GHOSH v1.0.0: a novel Gauss-Hermite High-Order Sampling Hybrid ensemble filter for computationally efficient data assimilation in geosciences - Part 1

Simone Spada[1], Anna Teruzzi[1], Stefano Maset[2], Stefano Salon[1], Cosimo Solidoro[1], and Gianpiero Cossarini[1]

[1]National Institute of Oceanography and Applied Geophysics - OGS
[2]University of Trieste

Correspondence: Simone Spada (sspada@ogs.it)

Abstract. Data assimilation is used in a number of geophysical applications to optimally integrate observations and model knowledge. Providing an estimation of both state and uncertainty, ensemble algorithms are ~~one of~~ among the most successful data assimilation approaches. Since the estimation quality depends on the ensemble, the sampling method is a crucial step in ensemble data assimilation. Among other options to improve the capability of generating an effective ensemble, a sampling method featuring a higher polynomial order of approximation represents a novelty. Indeed, the order of the most widespread ensemble algorithms is usually equal to or lower than 2. We propose a novel hybrid ensemble algorithm, the Gauss-Hermite High-Order Sampling Hybrid (GHOSH) filter, version 1.0.0~~, which we apply in a twin experiment (based on Lorenz96) and in a realistic geophysical application~~. In the ~~most error components~~directions where the uncertainty is larger, the GHOSH ~~sampling method can achieve~~ filter's sampling method achieves a higher order of approximation than in other ensemble based filters, without increasing the asymptotic computational complexity that is comparable to the one of second-order deterministic filters. To evaluate the benefits of the higher approximation order, a set of ~~thousands~~ twin experiments of Lorenz96 simulations has been carried out using the GHOSH filter and a ~~second order~~ second-order ensemble Kalman filter (SEIK; singular evolutive interpolated Kalman filter). The comparison between the GHOSH and the SEIK filter has been done by varying a number of data assimilation settings~~: ensemble size, inflation, assimilated observations, and initial conditions~~. The twin-experiment results show that GHOSH outperforms SEIK in most of the assimilation settings up to a ~~69%~~ 70% reduction of the root mean square error on assimilated and non-assimilated variables. ~~A parallel implementation of the GHOSH filter has been coupled with a realistic geophysical application: a physical-biogeochemical model of the Mediterranean Sea with assimilation of surface satellite chlorophyll. The simulation results are validated using both semi-independent (satellite chlorophyll) and independent (nutrient concentrations from an in-situ climatology) observations. Results show the feasibility of GHOSH implementation in a realistic three-dimensional application. The GHOSH assimilation algorithm improves the agreement between forecasts and observations without producing unrealistic effects on the non-assimilated variables. Furthermore, the sensitivity analysis on GHOSH setup indicates that the use of a higher order of convergence substantially improves the performance of the assimilation with respect to nitrate (i.e., one of the non-assimilated variables). In view of potential implementation of the GHOSH filter in operational applications, it should be noted that GHOSH and SEIK filters have not shown significant differences in terms of~~

~~time to solution, since, as in all ensemble-like Kalman filters, the model integration is by far more computationally expensive than the assimilation scheme.~~

1 Introduction

Data assimilation (DA) methodologies provide a conceptual framework to integrate the information contents embedded in observations and numerical models, and plays a pivotal role in deriving accurate estimates of the state of earth systems components and reducing the uncertainties of geophysical systems ~~(Bannister (2017a); Houtekamer and Zhang (2016); Edwards et al. (2015); Lah~~ ~~). In fact, the use of DA has been steadily increasing in the last decade and it is now ubiquitous in earth sciences, also thanks to the unprecedented -and always increasing- availability of both data and computational capability.~~ (among the others, see the methods and ap .

~~The Kalman filters and the variational approaches are the two main classes of DA algorithms (Kalman (1960), Talagrand and Courtier (19)). Both approaches are based on Gaussian approximation and can be interpreted in a Bayesian framework, and include an estimate of the most probable state and its evolution, and possibly the associated variance. Both methods are strictly valid only for linear models, but modified implementations have been proposed for application to non-linear ones (Carrassi et al. (2018)).~~

~~Recently, thanks~~ Thanks to the scalability of parallel implementations, the use of ensemble algorithms ~~(e.g., EnKF Evensen (1994) and EnVar Bannister (2017b))~~ (see e.g., Vetra-Carvalho et al., 2018; Houtekamer and Zhang, 2016; Bannister, 2017a) have been proposed to estimate uncertainty and improve assimilation skills in Kalman filters and variational methodologies. ~~Furthermore~~On the other hand, some of the strong points of the ~~two approaches~~ ensemble and variational have been merged ~~, leading to hybrid filters (e.g. , Hamill and Snyder (2000))~~in hybrid filters (e.g., Hamill and Snyder, 2000). Moreover, recent developments of EnKF have been conceived to face increasing non-linearities and model complexity that are also related to the expanding availability of observations and computational resources (see e.g., Vetra-Carvalho et al., 2018).

However, the definition of the strategy for ensemble generation is not a trivial task ~~(Moore et al. (2019)). Straightforward Montecarlo~~ (see e.g., Moore et al., 2019). Straightforward Monte Carlo approaches are not usually a viable option in geoscience applications, because they would require a too large number of ensemble members, and consequently, computational effort. The number of ensemble members can be reduced by adopting deterministic ~~(or semi-deterministic) sampling methods . These methods provide an ensemble mean of future state~~ sampling methods (as opposed to stochastic EnKF methods, see e.g. Carrassi et al . Examples of deterministic EnKF using second-order sampling methods are SEIK (Pham, 2001) and ETKF (Bishop et al., 2001) . Extending the definition of second-order exact methods in Pham (2001), with the term “order” we refer to the polynomial order of approximation of the filter or of its sampling method. Namely, in the case of a h^{th}-order approximation, the ensemble has the property of providing a forecast mean with no error ~~after evolving the ensemble,~~ as long as the evolution function used for forecasting (i.e., the model) is a polynomial of ~~the same order of the sampling method. Indeed, the use of deterministic or semi-deterministic ensembles, built using second-order sampling methods, are now quite common (as in SEIK, Pham (1996) , or ETKF, Bishop et al. (2001)~~order h. As proven in Appendix A, this is achieved by sampling an ensemble that matches the first h statistical moments of the uncertainty probability density function (pdf) before evolution (also called *prior* in Bayesian statistics).

~~At the same time, most~~ Most of the models used in geoscience applications are based on systems of differential equations that cannot be represented by a ~~second order~~ second-order polynomial and in all of these cases the ~~second order sampling methods provide only a second order approximation of the model~~second-order sampling methods provides a non-exact estimation of the forecast mean, which is affected by an error strictly related to the error made in approximating the model by a second-order polynomial. Furthermore, the ~~second order~~ second-order approximation is more effective the closer the ensemble members are to each other ~~, thus, the larger the ensemble spread~~ (i.e., small uncertainty), thus the higher the uncertainty the worse will be the approximation error in the mean computation. Since ~~a relatively large ensemble spread is often required in~~ the state estimation is often affected by a relatively high uncertainty in geosciences data assimilation applications, this approximation error may be not negligible.

A potential strategy to reduce this error is the use of a higher order of approximation, but this would require a larger ensemble with respect to the ~~second order case and consequently bigger~~ second-order methods and consequently larger computational costs, given that a higher order of approximation implies a larger number of ensemble members to represent the same ~~error subspace~~uncertainty subspace[1]. For instance, ~~second order~~ second-order methods use an ensemble of $r+1$ members to span an r-dimensional ~~error subspace (e.g., Pham (1996))~~uncertainty subspace (e.g., Pham, 2001), while it can be proven (see Appendix C, Section C4) that $2r$ ensemble members are needed to achieve order 3 in the same subspace ~~(an example of 3rd order method with $2r+1$ ensemble members is the unscented Kalman filter, Wan and Van Der Merwe (2000))~~(see e.g., Wan and Van Der Merwe, 2000; Ambadan and Tang, 2009; Luo and Moroz, 2009, for examples of third-order method with $2r$ -
.

The approximation order can ~~be increased further~~ also be increased by using properly weighted ensemble members. A weighted ensemble based on multi-dimensional Gauss-Hermite quadrature rule ~~(Mastroianni and Milovanovic (2008))~~ (Mastroianni and M could arbitrarily increase the order of approximation, but always at the cost of a larger ensemble size ~~(Ito and Xiong (2000))~~(Ito and Xiong, 2000). Weighted ensembles are also exploited in particle filter methods ~~(Bocquet et al. (2010))~~(van Leeuwen et al., 2019), where weights are used to evolve both the ensemble members and their probability.

In the present work, we propose a novel weighted ensemble method based on a new high-order sampling, that provides ensemble mean estimates of order higher than 2 without increasing the number of ensemble members, and therefore without major impacts on the computational cost.

The proposed high-order filter (as its sampling method) exploits a Gauss-Hermite-like quadrature rule, along with a principal component analysis (PCA), and we named it Gauss-Hermite high-order sampling hybrid (GHOSH) filter. Members and related weights are chosen in such a way as to guarantee accuracy of order higher than 2 for a subset of ~~relevant modes(a~~

[1] In the space of the possible state vectors, the uncertainty subspace is the subspace generated by the ensemble members (i.e., the smallest subspace that contains all the ensemble members). The uncertainty subspace was originally called error subspace (see Nerger et al., 2005, 2012, for an introduction to the error subspace concept), but in the context of this work it is more natural to interpret the ensemble as a proxy of the uncertainty, avoiding unnecessary confusion with, for instance, the approximation error.

~~relevant subspace)~~principal modes, and no less than 2 for the remaining ones. The GHOSH filter exploits a hybrid approach that considers the ~~model uncertainty described~~ uncertainty as composed by a constant and a time-evolving ~~part based on the evolution of an ensemble~~ensemble-based part.

The reliability of the new filter is proven ~~first~~ in a large set of twin experiments based on an idealized model commonly used to test DA methodologies ~~(Lorenz96, Lorenz (1996)) and then~~ (Lorenz, 1996). In a companion paper (Spada et al., 2024) the new filter is tested in the much more complex and realistic marine biogeochemical application currently used in Copernicus Marine CMEMS System ~~(Salon et al. (2019))~~(Salon et al., 2019), featuring assimilation of satellite observations.

~~The high non-linearity of the biogeochemical processes makes the chosen application a challenging framework suitable for a GHOSH test implementation. Indeed, biogeochemical DA is relatively recent both in terms of methods and assimilated variables (Fennel et al. (2019)). Surface chlorophyll concentration, which is estimated by satellite ocean colour observations, is the most commonly assimilated biogeochemcial variable since its relatively high coverage and its near real-time availability make it suitable also for operational applications (e.g., Song et al. (2016); Tsiaras et al. (2017); Santana-Falcón et al. (2020); Pradhan et al.).~~

Section 2 introduces the novel elements of the high-order sampling and Section 3 presents a synthetic description of the GHOSH algorithm and its localized version. ~~Additional mathematical details are provided in Appendix A. The two sets of numerical experiments (a twin experiment of Lorenz96 model and a realistic 3D marine biogeochemical application) are introduced (Section 4) and their results presented (Section 5)~~The Lorenz96 numerical experiment is introduced in Section 4 and its results presented in Section 5. The algorithm and the experimental results are discussed in Section 6. Finally, additional mathematical and algorithmic details along with examples are provided in appendices.

2 The high-order sampling

In this work we propose a novel method to sample the ensemble members in data assimilation applications. Compared to deterministic second-order square-root sampling methods ~~(Carrassi et al. (2018))~~(see e.g., Carrassi et al., 2018), the novel strategy achieves a higher order of approximation without increasing the ensemble size. The strategy is based on the choice, among all the ~~second order~~ second-order ensembles, of those that feature a higher order of approximation along the ~~most relevant~~ principal components of the ~~error~~ uncertainty (i.e., those directions where the ensemble spread is larger and the approximation is consequently worse). To better understand the ~~idea~~GHOSH approach, Fig. 1 shows an ~~error~~ uncertainty represented by a 3-dimensional standard normal distribution (~~in shadow the spherical errors isosurface~~spherical uncertainty isosurfaces are represented by shadowed areas). Using the ~~second order exact sampling (Pham (1996))~~second-order-exact sampling (Pham, 2001), 4 ensemble members (yellow points) are needed to represent ~~the uncertainty produced by this error distribution~~this uncertainty probability density function. In a xyz Cartesian coordinate system, the 4 ensemble members are the vertices of a regular tetrahedron, and all the possible ~~second order~~ second-order samplings correspond to a rotation of this tetrahedron. With the appropriate rotation, the vertices draw (project) a square in the xy plane ~~. In this way, we achieved a third order~~ (Appendix C, Section C3.1), achieving a third-order approximation in the xy plane~~, because~~ . Indeed, it can be ~~proved~~

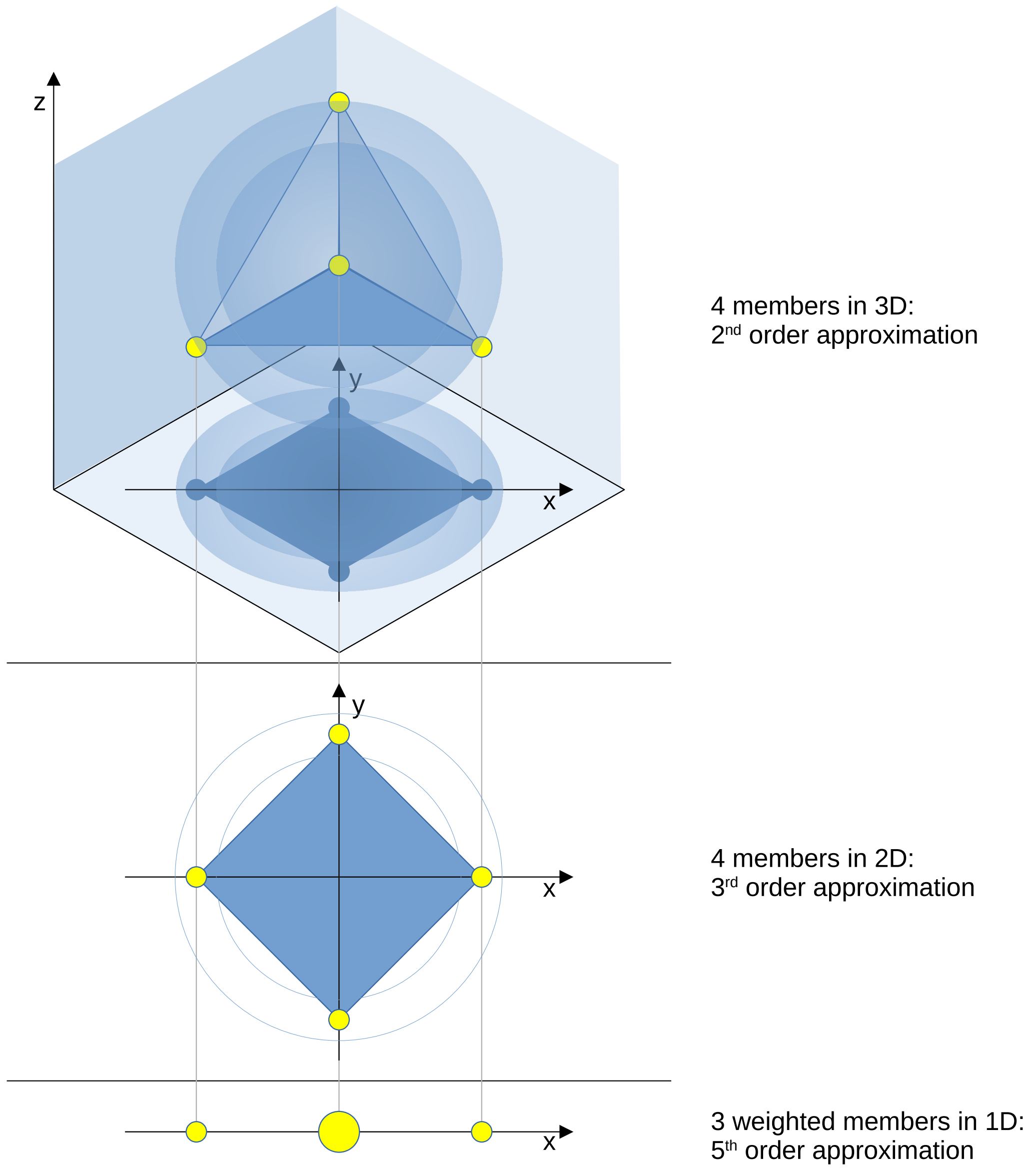


Figure 1. 4 tetrahedron-shaped ensemble members (yellow dots) in a 3-dimensional space (top) represent a second-order approximation of a standard normal distribution (in shadow the spherical isosurfaces). The same 4 members form a square in 2 dimensions (middle), producing a third-order approximation. By projecting further into 1 ~~direction~~ dimension (bottom) and by assigning proper weights, ~~they become~~ a fifth-order ~~weighted ensemble~~approximation is obtained.

