, S., and Murray J.: Functional responses for zooplankton feeding on multiple resources: a review of assumptions and biological dynamics, Deep Sea Res. Pt. II, 50, 2847–2875, 2003.

Gregg, W. W. and Casey, N. W.: Global and regional evaluation of the SeaWiFS chlorophyll data set, Remote Sens. Environ., 93, 463–479, 2004.

Lauvernet, L., Brankart, J.-M., Castruccio, F., Broquet, G., Brasseur, P., and Verron, J.: A truncated Gaussian filter for data assimilation with inequality constraints: Application to the hydrostatic stability condition in ocean models, Ocean Modell., 27, 1–17, 2009.

Nurmela, K. J.: Constructing spherical codes by global optimization methods, Research Report 32, Helsinki University of Technology, Finland, 1995.

Pätsch, J., Kühn, W., Moll, A., and Lenhart, H.: ECOHAM4 user guide, Technical Report 1-2009, Institut für Meereskunde, Hamburg, Germany, 2009.

Sakov, P. and Oke, P. R.: A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filter, Tellus, 60A, 361–371, 2008.

Sakov, P., Evensen, G., and Bertino, L.: Asynchronous data assimilation with the EnKF, Tellus, 62A, 24–29, 2010.

Simon, E. and Bertino, L.: Application of the Gaussian anamorphosis to assimilation in a 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment, Ocean Sci., 5, 495–510, doi:10.5194/os-5-495-2009, 2009.

Simon, E. and Bertino, L.: Gaussian anamorphosis extension of the DEnKF for combined state and parameter estimation: application to a 1-D ecosystem model, J. Mar. Syst., 89, 1–18, 2012.

Skogen, M. and Søiland, H.: A user's guide to NORWECOM v2.0. The NORWegian Ecological Model system, Technical Report Fisken og Havet 18, Institute of Marine Research, Norway,1998.

Stein, W. E. and Keblis, M. F.: A new method to simulate the triangular distribution, Math. Comp. Model., 49, 1143–1147, 2009.

Umlauf, L. and Burchard, H.: Second-order turbulence closure models for geophysical boundary layers, A review of recent work, Cont. Shelf Res., 25, 795–827, 2005.

Zhou, H., Gómez-Hernández, J. J., Hendricks Franssen, H.-J., and Li, L.: An approach to handling non-Gaussianity of parameters and state variables in ensemble Kalman filtering, Adv. Water Res., 34, 844–864, 2011.