proven that 4 square-shaped ensemble members provide a ~~third order~~ third-order approximation for a 2-dimensional normal distribution (Appendix C, Section C2.4). Similarly, a further projection on a single axis returns 4 points or, as in the figure, only 3 points, one of which represents 2 of the original points. In the 1D case, the highest approximation order that can be obtained using 4 members is still limited to 3, unless proper weights are assigned to each ensemble member. In fact, 4 weighted members (or even only 3, see Appendix C at Section C1.3) are sufficient to reach an approximation order as high as 5 on the x axis. ~~Finally, what we have obtained is a~~ In summary, Fig. 1 represents a weighted ensemble with approximation order 2 in the xyz 3-dimensional ~~second order weighted ensemble which is a third-order approximation~~ space, approximation order 3 in the xy plane and ~~a fifth order approximation~~ 5 along the x axis ~~.~~

~~This means~~ (Appendix C, Section C3.2). Meaning that, if the z component of the distribution was less relevant than the x and y components (e.g., a non-standard normal distribution with smaller variance on the z coordinate), then ~~we obtained a third order approximation in the~~ the GHOSH third-order approximation mitigates the approximation error in the x and y dimensions where the spread is higher. And, if the variance is ~~bigger~~ larger on x than y, then ~~we achieved a fifth order approximation~~ the achieved fifth-order approximation reduces the error more efficiently where the spread is maximum (x coordinate).

The idea described above can be generalized considering that any probability distribution (under reasonable hypothesis) can be sampled with an arbitrary high order of approximation with a weighted ensemble that, in the special case of the Gaussian distribution, is represented by the nodes and weights of the Gauss-Hermit quadrature rule ~~(Mastroianni and Milovanovic (2008))~~(Mastroianni and Milovanovic, 2008, cap. 2). Once the approximation order (h larger than 2) and the ensemble size ($r+1$) are fixed, the ~~h~~h^{th}-order weighted ensemble is computed for a relatively small s-dimensional subspace (s lower than r); then, a new $(r+1)$-sized ~~second order~~ second-order weighted ensemble is computed such that its projections along the ~~most relevant error directions~~ uncertainty principal components coincide with the ~~h~~h^{th}-order ensemble. The resulting ensemble has a ~~second order~~ second-order approximation in the r-dimensional subspace and a ~~h~~h^{th}-order approximation in the s-dimensional subspace of the ~~most relevant error~~ principal uncertainty directions. By comparison, a ~~second order exact~~ second-order-exact sampling (e.g., SEIK) with $r+1$ ensemble members would provide a ~~second order~~ second-order approximation for the whole r-dimensional subspace.

The two phases of the high-order sampling ~~is~~ (*initialization* and *sampling*) are described in the following from an algorithmic point of view. ~~Mathematical details, instead, can be found in Appendix A.The high-order sampling presents two phases:~~ The *initialization* ~~and *sampling*. The former~~ includes the steps that must be executed only once, while the ~~latter~~ *sampling* presents the ensemble generation procedure. Mathematical details, can be found in Appendix A.

2.1 Initialization

First, 3 hyper-parameters r, h and s are chosen in the inizialization: r is the dimension of the ~~error~~ uncertainty subspace, which implies an ensemble size of $r+1$ members; h is the higher approximation order reached in the most relevant directions; s represents the number of principal components that are approximated with order h and must be smaller than r.

Then, the real numbers $\cancel{u_j^m}$ $u_{m,j}$ and v_j, with $j \in \{1,\dots,s\}$ and $m \in \{1,\dots,r+1\}$, are calculated as a solution of the nonlinear system

$$\begin{cases} \vdots \\ \sum_{m=1}^{r+1} u_{m,j_1} \cdots u_{m,j_\xi} {v_m}^{2-\xi} = \mu_{j_1,\ldots,j_\xi} \\ \vdots \end{cases} \tag{1}$$

with one equation for each $\xi \in \{0,\ldots,h\}$ and for each $j_1,\ldots,j_\xi \in \{1,\ldots,s\}$~~.~~, for a total of $\left(s^{h+1}-1\right)/\left(s-1\right)$ equations. The system represents the matching between the statistical moments of the ensemble and the normalized uncertainty pdf. In fact, its first equation, i.e.,

$$\sum_{m=1}^{r+1} {v_m}^2 = 1,$$

ensures that the weights will sum to 1; the following block of s equations, i.e.,

$$\begin{cases} \sum_{m=1}^{r+1} u_{m,1} v_m = 0 \\ \vdots \\ \sum_{m=1}^{r+1} u_{m,s} v_m = 0, \end{cases}$$

ensures that the ensemble has 0-mean; the third block of s^2 equations, i.e.,

$$\begin{cases} \sum_{m=1}^{r+1} u_{m,1} u_{m,1} = \delta_{1,1} = 1 \\ \vdots \\ \sum_{m=1}^{r+1} u_{m,1} u_{m,s} = \delta_{1,s} = 0 \\ \\ \vdots \\ \\ \sum_{m=1}^{r+1} u_{m,s} u_{m,1} = \delta_{s,1} = 0 \\ \vdots \\ \sum_{m=1}^{r+1} u_{m,s} u_{m,s} = \delta_{s,s} = 1, \end{cases}$$

forces the ensemble covariance matrix to be equal to the identity matrix, and so on and so forth. The general equation reported in system (1) represents all the moment-matching equations up to order h, where $\mu_{j_1,\ldots,j_\xi}$ are the statistical moments of a

s-dimensional probability distribution that approximates the assumed ~~error~~ uncertainty distribution shape. Such probability distribution must be uncorrelated, normalized and have 0 mean, since the ensemble mean and covariance will be later adjusted by the sampling algorithm, which will also rescale the other moments accordingly (see equation (6)). A typical choice is the ~~Normal~~ standard normal distribution $\mathcal{N}(\boldsymbol{z};0,\mathbf{I}_s)$, i.e.,

$$\mu_{j_1,\ldots,j_\xi} = \int\limits_{\mathbb{R}^s} z_{j_1}\cdots z_{j_\xi}\mathcal{N}(\boldsymbol{z};0,\mathbf{I}_s)\,d\boldsymbol{z}. \tag{2}$$

In the Gaussian case, the moments $\mu_{j_1,\ldots,j_\xi}$, which are other hyper-parameters of the GHOSH algorithm, can easily be computed (i.e., without solving numerically the integral, see Appendix A, equations (A6) and (A7)). Changing the value of the moments $\mu_{j_1,\ldots,j_\xi}$ allows to explore other non-Gaussian probability distributions.

Note that, in order to actually have at least a solution of the system, the number of independent equations cannot be ~~bigger~~ larger than the number of variables, which implies that ~~a bigger~~ given r (and consequently the ensemble size) a larger approximation order h ~~requires~~ implies a smaller s ~~or a bigger r~~(i.e., a smaller number of components approximated with order h).

The system solution is used to define $\boldsymbol{\Omega}_h$, as the $(r+1)\times s$ matrix with coordinates $\cancel{u_j^m}$ $u_{m,j}$ (i.e., $u_{m,j}$ is the entry of the matrix $\boldsymbol{\Omega}_h$ at row m and column j), and the ensemble weight vector $\boldsymbol{w}\in\mathbb{R}^{(r+1)}$, as the element-wise square of $\boldsymbol{v}$, i.e.,

$$\boldsymbol{w} = \left({v_1}^2,\ldots,{v_{r+1}}^2\right). \tag{3}$$

2.2 Sampling

Given an N-dimensional state space, in the sampling phase a new ensemble is generated based on the ensemble mean $\boldsymbol{x}_i\in\mathbb{R}^N$ and the covariance matrix $\mathbf{L}_i\mathbf{L}_i^T$. Differently from non-varying quantities (like those defined in the *initialization* phase), i-indexed quantities are specific of each sampling. The columns of the $N\times r$ matrix $\mathbf{L}_i$ adopted to obtain a r-rank factorization of the covariance matrix represent the base of the ~~error~~ uncertainty subspace spanned by the ensemble members.

A random $s\times s$ orthogonal matrix $\boldsymbol{\Omega}_i^{rnd}$ is drawn, then the $(r+1)\times(s+1)$ matrix $\left(\begin{array}{c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h\boldsymbol{\Omega}_i^{rnd}\end{array}\right)$ is completed to an $(r+1)\times(r+1)$ orthogonal matrix by the $(r+1)\times(r-s)$ matrix $\boldsymbol{\Omega}_i^{\perp}$, i.e.,

$$\left(\begin{array}{c|c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h\boldsymbol{\Omega}_i^{rnd} & \boldsymbol{\Omega}_i^{\perp}\end{array}\right)^T\left(\begin{array}{c|c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h\boldsymbol{\Omega}_i^{rnd} & \boldsymbol{\Omega}_i^{\perp}\end{array}\right) = \mathbf{I}_{r+1}, \tag{4}$$

where $\boldsymbol{v}$ is the array of the square roots of the weights (as in equation (3)) and $\mathbf{I}_{r+1}$ is the identity matrix of rank $r+1$. ~~Many procedures (e.g., Pham (1996)) are available to generate a random orthogonal matrix (or to randomly complete a given set of orthonormal vectors to an orthogonal matrix)~~A procedure for randomly building or completing orthogonal matrices is described in Appendix B.

Now, the $(r+1)\times r$ matrix $\boldsymbol{\Omega}_i$, defined as

$$\boldsymbol{\Omega}_i = \left(\begin{array}{c|c} \boldsymbol{\Omega}^h\boldsymbol{\Omega}_i^{rnd} & \boldsymbol{\Omega}_i^{\perp}\end{array}\right), \tag{5}$$

can be used to build the $N \times r$ ensemble matrix $\mathbf{X}_i$:

$$\mathbf{X}_i = \boldsymbol{x}_i \mathbb{1}_{1\times(r+1)} + \mathbf{L}_i {\mathbf{C}_i}^T {\boldsymbol{\Omega}_i}^T \mathbf{W}^{-\frac{1}{2}}, \tag{6}$$

where $\mathbb{1}$ is a matrix (of the subscripted size) filled with ones, $\mathbf{W} = \mathrm{diag}(\boldsymbol{w})$ is the $(r+1)\times(r+1)$ diagonal matrix of weights and $\mathbf{C}_i$ is a $r \times r$ orthogonal change-of-basis matrix such that

$$ {\mathbf{L}_i}^T {\boldsymbol{\Lambda}_i}^{-1} \mathbf{L}_i = \mathbf{C}_i^T \mathbf{E}_i \mathbf{C}_i. \tag{7}$$

In the last equation, the right-hand side is an eigenvalue decomposition of the left-hand side, with $\mathbf{E}_i$ being an $r \times r$ diagonal matrix of eigenvalues in descending order and $\boldsymbol{\Lambda}_i$ being an appropriate $N \times N$ matrix. The purpose of $\boldsymbol{\Lambda}_i$ is to weight the product between $\mathbf{L}_i$ and its transposed~~and thereforeto define~~ ; therefore, $\boldsymbol{\Lambda}_i$ defines the criterion that decides the ~~relevance of each PCA component. In the simplest cases~~relative importance of different parts of the state vector, affecting the uncertainty principal components. In relatively simple applications, $\boldsymbol{\Lambda}_i$ can be the ~~identical~~ identity matrix, but in most scenarios, $\boldsymbol{\Lambda}_i$ needs to be properly designed ~~(see sections ?? and 6)~~(see Section 6 and Spada et al., 2024).

After $\mathbf{X}_i$ computation, the ensemble members can be retrieved from the column of the matrix, i.e.,

$$\mathbf{X}_i = \left(\boldsymbol{x}_i^1, \dots, \boldsymbol{x}_i^{(r+1)}\right). \tag{8}$$

3 The GHOSH filter

Similar to other data assimilation ensemble schemes, the GHOSH filter provides an estimate of the state of a system at some ~~time~~ pre-fixed times t_i in terms of the state vector and the covariance matrix ~~that represents the error estimate of the state vector~~representing the estimation uncertainty. Hence, at time t_i, a forecast is ~~produced in the form of~~ composed by the forecast state vector $\boldsymbol{x}_i^f \in \mathbb{R}^N$ (N is the dimension of the state space) and ~~the forecast error~~ by the forecast uncertainty covariance matrix $\mathbf{P}_i^f$. ~~Moreover, if~~ If an observation vector $\boldsymbol{y}_i$ is available at time t_i, the information is assimilated in the analysis state vector $\boldsymbol{x}_i^a$ with ~~error~~ uncertainty covariance matrix $\mathbf{P}_i^a$.

Being an ensemble-based filter scheme, the GHOSH filter represents these quantities by an ensemble of state vectors, e.g.,

$$\mathbf{X}_i{}^{\cancel{a}\,f} = \left(\boldsymbol{x}_i{}^{\cancel{a,1}\,f,1}, \dots, \boldsymbol{x}_i{}^{\cancel{a,(r+1)}\,f,(r+1)}\right) \tag{9}$$

of $r+1$ model state realizations. However, differently from other ensemble filters that uniformly weight the ensemble members, the GHOSH filter assigns to ~~$\mathbf{X}_i^a$~~ $\mathbf{X}_i^f$ a vector of corresponding weights

$$\boldsymbol{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_{r+1} \end{pmatrix}. \tag{10}$$

The state estimate is given by the weighted mean

$$\boldsymbol{x}_i{}^{a\,f} = \sum_{j=1}^{r+1} \boldsymbol{x}_i{}^{a,j\,f,j} w_j = \mathbf{X}_i{}^{a\,f} \boldsymbol{w}, \tag{11}$$

while the ~~error~~ uncertainty covariance is approximated by the ensemble covariance matrix

$$\mathbf{P}_i{}^{a\,f} \approx \mathbf{X}_i{}^{a\,f} \mathbf{W}\, \mathbf{X}_i{}^{a\,f\,T} - \boldsymbol{x}_i{}^{a\,f}\, \boldsymbol{x}_i{}^{a\,f\,T}, \tag{12}$$

with $\mathbf{W} = \mathrm{diag}\,(\boldsymbol{w})$ being the diagonal matrix of weights.

The GHOSH algorithm consists of 5 phases, i.e., *initialization*, *forecast*, *forecast high-order sampling*, *analysis* and *analysis high-order sampling* (Fig. 2). The *initialization* is performed once at the beginning of the filter application, the *forecast* and *forecast high-order sampling* phases are carried out at each time t_i, and *analysis* and *analysis high-order sampling* are executed only when ~~an observation is~~ observations are available.

The *analysis* equations are a novel weighted version of the ensemble Kalman filter equations ~~(Evensen (1994))~~(Evensen, 1994), while the *forecast* phase equations rely on a hybrid approach. In fact, the forecast uncertainty is obtained by combining the ensemble covariance (weighted by a forgetting factor) and a ~~model error~~ parametric covariance matrix $\mathbf{Q}_i$, which is built from the existing knowledge of the system, as done, for instance, for the background matrix in many variational schemes ~~(Bannister (2017a, 2008))~~(Bannister, 2017a, 2008).

The method includes two resampling phases, after *analysis* and after *forecast* (Fig. 2), at which the whole ensemble is rebuilt using the high-order sampling method (Section 2). The resampling produces a sample of state vectors that matches the moments of the ~~error~~ uncertainty probability density function with better precision than in the other commonly used ensemble DA algorithms~~. In fact, projecting the ensemble in the subspace spanned by the most relevant components of the ensemble principal components analysis (PCA), the corresponding discrete probability distribution function matches the moments of the projection of the error uncertainty up to order $h \geq 2$~~, resulting in an order of approximation greater than 2 (see Appendix A) ~~. When the sampling is applied after the forecast, GHOSH allows to take further~~ in the principal uncertainty modes. In this way the GHOSH filter takes advantage of the high ~~convergence~~ approximation order of the ~~method, especially when the model error is not neglected.~~ sampling method twice: before applying the model operator and before applying the observation operator.

3.1 The global GHOSH filter

3.1.1 Initialization

First, ~~following~~ as the high-order sampling *initialization* phase (Section 2.1), the three GHOSH hyper-parameters r (the dimension of the ~~error~~ uncertainty subspace), h (the higher approximation order reached in the most relevant directions) and s (the number of principal components that are approximated with order h) are used to compute the sampling matrix $\mathbf{\Omega}^h$ and the weights vector $\boldsymbol{w}$.

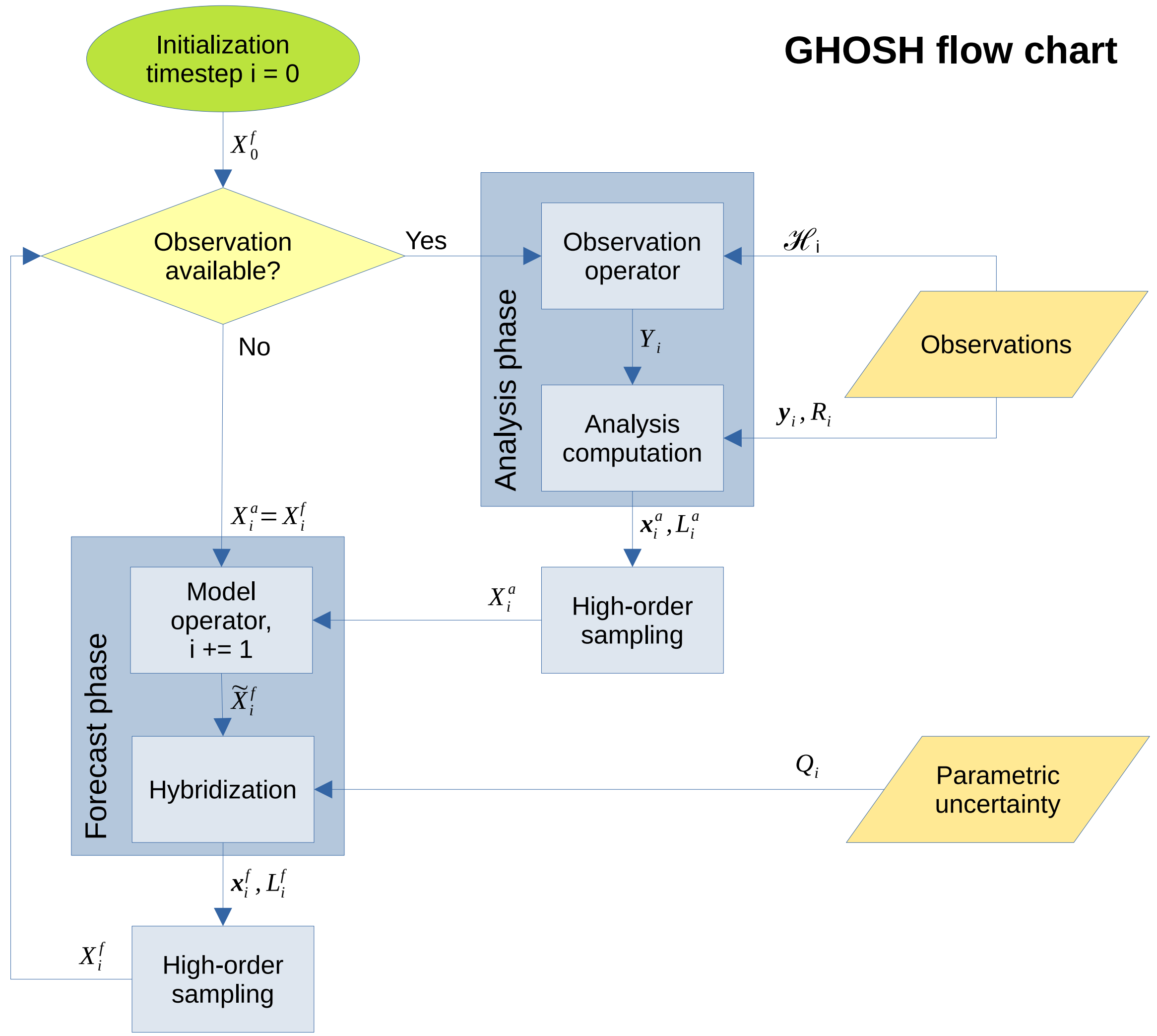


Figure 2. GHOSH filter flow chart. Subscripts represent the time steps of the numerical simulation.

A starting weighted ensemble must be provided, with the coordinates of $\boldsymbol{w}$ as weights. Such ensemble can come from any previous forecast/assimilation using the GHOSH filter, or it can be built from scratch with an ensemble generation method (e.g., a PCA on historical data followed by the GHOSH sampling method described in Section 2.2). The ensemble members are stored in the columns of ensemble matrix $\mathbf{X}_0^f$, i.e.,

$$\mathbf{X}_0^f = \left(\boldsymbol{x}_0^{f,1}, \ldots, \boldsymbol{x}_0^{f,(r+1)}\right), \tag{13}$$

with the subscripted index 0 representing the first time step.

The mean $\mathbf{x}_0^f$, can be computed as

$$\mathbf{x}_0^f = \mathbf{X}_0^f \mathbf{w}. \tag{14}$$

3.1.2 Analysis phase

~~If~~ When observations are available at time step i (Fig. 2), $\boldsymbol{y}_i \in \mathbb{R}^n$ represents the observation array and $\mathcal{H}_i$ is the observation operator, which incorporates all operations needed to obtain the ~~measured~~ observed quantities from the state vector. This operator is used on each ensemble member $\boldsymbol{x}_i^{f,k}$ to build the $n \times (r+1)$ matrix $\mathbf{Y}_i$, i.e.,

$$\boldsymbol{y}_i^k = \mathcal{H}_i\left(\boldsymbol{x}_i^{f,k}\right) \tag{15}$$

$$\mathbf{Y}_i = \left(\boldsymbol{y}_i^1, \ldots, \boldsymbol{y}_i^{(r+1)}\right), \tag{16}$$

with $k \in \{1, \ldots, r+1\}$.

The base $\tilde{\mathbf{L}}_i^a$ of the ~~error~~ uncertainty subspace spanned by the ensemble is computed by

$$\tilde{\mathbf{L}}_i^a = \mathbf{X}_i^f \mathbf{T}, \tag{17}$$

where $\mathbf{T}$ is an $(r+1) \times r$ full-rank matrix with zero column sums. A ~~reasonable choice is~~ matrix fulfilling these requirements, that also makes product in equation (17) computationally fast, is

$$\mathbf{T} = \left(\begin{array}{ccc} & \mathbf{I}_r & \\ \hline 0 & \cdots & 0 \end{array}\right) - \boldsymbol{w}\mathbb{1}_{1\times r}, \tag{18}$$

with $\mathbf{I}_r$ being the identity matrix of size r and $\mathbb{1}$ is a matrix (of the subscripted size) filled with ones. Equation (18) implies the use of the ensemble anomalies as basis members, and it is a common choice in SEIK implementations ~~(e.g., Triantafyllou et al. (2003))~~(e.g., Triantafyllou et al., 2003), but other $\mathbf{T}$s can be explored without affecting the ~~algorithm (e.g., see Nerger et al. (2012))~~main structure of the algorithm (see e.g., Nerger et al., 2012).

The analysis ~~error~~ uncertainty covariance matrix $\mathbf{P}_i^a$ is obtained by

$$\mathbf{P}_i^a = \tilde{\mathbf{L}}_i^a \mathbf{A}_i^a \tilde{\mathbf{L}}_i^{a\,T}, \tag{19}$$

where

$$\mathbf{A}_i^{a\,-1} = \mathbf{T}^T W^{-1} \mathbf{T} + \mathbf{Z}_i^T \mathbf{R}_i^{-1} \mathbf{Z}_i, \tag{20}$$

and

$$\mathbf{Z}_i = \mathbf{Y}_i \mathbf{T}, \tag{21}$$

$W = \mathrm{diag}\,(\boldsymbol{w})$ being the $(r+1) \times (r+1)$ diagonal matrix of weights and $\mathbf{R}_i$ the $n \times n$ ~~error~~ covariance matrix representing the uncertainty in the observation y_i.
An opportune change in basis is used to obtain the simpler decomposition

$$\mathbf{P}_i^a = \mathbf{L}_i^a \mathbf{L}_i^{aT}, \tag{22}$$

where

$$\mathbf{L}_i^a = \tilde{\mathbf{L}}_i^a \mathbf{S}_i^a \tag{23}$$

and $\mathbf{S}_i^a$ is the symmetric square root of $\mathbf{A}_i^a$, i.e.,

$$\mathbf{S}_i^a \mathbf{S}_i^a = \mathbf{A}_i^a. \tag{24}$$

Finally, the analysis state $\boldsymbol{x}_i^a$ is estimated by

$$\boldsymbol{x}_i^a = \boldsymbol{x}_i^f + \mathbf{L}_i^a \mathbf{S}_i^a \mathbf{Z}_i^T \mathbf{R}_i^{-1} \left(\boldsymbol{y}_i - \mathbf{Y}_i \boldsymbol{w}\right). \tag{25}$$

Equation (25) differs from the usual ensemble Kalman filter equations in the use of $\mathbf{Y}_i\boldsymbol{w}$ in the last term instead of $\mathcal{H}_i\left(\boldsymbol{x}_i^f\right)$. The two are the same if the observation operator $\mathcal{H}_i$ is linear. However, in the general case, $\mathbf{Y}_i\boldsymbol{w}$, which relies on the whole ensemble instead of the ensemble mean only, is a better estimator of the expected value of the observed quantities. Moreover, ensembles generated with GHOSH's sampling method lead to a further advantage due to the higher order of approximation.

3.1.3 Analysis high-order sampling

The $N \times (r+1)$ ensemble matrix $\mathbf{X}_i^a = \left(\boldsymbol{x}_i^{a,1}, \ldots, \boldsymbol{x}_i^{a,(r+1)}\right)$, whose columns are the ensemble members, is obtained by

$$\mathbf{X}_i^a = \boldsymbol{x}_i^a \mathbb{1}_{1\times(r+1)} + \mathbf{L}_i^a \mathbf{C}_i^{aT} \boldsymbol{\Omega}_i^{aT} W \mathbf{W}^{-\frac{1}{2}}. \tag{26}$$

The sampling procedure is described in Section 2.2; therefore, $\mathbf{C}_i^a$ and $\boldsymbol{\Omega}_i^a$ are computed as $\mathbf{C}_i$ and $\boldsymbol{\Omega}_i$ used in equation (6).

3.1.4 Forecast phase

In this phase, the time step index i is increased by 1 and the ensemble is evolved through the model operator $\mathcal{M}_i$ (which usually represents the numerical integration of a system of differential equations), i.e., for $k \in \{1, \ldots, r+1\}$,

$$\tilde{\boldsymbol{x}}_i^{f,k} = \mathcal{M}_i\left(\boldsymbol{x}_{i-1}^{a,k}\right), \tag{27}$$

$$\tilde{\mathbf{X}}_i^f = \left(\tilde{\boldsymbol{x}}_i^{f,1}, \ldots, \tilde{\boldsymbol{x}}_i^{f,(r+1)}\right). \tag{28}$$

The forecast state $\boldsymbol{x}_i^f$ is estimated as the weighted mean of the ensemble by

$$\boldsymbol{x}_i^f = \tilde{\mathbf{X}}_i^f \boldsymbol{w}. \tag{29}$$

To track the uncertainty of this estimation, the basis of the ~~error~~ uncertainty subspace is obtained by

$$\tilde{\mathbf{L}}_i^f = \tilde{\mathbf{X}}_i^f \mathbf{T}. \tag{30}$$

The covariance matrix $\mathbf{P}_i^f$ is approximated by

$$\mathbf{P}_i^f = \tilde{\mathbf{L}}_i^f \left(\mathbf{T}^T \mathbf{W}^{-1} \mathbf{T}\right)^{-1} \tilde{\mathbf{L}}_i^{f\,T} + \mathbf{Q}_i \approx \tilde{\mathbf{L}}_i^f \mathbf{A}_i^f \tilde{\mathbf{L}}_i^{f\,T}, \tag{31}$$

where $\mathbf{Q}_i$ is the $N \times N$ covariance matrix of ~~the model error and~~ an additive unbiased noise parameterizing some of the sources of uncertainty in the forecast estimation not already accounted for by the ensemble, while $\mathbf{A}_i^f$ is the $r \times r$ matrix approximating the ~~error~~ uncertainty covariance in the reduced basis expressed by $\tilde{\mathbf{L}}_i^f$. $\mathbf{A}_i^f$ can be computed in many ways, depending on the form of $\mathbf{Q}_i$ and on the chosen hybridization strategy. Here we consider the case of $\mathbf{Q}_i$ being a full rank matrix (e.g., diagonal), thus we suggest

$$\mathbf{A}_i^f = \left(\rho \mathbf{T}^T \mathbf{W}^{-1} \mathbf{T}\right)^{-1} + \left(\tilde{\mathbf{L}}_i^{f\,T} \mathbf{Q}_i^{-1} \tilde{\mathbf{L}}_i^f\right)^{-1}, \tag{32}$$

where ~~$\rho \leq 1$~~ ρ is the forgetting factor, which is added to the equation to introduce inflation ~~(Carrassi et al. (2018))~~(see e.g., Pham et al., 1998 . The last term in equation (32) is one of the novel ~~element~~ elements of the present work. It represents the projection of the ~~model error on the error~~ parametric uncertainty matrix $\mathbf{Q}_i$ on the uncertainty subspace defined by the ensemble. ~~This~~ While the use of $\mathbf{Q}_i$ in equation (31) follows Pham (2001), equation (32) projects $\mathbf{Q}_i$ in a novel non-orthogonal ~~projection,~~ way that is induced by the scalar product defined by ~~the matrix~~ $\mathbf{Q}_i^{-1}$~~, is~~ ~~.~~ . This approach can be interpreted as a form of hybridization ~~which aims to focus~~ that aims at focusing on the effects of the ~~model~~ parametric uncertainty in the ensemble ~~error~~ uncertainty subspace.

~~Since $\mathbf{Q}_i$ can have a very large size, it should be sparse or managed in a decomposed form.~~ $\mathbf{Q}_i$ should be built accordingly to some knowledge of the system, as done, for example, for the background ~~error~~ covariance matrix in variational methods ~~(Bannister (2017a, 2008))~~(Bannister, 2017a, 2008). Since $\mathbf{Q}_i$ can have a very large size, it should be sparse or managed in a decomposed form.

~~A new basis for the error subspace~~ Finally, equation (31) is conveniently rewritten as

$$\mathbf{P}_i^f \approx \mathbf{L}_i^f \mathbf{L}_i^{f^T}, \tag{33}$$

providing a covariance decomposition consistent with the high-order sampling formulation. Here, $\mathbf{L}_i^f$ is the new basis of the uncertainty subspace and it is obtained by

$$\mathbf{L}_i^f = \tilde{\mathbf{L}}_i^f \mathbf{S}_i^f, \tag{34}$$

where $\mathbf{S}_i^f$ is the symmetric square root of $\mathbf{A}_i^f$, i.e.,

$$\mathbf{S}_i^f \mathbf{S}_i^f = \mathbf{A}_i^f, \tag{35}$$

which can be computed by an eigenvalue decomposition of $\mathbf{A}_i^f$. ~~Finally, equation (31) can be rewritten as~~

$$\mathbf{P}_i^f \approx \mathbf{L}_i^f \mathbf{L}_i^{f^T}.$$

~~Writing $\mathbf{P}_i^f$ in this form is important for the consistency of the filter and is useful for the implementation of the localization (see Section 3.2).~~

3.1.5 Forecast high-order sampling

The $N \times (r+1)$ ensemble matrix $\mathbf{X}_i^f = \left(\boldsymbol{x}_i^{f,1}, \ldots, \boldsymbol{x}_i^{f,(r+1)}\right)$, whose columns are the ensemble members, is obtained by

$$\mathbf{X}_i^f = \boldsymbol{x}_i^f \mathbb{1}_{1\times(r+1)} + \mathbf{L}_i^f \mathbf{C}_i^{f^T} \boldsymbol{\Omega}_i^{f^T} \cancel{W}\mathbf{W}^{-\frac{1}{2}}. \tag{36}$$

The sampling procedure is similar to ~~that in~~ the one of Section 3.1.3; therefore, $\mathbf{C}_i^f$ and $\boldsymbol{\Omega}_i^f$ are computed as $\mathbf{C}_i$ and $\boldsymbol{\Omega}_i$ used in equation (6).

Compared to data assimilation schemes with resampling only after analysis, the GHOSH filter uses this sampling phase to produce a better representation of the uncertainty. In fact, it takes into account the ~~model error~~ $\mathbf{Q}_i$ effects in equations (31) and (32), which ~~are~~ would be otherwise neglected. If ~~the model error~~ $\mathbf{Q}_i$ is supposed to be not significant (i.e., $\mathbf{Q}_i = 0$) then the *forecast high-order sampling* phase can be widely simplified and the ensemble members can be obtained by

$$\boldsymbol{x}_i^{f,k} = \boldsymbol{x}_i^f + \frac{1}{\sqrt{\rho}}\left(\tilde{\boldsymbol{x}}_i^{f,k} - \boldsymbol{x}_i^f\right), \tag{37}$$

for $k \in \{1, \ldots, r+1\}$.

3.1.6 Asymptotic computational complexity

Here, the asymptotic computational complexity of the GHOSH filter does not include the cost of applying the model operator $\mathcal{M}_i$ and the observation operator $\mathcal{H}_i$. These operators can change dramatically the total computational cost, in particular in realistic applications, where the model operator can be very computationally expensive.

Further, the *initialization* cost is not taken into account, since it is done only once. This includes the computational cost of solving system (1), which can substantially vary depending on the adopted method. For instance, in the provided python code (see the *Code availability* section), an analytical solution for system (1) is built with a constructive method in $\mathcal{O}\left(r^2\right)$ steps.

Finally, the covariance matrices $\mathbf{Q}_i^{-1}$ and $\mathbf{R}_i^{-1}$ and the scaling matrix $\mathbf{\Lambda}_i^{-1}$ are considered sparse or low rank, such that, for instance, the computational complexity of $\mathbf{Z}_i^T\mathbf{R}_i^{-1}\mathbf{Z}_i$ is the same as $\mathbf{Z}_i^T\mathbf{Z}_i$.

Under these assumptions, the most expensive operations in the *analysis* phase are: the computation of ${\mathbf{A}_i^a}^{-1}$ (which is $\mathcal{O}\left(nr^2\right)$) and of its inverse ($\mathcal{O}\left(r^3\right)$), and the change of basis of equation (23), which is a $\mathcal{O}\left(Nr^2\right)$ operation. Excluding the cost of the observation operator $\mathcal{H}_i$, the analysis computational complexity sums to $\mathcal{O}\left(nr^2+r^3+Nr^2\right)$.

During the *forecast* phase, the computation of $\mathbf{A}_i^f$ and the following change of base (equation (34)) lead to $\mathcal{O}\left(r^3+Nr^2\right)$, without taking into account the cost of the model operator $\mathcal{M}_i$.

Finally, each sampling phase adds $\mathcal{O}\left(r^3+Nr^2\right)$ operations given the eigenvalue decomposition and the matrix products in equation (6).

All together, the asymptotic computational complexity of the GHOSH filter is $\mathcal{O}\left(nr^2+r^3+Nr^2\right)$. This is similar to SEIK and ETKF (see, e.g., Nerger et al., 2012; Tippett et al., 2003), meaning that the filters scale in the same way. However, even if not changing the asymptotic complexity, the GHOSH filter's eigenvalue decomposition and its re-sampling after *forecast* are GHOSH-specific operations that add computational time with respect to other ensemble filters that do not execute such operations. On the other hand, in the vast majority of realistic geoscience applications of ensemble data assimilation, the most demanding computational cost is represented by the model integration scaled by the ensemble size (see e.g., Vetra-Carvalho et al., 2018), making the cost of the other data assimilation operations almost negligible.

3.2 The local GHOSH filter

Localization is a widely used procedure that aims to avoid spurious correlations induced by an overly small ensemble while reducing the degrees of freedom of the system ~~(Janjic et al. (2011))~~(see e.g., Janjic et al., 2011).

A localized version of the GHOSH algorithm is obtained by modifying some of the equations in Sections 3.1.2 and 3.1.4. Three operators $\mathcal{L}_p$, $\mathcal{L}_p^{\mathcal{H}}$ and $\mathcal{D}$ are used to improve readability. The localization operator $\mathcal{L}_p$ takes a global (array or) matrix with N rows and returns its localized counterpart with l rows with respect to the domain point p. The simplest ~~example of the localization operator takes~~ and most common example of localization operator is taking just the variable values of the cells in a small radius around p. $\mathcal{L}_p^{\mathcal{H}}$ works in the same way in the observation space, reducing the number of rows from n to l'. The delocalization operator $\mathcal{D}$ takes a set of local matrices (or arrays), ideally one for each point of the domain, and returns the global counterpart. Building the global matrix by taking the values at the central points of each local matrix is a common example of ~~the~~ delocalization operator.

3.2.1 Local forecast

Instead of equations (31-35), the localized version of the forecast step of the algorithm computes, for each point p of the domain, the local covariance matrix $\mathbf{A}_i^{p,f}$ by

$$\mathbf{A}_i^{p,f} = \left(\rho \mathbf{T}^T \mathbf{W}^{-1} \mathbf{T}\right)^{-1} + \left(\mathcal{L}_p\left(\tilde{\mathbf{L}}_i^f\right)^T \left(\mathbf{Q}_i^p\right)^{-1} \mathcal{L}_p\left(\tilde{\mathbf{L}}_i^f\right)\right)^{-1}, \tag{38}$$

where $\mathbf{Q}_i^p$ is the $l \times l$ ~~model error~~ covariance matrix at the point p of the parametric uncertainty non-dependent on the ensemble. The global basis is built by changing the basis and delocalizing, i.e.,

$$\mathbf{L}_i^f = \mathcal{D}\left(\left\{\mathcal{L}_p\left(\tilde{\mathbf{L}}_i^f\right)\mathbf{S}_i^{p,f}\right\}_p\right), \tag{39}$$

where $\mathbf{S}_i^{p,f}$ is the symmetric square root of $\mathbf{A}_i^{p,f}$. Note that the change in basis induced by $\mathbf{S}_i^{p,f}$ is continuous ~~(see, e.g., Nerger et al. (2012))~~(see, e.g., Nerger et al., 2012); thus, the delocalization operator does not need to include averaging or other smoothing procedures to avoid discontinuities.

3.2.2 Local analysis

The localized version of the analysis step substitutes equations (19) and (20) by calculating, for each point p of the domain, the local covariance matrix $\mathbf{A}_i^{p,a}$, i.e.,

$$\mathbf{A}_i^{p,a\,-1} = \mathbf{T}^T \mathbf{W}^{-1} \mathbf{T} + \mathcal{L}_p^{\mathcal{H}}\left(\mathbf{Z}_i\right)^T \left(\mathbf{R}_i^p\right)^{-1} \mathcal{L}_p^{\mathcal{H}}\left(\mathbf{Z}_i\right), \tag{40}$$

where $\mathbf{R}_i^p$ is the $l' \times l'$ observation ~~error~~ uncertainty covariance matrix at the point p. This matrix can be extracted by $\mathbf{R}_i^{-1}$, but it is convenient to increase the uncertainty at the points far from p ~~(as in Nerger and Gregg (2007)). A reasonable choice is a polynomial weight function that~~ (as in Nerger and Gregg, 2007). A sound choice is to rescale by a function that is equal to 1 in p and goes smoothly to zero out of the localization radius, e.g.,

$$f(d) = \cancel{\frac{d}{d_l}^2 - 1^2,} \begin{cases} \left(\left(\frac{d}{d_l}\right)^2 - 1\right)^2 & \text{if } d < d_l \\ 0 & \text{otherwise,} \end{cases} \tag{41}$$

with d being the distance from p and d_l the localization radius. Equation (41) is the simplest (i.e., lowest order) polynomial function with the required properties, but other options are viable (e.g., the fifth-order piecewise rational function of Gaspari and Cohn, 199

Similarly to equation (~~34~~39) for local forecast, equation (23) becomes

$$\mathbf{L}_i^a = \mathcal{D}\left(\left\{\mathcal{L}_p\left(\tilde{\mathbf{L}}_i^a\right)\mathbf{S}_i^{p,a}\right\}_p\right), \tag{42}$$

where $\mathbf{S}_i^{p,a}$ is the symmetric square root of $\mathbf{A}_i^{p,a}$.

Finally, the analysis state $\boldsymbol{x}_i^a$ in equation (25) is instead estimated by

$$\boldsymbol{x}_i^a = \boldsymbol{x}_i^f + \mathcal{D}\left(\left\{\mathcal{L}_p\left(\tilde{\mathbf{L}}_i^a\right)\mathbf{A}_i^{p,a}\mathcal{L}_p^{\mathcal{H}}\left(\mathbf{Z}_i\right)^T\left(\mathbf{R}_i^p\right)^{-1}\mathcal{L}_p^{\mathcal{H}}\left(\boldsymbol{y}_i - \mathbf{Y}_i\boldsymbol{w}\right)\right\}_p\right). \tag{43}$$

4 Experimental setup

The GHOSH filter has been tested in ~~two sets of numerical experiments of different complexity. First, in Section ??,~~ a very large set of twin ~~experiment~~ experiments based on the Lorenz96 model ~~(Lorenz (1996)) aims~~ (Lorenz, 1996) to evaluate the performance of GHOSH compared with a ~~second order~~ second-order filter (SEIK). The Lorenz96 model, which has a chaotic behaviour that is comparable to the one of fluid dynamics equations at a largely lower computational cost, is a standard choice for testing new data assimilation schemes ~~(e.g., Brajard et al. (2020); Fertig et al. (2007); Gharamti (2018); Grooms and Robinson (2021); Ner~~ ~~).~~

~~The second experiment aims to assess the feasibility of the GHOSH implementation in a realistic geophysical application (Section ??). In particular, the GHOSH filter has been coupled with the Mediterranean biogeochemical model system currently used in Copernicus Marine CMEMS System (OGSTM-BFM: Salon et al. (2019)) to assimilate ocean color satellite data.~~

4.1 ~~Twin Experiment~~

~~The set of twin experiments is based on the Lorenz96 model with random initial conditions.~~(e.g., Brajard et al., 2020; Fertig et al., 2007; Gh

In each experiment, the *truth* trajectory is provided by integrating the model for a ~~fixed~~ certain time interval. Then, at regular time intervals, observations of a subset of the system variables have been extracted ~~by~~ from the *truth*, adding a random error to each of them to represent the observation uncertainty. The same set of observations is then used for two assimilation experiments, one using the SEIK filter ~~(as a second order approximation filter Nerger et al. (2005))~~(as a second-order approximation filter Nerger , and one using the GHOSH filter with fifth approximation order (i.e., $h = 5$, see Section 3). ~~Both the filters are initialized from~~ The SEIK and GHOSH assimilation experiments are initialised with the the same initial condition, chosen randomly around the *truth* initial condition. After an initial spin up of 10 time units, the skill of each filter is ~~computed by averaging~~ evaluated by computing, during the whole simulation, the root mean square error (RMSE) between the filter forecast and the *truth* before the assimilation step. ~~RMSE evolution in time and a synthetic RMSE value for the whole simulation time has~~ The RMSEs on assimilated and non-assimilated variables have been evaluated for each filter. ~~Then, in~~ In order to produce reliable statistics, the same procedure has been repeated ~~56000 times by changing~~ 90000 times varying the *truth* ~~initial conditions, observations,~~ trajectory, the observations, the filters initial conditions and some filters hyper-parameters such as the ensemble size and the forgetting factor.

4.0.1 ~~The Lorenz96 model~~

4.1 The Lorenz96 model

The Lorenz96 model ~~(Lorenz (1996))~~ (Lorenz, 1996) is a dynamical system commonly used as a testing framework in data assimilation. The state vector $\boldsymbol{x} = (x_1, \dots, x_N) \in \mathbb{R}^N$ is evolved according to the system of differential equations given by, for $1 \leq j \leq N$,

$$\frac{dx_j}{dt} = (x_{j+1} - x_{j-2})\,x_{j-1} - x_j + F, \tag{44}$$

with periodic conditions, i.e., $x_{-1} = x_{N-1}$, $x_0 = x_N$ and $x_{N+1} = x_1$. In our implementation, the size of the state ~~array was~~ vector is $N = 62$ and the forcing term ~~was~~ is $F = 8$, which is a common value used to generate chaotic behaviour.
The equations have been implemented in python and ~~solved numerically~~ numerically solved using SciPy's *solve_ivp* routine with its default solver method (i.e., *RK45*).
At time $t = 0$, the state vector has been initialized by adding a small perturbation ($x_N = 0.01$) to the null solution $\boldsymbol{x} = 0$. The ~~twin experiment truth has been obtained by integrating the model for 100 time units . In particular, after a~~ model has been integrated for 20 time units as spin-up, then, the following 1000 time units have been used as historical data to compute the climatological mean and covariance (performing a PCA on 10000 snapshots taken every 0.1 time units). The model has been further integrated for other 20 time units of spin-up, ~~the obtained solution in the time intervals $t \in [20,40]$, $t \in [40,60]$, $t \in [60,80]$, $t \in [80,100]$ has been referred as~~ followed by 80 time units, divided in 4 blocks of 20 time units. The model trajectory of each one of these 4 blocks has been taken as reference in a set of twin experiments and referred as *truth*~~*_20, truth_40, truth_60,*~~ . In the same way, a longer block of 150 time units has been used as *truth*~~*_80*~~ in a set of twin experiments aimed to test long-run performances.

4.1.1 ~~Observations~~

4.2 Observations

Observations are generated for each ~~truth period (*truth_20 , truth_40, truth_60, truth_80*)~~ trajectory by extracting from the state vector ~~,~~ the values of the even indexed variables (i.e., $x_2, x_4, \dots, x_{62}$) every Δt time units and adding to each observed variable a random number sampled from a standard normal distribution (i.e., with variance equal to 1). Tests have been made for $\Delta t \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$.

4.2.1 ~~Filters setup~~

4.3 Filters setup

The same settings are used in both SEIK and GHOSH filters. The ensemble size ~~was~~ is chosen among $7, 15, 31$ and 63 members.~~The initial condition~~

The initial conditions of the ensemble mean is chosen randomly around ~~each~~ a *truth* initial condition, adding a random Gaussian noise with 0~~mean and a standard deviation of 5. Initial conditions of the ensemble members are~~ -mean and same covariance matrix as the climatological covariance of the Lorenz96 model (Section 4.1). The initial ensemble is sampled with each filter's own sampling method (~~second order~~ second-order-exact sampling for SEIK, high-order sampling with $h = 5$ for GHOSH), setting the prior ~~error covariance to a diagonal matrix representing a standard deviation of 5 in each variable~~covariance matrix accordingly to the PCA approximation of the climatolgical covariance (i.e., r principal components are taken into account if the ensemble size is $r + 1$).

Both filters ~~have~~ use the same forgetting factor (equation (32)), chosen among ~~$0.7, 0.75, 0.8, 0.85, 0.9, 0.95$~~ $0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8,$ and 1 (the last value describes a condition without inflation). The GHOSH filter is run without hybridization ($\mathbf{Q}_i = \mathbf{0}$) to keep GHOSH and SEIK as similar as possible.

The order of the GHOSH filter was fixed at $h = 5$, with s (i.e., the number of principal components approximated with order $h = 5$) as big as possible, depending on the ensemble size. ~~The~~ Hence, the values adopted for s are $2, 3, 4$ and 5, respectively for $7, 15, 31$ and 63 ensemble members.

Localization has not been applied, since the number of variables N is relatively small and comparable with the ensemble size.

4.3.1 ~~Experiments design~~

~~100~~

4.4 Experimental design

The twin experiments have been ~~launched~~ performed in two phases, the first one for the 20-time-units-long *truth* trajectories and the second one for the 150-time-units-long trajectory.

In the first phase, for each of the 4 ~~truth periods (~~20-time-units-long *truth*~~_20 , truth_40, truth_60, truth_80),~~ trajectories, 100 twin experiments have been launched for each of the 5 observation frequencies, 4 ensemble sizes and ~~7 forgetting factors~~ 11 forgetting factors (for a total of 88000 twin experiments). Each of the 100 twin experiments has its own set of random observations and initial conditions. ~~Each experiment lasts 20 time units (as in the *truth* simulations), using the first half for spin-up. During the second half of the time interval, the root mean square error (RMSE) for assimilated and non-assimilated variables is separately measured before each assimilation.All the code was written in python and executed in a Linux computer equipped with an Intel(R) Core(TM) i7-11800H @2.30GHz.~~

4.5 ~~Three-dimensional marine biogeochemistry application~~

~~The GHOSH filter has been implemented in the OGSTM-BFM biogeochemical model system of the Mediterranean Sea at $\frac{1}{4}^o$ resolution. The GHOSH implementation features a weekly assimilation of surface chlorophyll from satellite ocean color (as in Teruzzi et al. (2018a)).Consistently, an OGSTM-BFM model run without assimilation (control run)~~ The RMSE score of the first phase experiments has been used ~~as a reference to investigate the GHOSH filter sensitivity to some different parameterizations.~~

4.4.1 ~~The OGSTM-BFM model~~

~~The biogeochemical model used in this study is the coupled transport-biogeochemical model OGSTM-BFM, which is the biogeochemical component of the Mediterranean Copernicus Marine Service (Salon et al. (2019)). The OGSTM transport model is a modified version of the OPA 8.1 transport model (Foujols et al. (2000)), which resolves the advection, the diffusion and the sinking terms of the biogeochemical variables (Lazzari et al. (2010)). The biogeochemical BFM model describes the biogeochemical cycles of carbon, nitrogen, phosphorus and silicon through the dissolved inorganic, particulate living organic and non-living organic compartments, and includes nine plankton functional types, as well as dissolved oxygen, pH and the carbonate system. The OGSTM-BFM model has been applied to study the chlorophyll and primary production (Lazzari et al. (2012); Teruzzi et al. (2018a)), nutrient cycles (Lazzari et al. (2016)), alkalinity, the carbon cycle and carbon fluxes (Cossarini et al. (2015); Melaku Canu et al. (2015)) and oxygen dynamics (Di Biagio et al. (2022)).~~

~~The OGSTM-BFM is forced by the output of an ocean general circulation model (i.e., NEMO, from the Copernicus Marine Mediterranean system in the present application; Salon et al. (2019)) that provides fields of physical quantities (i.e., velocity, temperature, salinity, diffusivity, and solar radiation and wind at the surface from a physical reanalysis) driving the transport processes, the kinetic rates of chemical reactions, and energy and matter exchanges at the air-sea interface.~~

~~The present setup (terrestrial and atmospheric inputs, boundary conditions) is described in Salon et al. (2019).~~

4.4.1 ~~Observations~~

~~The assimilated observations are satellite chlorophyll maps obtained from the CMEMS multisensor productOCEANCOLOUR_MED_CHL (Volpe et al. (2017)). The original data at 1 km resolution were checked for spikes, excluding values exceeding by two standard deviations the climatological mean and remapped on the model grid resolution.~~

~~Additionally, in situ climatological data for nitrate and phosphate, to be used for validation, is computed from a dataset that integrates the in situ EMODnet data collections (Buga et al. (2018)) and the datasets listed in Lazzari et al. (2016). The climatology is calculated for the 0-30 m layer on 9 sub-basins (Salon et al., 2019) and for 4 different seasons (January-March; April-June; July-September; October-December).~~

4.4.1 ~~GHOSH setup~~

~~**Test name Order** (h) **Forgetting factor** (ρ) **Forecast frequency** *ctrl T_h2_rho1_d* 2 *T_h3_rho1_d* 3 *T_h5_rho1_d* 5 *T_h2_rho0.9_d* 2 *T_h3_rho0.9_d* 3 *T_h5_rho0.9_d* 5 *T_h2_rho0.8_d* 2 *T_h3_rho0.8_d* 3 *T_h5_rho0.8_d* 5 *T_h2_rho1_w* 2 *T_h3_rho1_w* 3 *T_h5_rho1_w* 5 *T_h2_rho0.7_w* 2 *T_h3_rho0.7_w* 3 *T_h5_rho0.7_w* 5 *T_h2_rho0.5_w* 2 *T_h3_rho0.5_w* 3 *T_h5_rho0.5_w* 5 Experimental setup.~~

~~In the 3D simulations, GHOSH assimilated surface chlorophyll and update the whole state vector (i.e., 51 state variables) of the OGSTM-BFM model by using an ensemble of 16 members (i.e., an error subspace of dimension $r = 15$). Tests were carried out~~ to identify the best forgetting factor for each filter in each configuration of observation frequency and ensemble size. Then, in the second phase, the 150-time-units-long *truth* trajectory has been employed in a set of 100 twin experiments for each of ~~three different maximum polynomial orders h (and corresponding numbers s of principal components approximated with such orders):~~

- ~~$h = 2$, i.e., a second order of approximation on the whole error subspace ($s = 15$);~~
- ~~$h = 3$, i.e., third order of approximation on the 8 most relevant components ($s = 8$) and second order on the remaining 7 principal directions;~~
- ~~$h = 5$, i.e., fifth order of approximation on the top 3 principal components ($s = 3$), and second order on the rest (12 components).~~

~~The ensemble weights and $\mathbf{\Omega}^h$ were computed (usually during the *initialization* phase) by solving system (1). In the case with $h = 2$, the solution is easily obtained starting from the constant vector v of the square root of the weights and adding orthonormal column vectors in order to form an orthogonal matrix (see equation (A10) in Appendix A).In the case with $h = 3$, a solution with constant weights ($v_m = \frac{1}{4}$ for each $m \in \{1, \ldots, 16\}$) was chosen, namely,~~

$$\mathbf{\Omega}^{h=3} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}^T .$$

~~In the case with $h = 5$, the chosen solution was~~

$$\boldsymbol{v} = \Big(\begin{matrix} \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{1}{2\sqrt{5}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \cdots \end{matrix}$$

$$\begin{matrix} \cdots & \frac{1}{2\sqrt{5}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} \end{matrix} \Big)^T$$

~~and~~

$$\boldsymbol{\Omega}^{h=5} = \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \cdots \\ 0 & 0 & 0 & \frac{1}{\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & \cdots \\ 0 & 0 & 0 & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \cdots \end{pmatrix.}$$

$$\begin{matrix.} \cdots & -\frac{1}{2} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} \\ \cdots & 0 & \frac{1}{\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} \\ \cdots & 0 & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} \end{pmatrix}^T.$$

~~The scaling matrix $\mathbf{A}_i$ (equation (7)) is a diagonal matrix representing the error variance (i.e., the diagonal of $\mathbf{P}_i^a$), and it is computed as the squared Euclidean norms of the rows of $\mathbf{L}_i^a$. The $\mathbf{A}_i$ matrix has been chosen valuing simplicity and generality over skill performance. The implemented scaling matrix is probably a suitable choice for many systems, since it implies that the PCA measures the Pearson correlation and that it is not affected by the different amplitudes of the variable standard deviations.~~

~~The simulations cover the period 1 January 2013 - 1 January 2014, with daily or weekly forecast frequency (**??**).The ensemble covariance was inflated as in equation (32). Tests were conducted with different forgetting factor values, namely, $\rho \in \{1, 0.9, 0.8\}$ for daily forecasts and $\rho \in \{1, 0.7\}$ for weekly forecasts.~~

~~The localized version of the filter used cylindrical and circular patches for $\mathcal{L}_p$ and $\mathcal{L}_p^{\mathcal{H}}$, using a 90 km radius (similarly to values adopted by other EnKF applications, e.g., Hu et al. (2012); Ciavatta et al. (2016)) centred at the point p. The delocalization operator $\mathcal{D}$ builds the global array by taking the values in the central cell of each localized array (Section 3.2). However, since the change in basis in the GHOSH algorithm was continuous (as noted in Section 3.2.1), more demanding localization strategies, such as averaging over the weighted values of each cell local array, would not introduce relevant improvements.~~

~~Given the non-Gaussianity of biogeochemical tracer concentrations, the GHOSH filter was applied to the log-transformed biogeochemical variables (Ciavatta et al. (2016); Pradhan et al. (2019); Song et al. (2016); Ford (2020)). To avoid instabilities due to large negative values of logarithms of very small concentrations, a cut-off was applied to values lower than the threshold of 10^{-4}. This solution can be seen as a localization that dynamically removes unwanted correlations, which can typically occur at deeper layers, where some tracers (such as, for example, phytoplankton biomass and chlorophyll) tend to very low concentrations.~~

~~The mean initial condition $\boldsymbol{x}_0^f$ and the starting ensemble $\mathbf{X}_0^f$ were produced based on a 10-year hindcast simulation of the OGSTM-BFM model without assimilation. For each year, 31 days centred on the 1st of January were log-transformed, after removing values smaller than 10^{-4}, for a total of 310 simulated multivariate states. The mean of the multivariate states was used as $\boldsymbol{x}_0^f$, while a PCA of the ensemble of normalized states provided the 15 (i.e., the dimension of the error subspace r) most relevant components, which are used to build the base matrix $\mathbf{L}_0^f$. Then, the GHOSH sampling method has been used to extract the 16 starting ensemble members, as in Section 3.1.5.~~

~~The local model error covariance matrix $\mathbf{Q}_i^p$ has been assumed diagonal and invariant in time. The variances in $\mathbf{Q}_i^p$ are assumed inversely proportional to the height of the cells. Given that the resolution of the present model setup is not high enough to adequately resolve coastal processes, assimilating observations in coastal areas could lead to instabilities. Therefore, the model error variance in the coastal areas was reduced by applying a 4th-degree polynomial smoothing function and setting the covariance values at zero in the grid cells closest to the coast. More precisely, the uncertainty starts to decrease where sea depth is 200 metres and reaches zero in areas shallower than 50 metres. The local observation error covariance matrix $\mathbf{R}_i^p$ was assumed diagonal and constant in time. Using the same 4th-degree polynomial smoothing applied for $\mathbf{Q}_i^p$, the inverse of the $\mathbf{R}_i^p$ variances was smoothed to zero in the coastal areas (i.e., the $\mathbf{R}_i^p$ variances increase in the coastal areas oppositely to $\mathbf{Q}_i^p$).Moreover, the matrix $(\mathbf{R}_i^p)^{-1}$ was weighted with the polynomial function in equation (41) to apply the local version of the filter.~~

~~A tuning of the model and observation error variance may be necessary to produce an effective assimilation when system errors are not fully known and parametrized (e.g., Cossarini et al. (2019); Ford and Barciela (2017); Janjić et al. (2018); Mattern et al. (201). The actual values of model and observations error variances were obtained testing different values and choosing those that provided effective assimilation increments without introducing biogeochemical variables degradation.~~

~~In total, 18 assimilation experiments were run, in addition to the control run, as summarized in Tab. ??. Overall, the experimental design features a number of different forgetting factors ρ, 3 values of the approximation order h (from 2 to~~ the 5 ~~) and 2 different time intervals between two *forecast* phases (1 and 7 days) to test the sensitivity of the assimilation scheme.~~ observation frequencies and 4 ensemble sizes (for a total of 2000 longer twin experiments), using for each filter its best forgetting factor in each configuration.

~~The tests were carried out in the GALILEO cluster at the Italian HPC centre CINECA, mounting 2 Intel Broadwell CPU (~~All the code was written in python and executed in a Linux computer equipped with an Intel(R) ~~Xeon~~Core(TM) ~~E5-2697 v4~~ i7-11800H @~~2.3GHz, 18 cores each) per each node. The parallel Fortran 2003 implementation of the GHOSH algorithm ran on 46 nodes (corresponding to 1656 cores)~~2.30GHz.

5 Results

5.1 ~~Twin ExperimentFigure 3 (a and b) reports the~~ 20-time-units-long experiments

Figure 3 reports an example of evolution of the first ~~2~~ two variables (one assimilated and one non-assimilated) of the Lorenz96 model for the two filters (GHOSH and SEIK). The selected simulation out of ~~56000~~ 88000 experiments shows that GHOSH filter evolution is closer to truth evolution (black line in Fig. 3) than SEIK. As expected, the uncertainty of variable 2 (shaded areas) of both filters is lower than in variable 1 because of the assimilation of observations in even variables. ~~The RMSE computed at every time step over the two sets of variables (assimilated and non-assimilated, Fig. 3) shows that GHOSH outperforms SEIK after the spin-up time.~~

In order to be statistically reliable, the RMSE of the ~~56000~~ 88000 experiments under different conditions have been computed and reported in aggregated form in Fig. 4. Each small square reports the RMSE value computed over a set of 400 experiments,

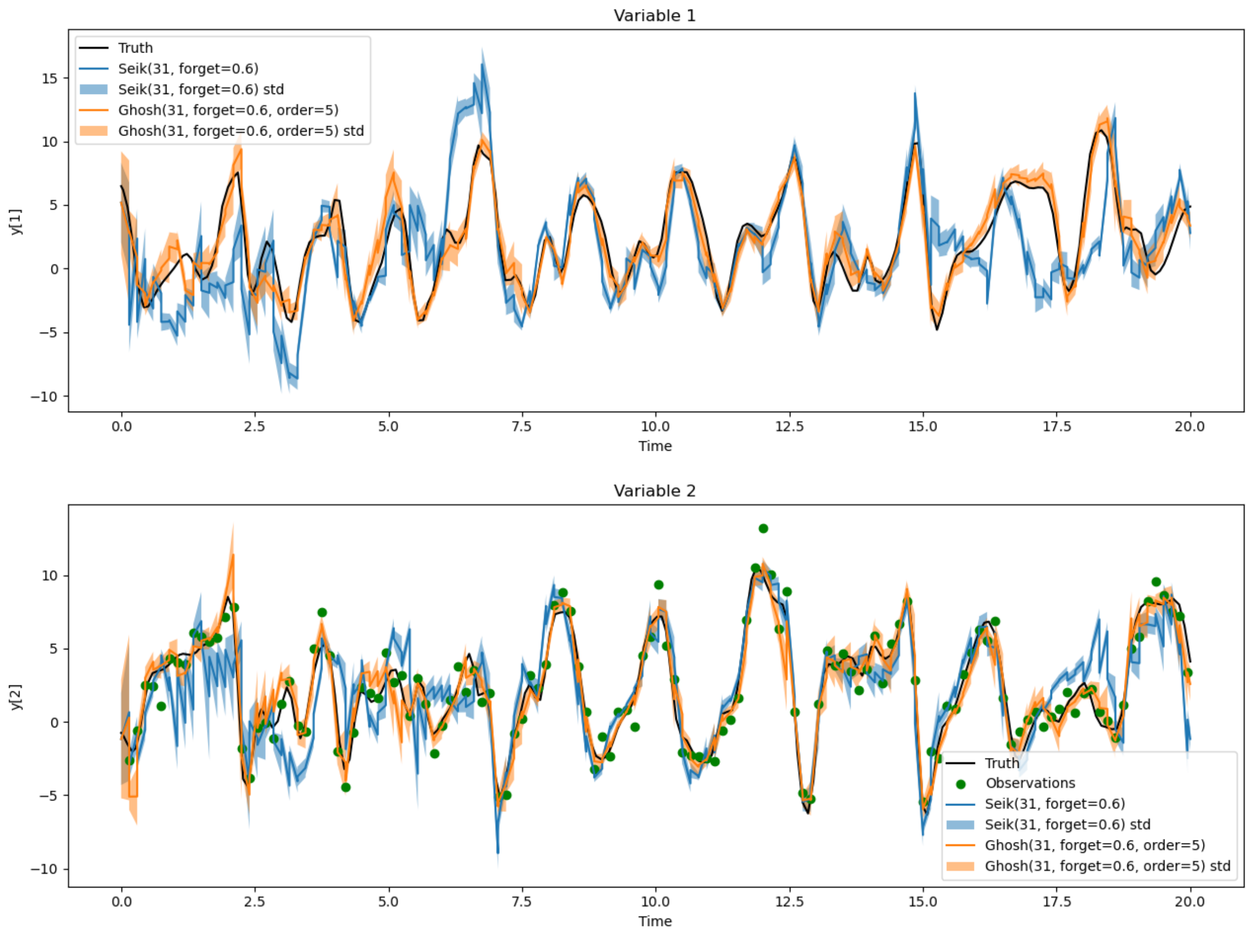


Figure 3. Results from one twin experiment example: ~~*truth* (black line) is *truth_60*,~~ time between observation (green dots) is 0.15, forgetting factor is ~~0.8~~0.6, ensemble size is 31. Shady areas around blue (SEIK) and ~~organge~~ orange (GHOSH) lines represent the ensemble standard deviation, *truth* is the black line. ~~From top to bottom~~Top panel: time evolution of the first variable (non-assimilated); bottom panel: time evolution of the second variable (assimilated)~~; RMSE along time based on all the assimilated variables; RMSE along time based on all the non-assimilated variables~~.

which includes, for each of the four available truths, 100 experiments with different random observations and initial conditions. Aggregating 400 experiments together reduces the outcome variability by a factor 20 (i.e., $\sqrt{400}$) with respect to a single twin experiment, removing stocasticity concerns in the analysis of the twin experiment results.

The first ~~2 lines of color maps~~ two rows of colour-maps refer to the SEIK filter ~~observed and unobserved~~ assimilated and non-assimilated variables. In each ~~color map~~colour-map, the time interval between assimilated observations increases from left to right, while the different forgetting factor values decrease from top to bottom. The first line in each ~~color map~~colour-map, labelled "*best*", represents the best result~~obtained between all the~~, in terms of lowest RMSE, obtained among the set of tested

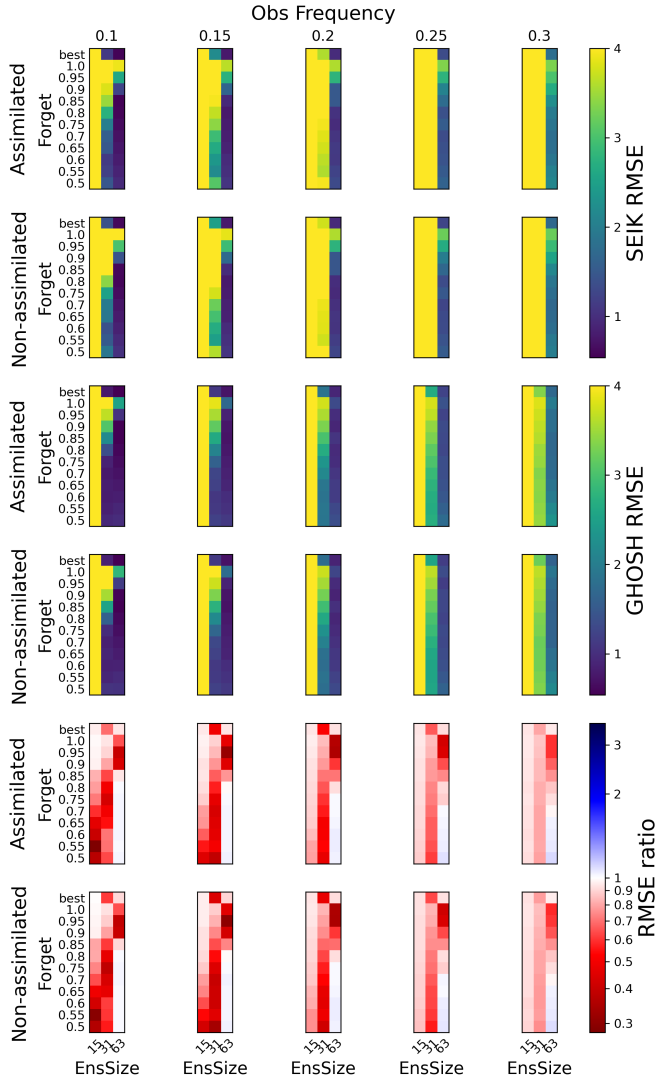


Figure 4. Result summary of 20-time-units-long twin experiments: each square in the ~~color maps~~ colour-maps represents the aggregated results of ~~400~~ 400 twin experiments, changing *truth*, observations and initial conditions. The results are summarized with ~~color maps~~ colour-maps aggregated in six different rows, from top to bottom: SEIK RMSE of assimilated variables, SEIK RMSE of non-assimilated variables, GHOSH RMSE of assimilated variables, GHOSH RMSE of non-assimilated variables, the ratio of GHOSH RMSE over SEIK RMSE of assimilated variables, the ratio of GHOSH RMSE over SEIK RMSE of non-assimilated variables (red color implies that GHOSH is better than SEIK). Each column of ~~color maps~~ colour-maps has different observation frequency, with the numbers on the top indicating the time elapsed between each observation/assimilation. Each ~~color map~~ colour-map shows different forgetting factors (Forget, along the y axis) and ensemble sizes (EnsSize, along the x axis).

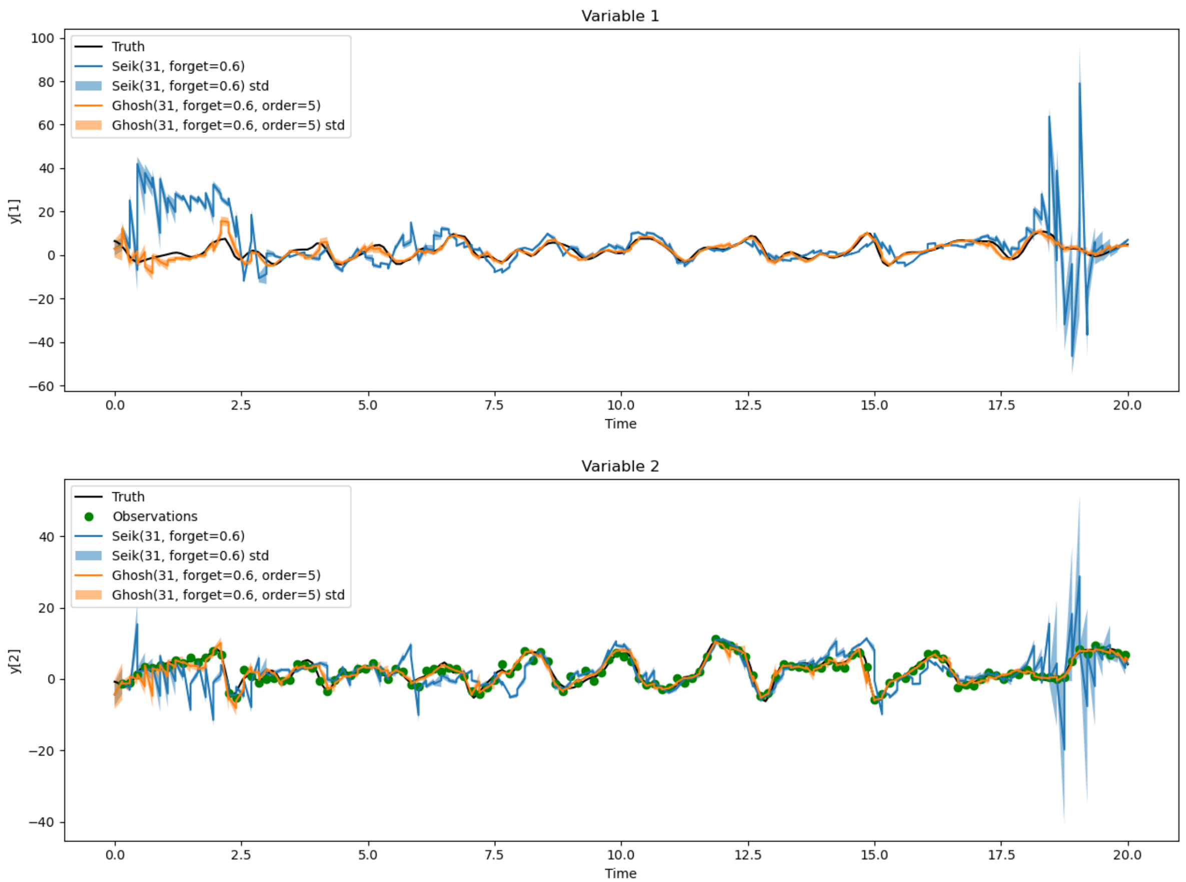


Figure 5. Example of a single twin experiment, with SEIK (blue line) diverging behaviour: time between observations (green dots) is 0.15, forgetting factor is 0.6, ensemble size is 31. Values over time of a non-assimilated (top) and an assimilated (bottom) variable.

forgetting factors. The color scale is capped to an RMSE of 4, the climatological standard deviation of the model, that represents the threshold under which the filter performs a successful assimilation.

The GHOSH experiments are summarized in the same way in the 3rd and 4th ~~lines~~ rows of Fig. 4, while the last ~~2 lines~~ two rows show the ratio between GHOSH and SEIK RMSE values, with red color indicating RMSE reduction of GHOSH with respect to SEIK.

The column corresponding to the smallest ensemble size (i.e., 7 members) is not shown, since in this case SEIK and GHOSH behave very similarly, ~~showing~~ with very poor performances, due to lack of capability of describing the system complexity with an overly small ensemble size.~~In these conditions the GHOSH showed a very small improvement with respect to SEIK (0-3%)~~

The yellow 15-members columns also show poor performances in all configurations but, looking at the last two rows of Fig. 4, the GHOSH RMSE is significantly better than SEIK RMSE (red squares), specially for low values of forgetting factor. In fact, while the GHOSH filter keeps the RMSE around the climatological values, the SEIK filter is more prone to numerical divergence landing in trajectory far from the *truth* (see Fig. 5 for an example of diverging behaviour).

Considering the cases with 31 and 63 ensemble members, the GHOSH filter produce a successful assimilation in almost all configurations (except two yellow squares representing experiments without inflation in the case of observation interval smaller than 0.15 time units). The SEIK filter instead shows a more limited capacity of producing a successful assimilation, as proved by a larger number of yellow squares in Fig 4. Remarkably, the GHOSH filter with 31 ensemble members improves the state estimation error compared to the climatological standard deviation of the model for every observation interval. The same is not true for the SEIK filter, which achieves convergence (i.e., RMSE lower than the climatological standard deviation, at least in its *best* forgetting factor configuration) only for observation intervals shorter than 0.2 time units.

Looking at the last ~~2~~ two rows of Fig. 4, the GHOSH filter outperforms the SEIK filter in most conditions, up to a ~~three~~ 3 times RMSE reduction. Furthermore GHOSH proves to be always at least as good as SEIK. ~~The~~ In the range of explored forgetting factors, the largest improvements (dark red, RMSE ratio ~~0.31~~0.30) occur when: i) inflation is low (forgetting factor near to 1); ii) large amount of information is injected into the assimilation scheme (i.e., high frequency observations, left side of Fig. 4); iii) the filter can take into account a high dimensional ~~error~~ uncertainty subspace (i.e., the ensemble size is large).

On the other hand, the comparison of each filter tuned with its best forgetting factor (the "*best*" top line in each ~~color map~~colour-map) clearly shows that the GHOSH filter has considerably better performances than the SEIK filter (up to a RMSE ratio of ~~0.43~~0.44) with a moderate number of ensemble members (31). In case of maximum ensemble size the improvement is still ~~relevant~~ significant but less intense (up to ~~0.9~~ 0.90 RMSE ratio).~~The improvement is comparably small, up to 7%, in the case of small ensemble size (15 members).~~

~~In general, it is~~ The first four lines of Fig. 4 makes evident that the GHOSH filter, compared to SEIK, is less dependent on inflation, reaching its best performance with a forgetting factor closer to 1.

The skill of both the filters is closely related with the observations frequency: the shorter the time between observations, the higher the accuracy.

Finally, as expected, for both the filters the skill is better on assimilated variables than non-assimilated ones but, quite interestingly, the RMSE ~~reductions of GHOSH with respect to SEIK are~~ ratio shows that the GHOSH advantage over SEIK is slightly larger in ~~the unobserved variables~~ all the non-assimilated variables compared to the assimilated ones with the same settings (up to ~~0.06~~ 0.08 RMSE ratio improvement ~~, from 0.67~~ in the *best* forgetting factor configuration, from 0.69 for assimilated to 0.61 for non-assimilated variables).

~~From the computational point of view, the whole experiment set needed around 7 computational hours. Since the computational cost is dominated by the integration of the model, the time to solution of the two tested filters should be nearly the same, however, unexpectedly, the average time to solution (averaged every~~

5.2 150-time-units-long experiments

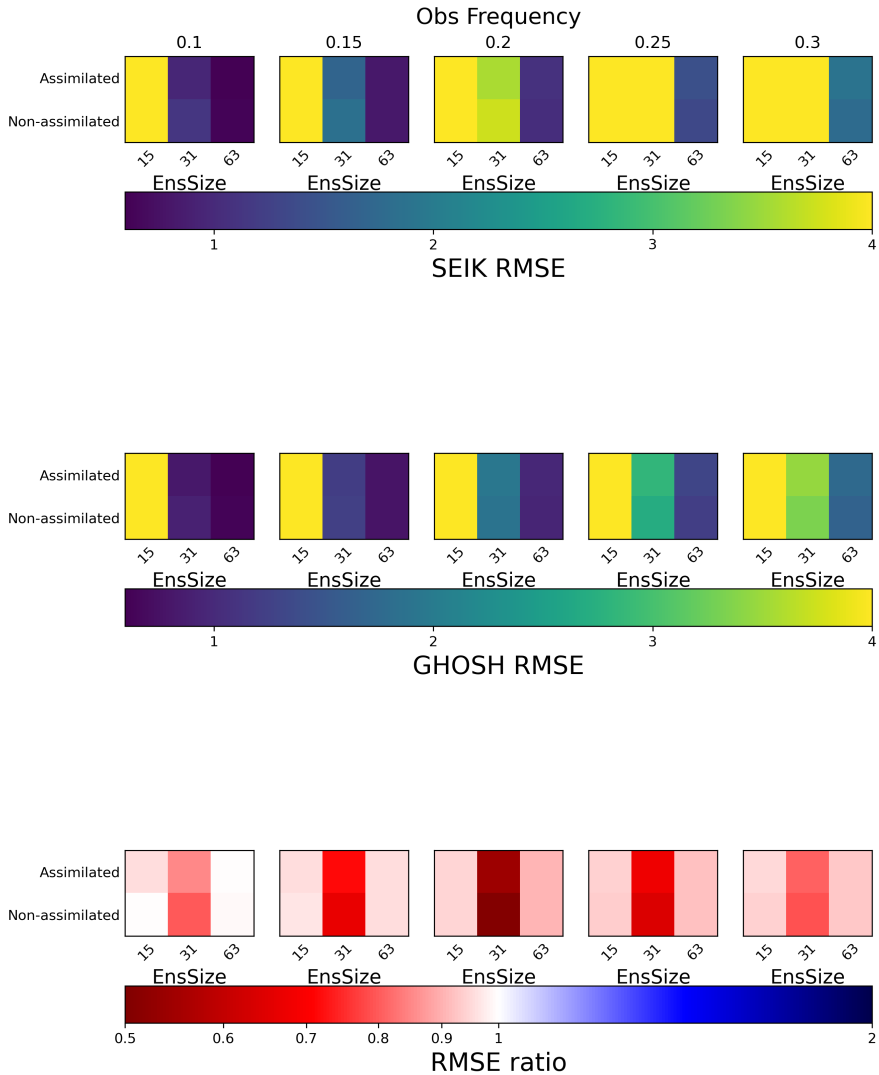


Figure 6. Result summary of 150-time-units-long twin experiments: each square in the colour-maps represents the aggregated results of 100 twin experiments, changing observations and initial conditions. The results are summarized with colour-maps aggregated in 3 different rows, from top to bottom: SEIK RMSE, GHOSH RMSE, ratio of GHOSH RMSE over SEIK RMSE (red color implies that GHOSH is better than SEIK). Each column of colour-maps has different observation frequency, with the numbers on the top indicating the time elapsed between each observation/assimilation. Each colour-map shows, for assimilated and non-assimilated variables, different ensemble sizes (EnsSize, along the x axis).

~~Example of a single twin experiment, with SEIK (blue line) diverging behaviour: *truth* (black line) is *truth_60*, time between observations (green dots) is 0.15, forgetting factor is 0.75, ensemble size is 31. Values over time of a non-assimilated (top) and an assimilated (bottom) variable.~~

Similarly to the case of the 20-time-units-long twin experiments, the results of the 150-time-units-long tests are summarized in Fig. 6. Each small square reports the RMSE value computed over a set of 100 ~~twin experiments) of the SEIK filter sometimes is longer than the time to solution of the GHOSH filter, up to twice the time, depending on some settings and random factors. This difference in computational time is more evident and occurs more often when the forgetting factor is lower (i.e., when the inflation is more pronounced). A possible explanation comes from the model integration algorithm used in the SciPy's *solve_ivp* routine: the integration time step is not constant and depends on the stability of the equations (stiffness). Which means that, if the filter is less accurate, the output can diverge from the expected trajectory in areas where the equations are stiffer, and the integration needs more time steps to converge rising the time to solution. This is supported by the fact that the SEIK filter can show a diverging behaviour when the forgetting factor is too low (Fig. 5 shows an example). Remarkably, we did not observe any diverging behaviour in the GHOSH filter runs.~~

5.3 ~~Three-dimensional realistic complexity marine biogeochemistry application~~

~~**Test name Order Chlorophyll Phosphate Nitrate** (h) **RMSD** (mg/m^3) **RMSD** $(mmol/m^3)$ **RMSD** $(mmol/m^3)$ *ctrl* no 0.084 0.036 0.99 *T_h2_rho1_d* 2 0.068 0.035 0.93 *T_h3_rho1_d* 3 0.068 0.035 0.85 *T_h5_rho1_d* 5 0.068 0.035 0.83 *T_h2_rho0.9_d* 2 0.059 0.033 1.50 *T_h3_rho0.9_d* 3 0.060 0.033 1.14**_T_h5_rho0.9_d_ 5 0.059 0.033 0.83** *T_h2_rho0.8_d* 2 0.059 0.033 1.96 *T_h3_rho0.8_d* 3 0.058 0.032 2.15 *T_h5_rho0.8_d* 5 0.058 0.032 1.76 *T_h2_rho1_w* 2 0.071 0.035 0.94 *T_h3_rho1_w* 3 0.071 0.035 0.89 *T_h5_rho1_w* 5 0.071 0.035 0.91 *T_h2_rho0.7_w* 2 0.063 0.034 0.94 *T_h3_rho0.7_w* 3 0.063 0.035 0.93 *T_h5_rho0.7_w* 5 0.063 0.034 0.91 *T_h2_rho0.5_w* 2 0.060 0.033 1.35 *T_h3_rho0.5_w* 3 0.060 0.032 1.02 *T_h5_rho0.5_w*~~ experiments with different random observations and initial conditions. In the top panel, the 5 ~~0.060 0.033 0.91 Results: summary of the root mean square difference (RMSD) indicators for each test. Best GHOSH configuration in bold.~~

~~The GHOSH capability to reproduce expected dynamics in a complex and real geophysical application has been evaluated by looking at: the skill performance of forecast of the assimilated variable (chlorophyll) and of the non-assimilated variables (nutrients) and the quality of the system state uncertainty estimation.~~

~~Furthermore, the GHOSH performances have been evaluated varying the order of approximation h. Table **??** shows that the performance of the assimilated variable is substantially unaffected by the GHOSH polynomial order, whilst h demonstrates a relevant impact on a~~ colour-maps correspond to different time intervals between observations, increasing from left to right. Each colour-map shows, for assimilated and non-assimilated ~~variable, namely the nitrate~~variables and for different ensemble sizes, the RMSE of the SEIK filter. The RMSE of the GHOSH filter is represented in the same way in the middle panel, while the ratio of the GHOSH RMSE over SEIK RMSE is shown in the bottom panel.

The longer experiments confirms the result obtained by the shorter ones, with GHOSH performing better the SEIK in every configuration. In particular, ~~when the order h is 5, nitrate root mean square difference (RMSD) between GHOSH and independent observations improves up to 81% with respect to order $h=2$ (scored in the set of tests with $\rho = 0.9$ and daily forecasts). Every $h=5$ test has a smaller RMSD than every corresponding $h=2$ and, in general, data clearly show that increasing h implies an improvement in RMSD. In fact, the nitrate's RMSD values, centred and normalized over their standard deviation, present a strong negative correlation with h (Pearson's correlation coefficient is -0.80).~~

~~Among the 18 assimilation experiments, *T_h5_rho0.9_d* proved to have the best GHOSH configuration (Tab. **??**). The results from this run are discussed in detail in the following sections, through a qualitative and quantitative comparison with the control run (i.e., a hindcast run with no assimilation).~~

5.2.1 ~~Surface chlorophyll~~

~~Surface chlorophyll (mg/m^3) comparison: satellite, control run and data assimilation (before and after assimilation) on 19.03.2013.~~

~~Surface chlorophyll RMSD (mg/m^3) and RRMSD over time.~~

~~The surface chlorophyll estimated by the free run and~~ the ~~GHOSH filter has been compared with satellite data from January to December 2013. A detailed analysis of a specific relevant event (winter bloom, 19 March 2013) is shown in Fig. **??**). The RMSD time series is shown in Fig. **??**.~~

~~The March map describes a typical winter bloom in the western Mediterranean, with the northernmost part still under the convective phase (i.e., blue hole surrounded by high chlorophyll concentration patches, Fig. **??**a). Additionally, high chlorophyll levels are present along the Algerian Current as a result of the active mesoscale dynamics, whereas the eastern Mediterranean Sea has a much lower chlorophyll concentration due to the lower intensity of the winter vertical mixing (Fig. **??**a). While the free run overestimates the surface chlorophyll in the whole Mediterranean Sea (Fig. **??**b), the pre-assimilation picture of GHOSH run shows a more consistent condition (Fig. **??**c), which is further improved by the assimilation (Fig. **??**d)~~ improvement is maximum (0.51 RMSE ratio) in the case of 31 ensemble members. Both filters show poor performance at 15 or less ensemble members, and both of them achieve a good convergence at the maximum ensemble size. In the case of 31 ensemble members, the behaviour of GHOSH and SEIK differs when the observation interval is 0.25 time units or longer, with GHOSH achieving a successful assimilation (blue-green color) while SEIK performs worse than the climatological error (yellow color).

~~Time series of RMSD computed between satellite and model before the assimilation are plotted in Fig. **??**. Since the amount of surface chlorophyll can vary considerably during the year, the relative root mean square difference (RRMSD) is also provided (i.e., RMSD over concentration), as a normalized proxy of the skill.Figure **??** shows that the GHOSH assimilation produces an improvement in the assimilated variable (i.e., surface chlorophyll) with respect to the control run. The RMSD of the control run follows a seasonal cycle, with maximum values in late winter-early spring, corresponding to the bloom period, and minimum values during summer when the surface layer is mostly nutrient depleted and chlorophyll is at its lowest level.~~

The ~~winter-spring RMSD is significantly reduced by the GHOSH filter whereas the RMSD reduction results fairly small during summer, due to a general smaller amount of surface chlorophyll in that season. In fact, when the RRMSD is considered, the summer improvement due to assimilation is more detectable. As a proxy of the overall performance of each 1-year experiment, the RMSD between the simulation and observation, averaged in space and time, is summarized in Tab. **??**: the chlorophyll RMSD for *ctrl* (control run) is 0.084, while it is 0.055 for the GHOSH test, resulting in a 53% improvement.~~

5.2.1 ~~Nutrients~~

~~The effect of DA on~~ long-run experiments also confirm that the improvement is slightly larger for non-assimilated ~~variables has been assessed using independent observations to validate phosphate and nitrate (Tab.??), which can both act as limiting nutrient for phytoplankton growth in the Mediterranean Sea (Lazzari et al. (2016)) .~~

~~Hovmöller diagrams of nutrients (phosphate and nitrate, $mmol/m^3$): comparison between the control run *ctrl* (up) and data assimilation test *T_h5_rho0.9_d* (bottom)~~ variables than for assimilated ones, scoring a better RMSE ratio (maximum ratio improvement of 0.05) in almost all configurations.

~~The Hovmöller diagrams of phosphate and nitrate in~~

5.3 Computational cost

From the computational point of view, the whole experiment set needed around 9 computational hours. SEIK and GHOSH schemes used 12% and 16% of the total time respectively, while the ~~DA and the control runs show the effects of the assimilation on the two nutrient compartments (Fig. ??). Both nitrate and phosphate show lower concentrations in the 150–400 m layer from April in the simulation with satellite assimilation, with highest impact at the beginning of April. At the surface (0–100 m layer), the assimilation slightly enhances the phosphate concentrations, particularly between April and July, while it reduces the nitrate concentrations~~rest was committed to model integration. The GHOSH filter executes more operations than SEIK (e.g., an eigenvalue decomposition) resulting in more computational time even if the asymptotic computational complexity of the two methods is the same (Section 3.1.6). However, ~~these effects do not degrade the typical vertical distribution of nutrients in the Mediterranean Sea, and the seasonal sequence of late-spring/winter mixed conditions and of summer stratification is preserved.Interestingly, the Hovmöller diagrams of Fig. ?? show that the DA is able to correct the profiles of the nutrients down to 300–400 m by changing the shape of the nutricline (i.e. , deepening and creating a gentler shape) during the month of the deep chlorophyll maximum formation (from April to June) and by re-setting the nutricline to a shallower level after summer.~~

5.3.1 ~~Uncertainty estimation~~

~~In~~ even in the case of a relatively simple model like the Lorenz96, the time to solution is dominated by the model integration and the difference between GHOSH and SEIK only accounts for 4% of the total time. Unexpectedly, the model integration time (averaged every 100 twin experiments) when applying the SEIK filter sometimes is longer than the GHOSH filter case, up to twice the time, depending on some settings and random factors. This difference in computational time is more evident and occurs more often when the forgetting factor is lower (i.e., ~~a weighted ensemble is used to estimate the uncertainty of the forecast and analysis states (equation (12)). The weighted ensemble standard deviation is used to quantify the uncertainty, whereas a comparison of the standard deviations before and after assimilation is calculated to evaluate the uncertainty reduction produced by the assimilation.Moreover, the ensemble correlation matrix is analyzed to investigate the relationships between variables. The standard deviation and correlation are shown for the assimilated variable and one of the nutrients during a bloom event in winter.~~

~~During the winter event (March), the spatial patterns of chlorophyll and phosphate standard deviations at the surface (second row of Fig. ??) reflect those of their concentration maps (first row of Fig. ??): areas with high surface concentrations in the western Mediterranean Sea are associated with the highest values of the ensemble standard deviation given the high variability of the surface bloom dynamics. The map also shows an area with low chlorophyll and high phosphate concentration in the northwestern Mediterranean where winter convection is still active and standard deviation is low. As an effect of the assimilation, the areas with the highest standard deviation for both variables in the western Mediterranean Sea decrease, proving the effectiveness of the surface assimilation in constraining the surface phytoplankton bloom dynamics. Moreover, a relevant reduction in the standard deviation, in relative terms, is also observed in the less productive eastern Mediterranean Sea. On average, the assimilation reduces the standard deviation at the surface by 33% and 22% for chlorophyll and phosphate, respectively.A few spots of high standard deviations remain only in some coastal areas in the western Mediterranean Sea for both variables. The system has been tuned to focus on open sea and the assimilation in shallow waters is less effective due to imposed higher observation errors in the coastal areas (see Section ?? for details).~~

~~Maps of chlorophyll and phosphate (a and b) and maps of their uncertainty (ensemble standard deviation) before assimilation (c and d) and after assimilation (e and f) for the day 19.03.2013.~~

5.3.1 ~~Three-dimensional Feasibility and computational cost~~

~~The feasibility of the GHOSH scheme in a realistic 3D operational system has been confirmed by the experiments: each 1-year assimilation test, consisting of 16 ensemble members, needed approximately 4 hours to be computed in the CINECA's GALILEO super computer, running in parallel on 46 computational nodes (1656 cores).The computational cost of the GHOSH filter itself is negligible compared with the model integration timeand the I/O operations, being around 1% of the total computational time.~~ when the inflation is more pronounced). The reason can be understood by looking at Fig. 5: in this experiment with forgetting factor equal to 0.6, the SEIK filter presents an unstable and diverging behaviour. When it happens, the SciPy's *solve_ivp* integrating routine reduces the time step size for granting the required accuracy, which comes with longer integration time.

6 Discussion

The name "Gauss-Hermite high-order sampling hybrid filter" was chosen to indicate the two main features of this novel filter. First, it is a hybrid filter, in the sense that the ~~error~~ uncertainty covariance matrix is obtained by averaging the ensemble covariance and a ~~background error covariance matrix~~ parametric covariance matrix non-depending by the ensemble, as described in~~Carrassi et al. (2018)~~, e.g., Carrassi et al. (2018). Second, it exploits a novel sampling method based on a Gauss-Hermite-like quadrature rule to reach an arbitrary high (depending on the ensemble size) polynomial order of approximation.

The latter feature introduces, to the best of our knowledge, a completely new class of DA algorithms with an order of approximation higher than 2. The advantage of moving to a higher order of approximation has previously been documented in the literature, in particular ~~Nerger et al. (2005)~~ Nerger et al. (2005) compares order 1 algorithms ~~(e.g., SEEK Pham et al. (1998)~~

~~)~~ (e.g., SEEK Pham et al., 1998) and order 2 algorithms ~~(e.g., SEIK Pham (1996))~~(e.g., SEIK Pham, 2001). Results of Section 5 confirm the expected benefits of the ~~high order~~ high-order sampling adopted in the GHOSH filter. In particular, in the Lorenz96 idealized and controlled conditions, the novel method outperforms (up to ~~3 times better~~ 70% lower RMSE) a typical ~~second order~~ second-order method like SEIK and features higher stability and better performances. ~~At the same time, in a realistic 3D high dimensional parallel implementation with assimilation of real satellite data, the GHOSH filter proved (in Section ??) to be a reliable data assimilation scheme with 30% RMSD reduction for the assimilated variable. Moreover, the GHOSH assimilation does not degrade non-assimilated variables, actually improving their RMSDs with respect to independent observations (figures ?? and ??). Consistently, the GHOSH assimilation results allowed catching some relevant system dynamics (Fig. ??). Further, the sensitivity analysis performed in the 3D application underlines the role of the high order sampling in improving performances in a non-assimilated variable (Tab. ??).~~

Thanks to the large number of settings tested in the Lorenz96 twin experiment, it is possible to suggest the conditions where the GHOSH filter considerably increases the assimilation skill and reliability. Concerning the ensemble size, the larger improvements occur when the dimension of the ~~error~~ uncertainty subspace spanned by the ensemble is big enough to be representative. ~~In other words, if~~ If the number of ensemble members is too small, the assimilation simply fails independently from the ~~assimilation algorithm.~~ tested assimilation algorithm, but even in that case, the GHOSH filter error is less detrimental than the SEIK error. Moreover, it has been shown that the GHOSH increased accuracy reduces instabilities and numerical divergence in our Lorenz96 implementation (Fig. 5). This aspect is consistent with the fact that a higher order of approximation helps to manage non-linearities ~~(Nerger et al. (2005))~~(Nerger et al., 2005). GHOSH filter showed a lower need for inflation with respect to SEIK, removing one of the instability causes observed in the twin experiment (Section ~~??). Together with the GHOSH's~~ 5). The lower need of inflation ~~, the hybridization adopted in the 3D application prevented the "collapse into the mean" effect (not shown), which often affects ensemble algorithms (Anderson (2007)). Indeed, the hybridization adopted in each *forecast*~~ can be also seen as a GHOSH improved capacity to take into account non-linearity. Indeed, non-linearity typically introduces errors that are not considered by ensemble filters, which needs inflation to compensate for this uncertainty covariance underestimation (see e.g., Raanes et al., 2019; Bocquet et al., 2015; Rainwater and Hunt, 2013). Finally, the GHOSH filter could remarkably perform a successful assimilation even when the SEIK filter is failing due to observation sparsity.

Even if a large number of settings have been tested in the twin experiments, other Lorenz96 literature applications (see e.g., Brajard et al., tested ensemble DA methods applying different conditions and strategies. In particular, longer time series, different observation operators or different state vector dimensions have been considered. The shorter time series in the first experiment phase in the ~~GHOSH 3D application has the beneficial effect of maintaining a reasonable spread, even without the forgetting factor (i. e. , $\rho = 1$).~~ present work (Section 4.4) are motivated by the aim to focus on comparing DA convergence in the phase after the assimilation and not on long-term convergence. Thank to this approach, our results provide information for possible GHOSH applications in realistic applications in geoscientific operational simulations. In this framework, the shown initial fast convergence of the GHOSH filter is strongly beneficial to improve the simulation accuracy on temporal scales of interest.

~~The results obtained coupling~~

In addition to a higher polynomial order, the GHOSH filter ~~to the OGSTM-BFM model show a major improvement in the nitrate performance when increasing the approximation order h~~features a resampling taking place twice per step. Ensemble Kalman-like filters usually adopt an interpolation strategy by applying the observation operator directly to the $\tilde{\mathbf{X}}_i^f$ ensemble of the states after the model evolution (see equation (28)). However, ~~chlorophyll and phosphate RMSDs are not substantially affected by h. Consistently with the twin experiments results (Section **??**) , this behaviour may depend on the relatively low number of ensemblemembers (16, $r = 15$), which imply that only a limited number of principal components ($s = 3$) takes advantage of the higher order $h = 5$. Indeed, the negligible effects of the higher approximation order on chlorophyll and phosphate could be motivated by the fact that, with the normalization rules chosen in this experiment (i.e. , Λ_i), they possibly contribute little to the few most relevant components of the state uncertainty. A larger number of ensemble members (i.e.~~ this strategy might lead to inaccurate estimations because the covariance matrix $\mathbf{Q}_i$ of equation (31) is not taken into account in the interpolation. For this reason, if Q_i is not zero, GHOSH applies the observation operator to a new ensemble, ~~bigger r) would allow a larger s-subspace, probably extending the high-order effects on further state variables. This view is consistent with the generally observed benefits of a larger ensemble in EnKF-based data assimilation (Edwards et al. (2015); Nerger et al. (2005); Sacher and B)~~resampled between the forecast and analysis phase (Fig. 2). In this way, $\mathbf{Q}_i$ is taken into account and, especially in the case of a nonlinear observation operator, the high order of the sampling method is exploited to improve the accuracy of the projection into the observation space (as in Part 2 Spada et al., 2024).

The statistical moments μ in equation (1) are key hyper-parameters that describe the uncertainty pdf moments for orders higher than 2. In our experiments we used the moments of a Gaussian distribution, but it is worth noting that the GHOSH algorithm does not enforce this choice and other pdfs could be used. However, Kalman filter's analysis equations somehow prescribe a Gaussian approximation, thus the GHOSH filter keeps a link to Gaussianity, even if the GHOSH sampling does not. Interestingly, other filters, like Hodyss (2011), try to overcome this limitation and could represent good candidates to study the effects of a non-Gaussian GHOSH sampling.

The scaling matrix $\mathbf{\Lambda}_i$ used in the sampling procedure plays ~~a relevant~~ an important role in the filter, since $\mathbf{\Lambda}_i$ defines how variables are compared to each others in the computation of the ~~state error~~ most relevant components of the state uncertainty (Equation (7)). In the Lorenz96 case, it is straightforward to set $\mathbf{\Lambda}_i$ equal to the identity matrix, since each variable is equivalent to any other. In ~~the realistic OGSTM-BFM application, we adopted a normalization matrix to represent $\mathbf{\Lambda}_i$ (Section **??**) to properly scale the different biogeochemical variables . However, this strategy implies that the most relevant components are composed by variables that have a high correlation with as many other variables as possible, including processes that can be less relevant for an analysis more interested in the photic layer (e.g. , deep layer processes). Alternative options can be investigated and, for instance, similarly to scaling strategies applied in sampling methods proposed in literature (Dovera and Della Rossa (2012)),~~ more complex applications, $\mathbf{\Lambda}_i$ ~~could be designed to focus on the most relevant processes. In particular, $\mathbf{\Lambda}_i$ can be a sparse matrix with non-zero coordinates corresponding to variables and regions of interest based on information derived by a climatology or a multi-annual simulation.~~

~~As for $\mathbf{\Lambda}_i$, a preliminary approach is applied for the definition of the model error covariance matrix $\mathbf{Q}_i$ (with its local counterpart $\mathbf{Q}_i^p$) used in the 3D realistic GHOSH implementation (Section **??**) . The diagonal matrix $\mathbf{Q}_i$ is not very representative~~

~~of the model error covariances of the relatively complex OGSTm-BFM model. A more advanced parameterization of $\mathbf{Q}_i^p$ deserves to be further investigated for future applications (e.g., using a non-diagonal low rank matrix as in 3D-variational methods, Teruzzi et al. (2014)).~~ should be carefully designed to focus on the variables relevant for the processes of interest (see also Spada et al., 2024, for a non-identity $\mathbf{\Lambda}_i$ and further discussion).

In the present implementation, the GHOSH filter derives the uncertainty subspace basis (equation 17) using a $\mathbf{T}$ matrix (equation 18) inherited by the SEIK filter. This is an advantage from the computational point of view, but ensemble members after the resampling may be very different from the original ensemble. For this reason, in some applications issues can arise, for instance, in the case of the physical balance constraints. As a solution, Nerger et al. (2012) propose a filter (the error subspace transform Kalman filter, ESTKF) that minimizes the differences after the resampling procedure, and such approach can also be applied to the GHOSH filter algorithm by choosing an opportunely designed $\mathbf{T}$ matrix. On the other hand, a GHOSH filter built with the ESTKF approach would present an additional complexity layer, since the $\mathbf{T}$ matrix should change at every forecast step. Given the additional complexity of the ESTKF approach and considering that: i) its effects are less important if the state vector does not strongly need to preserve specific non-linear constrains; ii) Nerger et al. (2012) showed that, in Lorenz96, SEIK with random rotation (adopted also in GHOSH) results to be as good as ESTKF; in the present work we limited to adopt a GHOSH filter that relies on a SEIK-like $\mathbf{T}$ matrix.

Compared to other ensemble filters, the key feature of the GHOSH filter is the higher polynomial order of its sampling method. The advantage of a higher order is not limited to a better estimation of the mean state (Appendix A) but it extends to the covariance matrix of the ~~error~~ uncertainty probability distribution. In fact, the polynomial order in the estimation of the covariance matrix is half the order of the mean (not shown). Since the covariance matrix is involved in the state estimation, the more accurate GHOSH results shown in Section 5 can be partly related to an improvement in the ~~error covariance matrix order of approximation .~~

~~In addition to an higher polynomial order, the GHOSH filter features a resampling taking place twice per step. Ensemble Kalman-like filters usually adopt an interpolation strategy by applying the observation operator directly to the $\tilde{\mathbf{X}}_i^f$ ensemble of the states after the model evolution (see equation (28)). However, this strategy might lead to inaccurate estimations because the model error covariance matrix $\mathbf{Q}_i$ is not taken into account in the interpolation. For this reason, GHOSH applies the observation operator to a new ensemble, resampled between the forecast and analysis phases (Fig. 2). In this way, $\mathbf{Q}_i$ is taken into account and, especially in the case of a nonlinear observation operator, the high order of the sampling method is exploited to improve the accuracy of the projection into the observation space~~approximation error affecting the uncertainty covariance matrix.

Similarly to other ensemble filters, a weighted ensemble is used in the GHOSH filter. However, the GHOSH filter differs substantially from the particle filter ~~(Bocquet et al. (2010)) and~~ (e.g., van Leeuwen et al., 2019) and from other types of weighted ensemble filters, ~~since the latters evolve the value of~~ e.g., Hoteit et al. (2008) and Stordal et al. (2011). In the listed filters, weights (representing likelihood) change over time, while particles remain the same (and strategies are applied to resample the particles wh . In the case of GHOSH, the ensemble is used to estimate specific moments and then resampled to keep constant the weights, ~~whilst GHOSH keeps them fixed and resamples the ensemble to find new members that are well-suited for each (not evolving) weight. Adopting this approach, the problem of some particle weights quickly tending to zero, that usually affects particle filters~~

~~(van Leeuwen et al. (2019))can be avoided.~~ because those weights have specific desired properties that help reduce the error of the mean estimation. In this, GHOSH has similarities with the nonlinear ensemble transform filter (Tödter and Ahrens, 2015), which does a resampling to keep the weights constant but with uniform weights and a second-order sampling procedure.

~~Finally~~Finally, the computational complexity of the GHOSH filter (Section 3.1.6) is comparable to second-order deterministic filters (e.g., SEIK, ETKF). The computational cost, as in most ensemble methods, ~~the computational cost of the GHOSH filter~~ mainly depends on the ensemble size~~(Nerger et al. (2005)). The performance of GHOSH and SEIK implemented in Lorenz96 and the profiling of the GHOSH implemented in OGSTM showed that the~~ , which multiplies the most demanding model integration cost (e.g., Nerger et al., 2005). Hence, the computational cost of the GHOSH filter is no more a concern than in other ~~order 2~~ second-order data assimilation schemes~~(5~~, affecting the total computational time only by a 4% even in the case of the simple Lorenz96 toy model implementation (Section 5.3). The benefits of the higher order offered by the GHOSH filter ~~do not depend on a bigger computational effort, instead they~~ rely on a mathematical advantage with limited impacts on computational effort. In fact, the GHOSH allows the exploitation of the degrees of freedom in the randomness of the sampling process, choosing ensembles that have a better consistency with the ~~error~~ uncertainty distribution.

7 Conclusions

This work presented a novel ensemble hybrid DA filter (GHOSH). The GHOSH filter features a higher order of approximation compared to other ensemble filters, which resulted in better data assimilation performances.

The order of an ensemble method is one of the most relevant proxies of a filter skill. In fact, the order 2 schemes (e.g., SEIK, ETKF, ESTKF) are ~~peferred~~ preferred nowadays to older order 1 methods (e.g., SEEK). This study represents a natural further step in this direction, opening to a new class of data assimilation schemes with order greater than 2.

~~The outstanding improvements of the high order GHOSH filter with respect to the SEIK filter are demonstrated in an extensive set of twin experiments based on the Lorenz96 model. Higher RMSE reductions (GHOSH error up to three times smaller than SEIK) occurred in case of small inflation, large number of observations or large ensemble size. Considering each filter tuned with its optimal forgetting factor~~While increasing the order usually involves increasing the number of ensemble members (and consequently the computational cost), the GHOSH ~~filter showed improved capacity to take into account non-linearities by a lesser need of inflation with respect to SEIK , and achieved the best RMSE reductions (GHOSH error lower than half of the SEIK error) in cases of moderate ensemble size.~~

approach keeps the same computational complexity and number of ensemble members as SEIK or ETKF and exploits the advantage of a higher order only where the approximation errors are larger.

The feasibility of the GHOSH filter in a realistic geophysical application has also been assessed in a companion paper (Spada et al., 2024) by testing a ~~3D~~ 3D parallel implementation of a Mediterranean physical-biogeochemical model. ~~The GHOSH improved the agreement between the forecast and observations without producing unrealistic effects on the non-assimilated variables. Considering relevant marine ecosystem dynamics, GHOSH filter successfully constrained the uncertainty of the~~

~~phytoplankton winter bloom events while assimilating ocean color. Furthermore, a sensitivity analysis indicates that the use of a higher order has positive impacts on the performance of a non-assimilated variable (e.g., nutrients).~~

~~From the computational point of view, the implementation of the~~

The outstanding improvements of the GHOSH filter with respect to the SEIK filter are demonstrated in an extensive set of twin experiments based on the Lorenz96 model. The GHOSH filter showed improved capacity to take into account non-linearities by a lesser need of inflation with respect to SEIK, along with the remarkable capability of performing a successful assimilation even when observation sparsity was producing a SEIK failure. In every configuration, the GHOSH filter proved to be as ~~demanding as the SEIK , given that the computational cost is dominated by the integration of the model and by the ensemble size~~good as SEIK or better, reaching up to a 70% RMSE reduction with respect to SEIK when tuned with the same forgetting factor. Considering each filter tuned with its optimal forgetting factor, the GHOSH filter significantly improved the assimilation skill with respect to SEIK, achieving its best RMSE reduction (54%) in the setting with moderate ensemble size (31 members).

Code availability. A GHOSH python implementation is available from the GitHub page: https://github.com/Sword-Code/PythonDA under the licence GNU GPLv3. The exact version used to produce the twin experiment results and the related plots (Section 5) is archived on Zenodo (https://doi.org/10.5281/zenodo.12960765).

Appendix A: High-order sampling

Sampling a limited number of ensemble members that effectively represent the uncertainty of the state estimation is a crucial part of any ensemble algorithm. The GHOSH filter uses multi-dimensional Gauss-Hermite-like quadrature rules to choose the ensemble members, achieving a high polynomial order of convergence.

The core idea behind the GHOSH sampling can be proved by taking two independent random variables with the same moments up to a certain order h. ~~Then~~Thus, they have the same mean after applying any polynomial operator of order h. In fact, given a probability density function (pdf) $p : \mathbb{R}^s \to \mathbb{R}$, if $\mathcal{F} : \mathbb{R}^s \longrightarrow \mathbb{R}$ is a polynomial operator such that

$$\mathcal{F}(\boldsymbol{x}) = \sum_{\xi=0}^{h} \sum_{j_1,\ldots,j_\xi=1}^{s} f_\xi^{j_1,\ldots,j_\xi} x_{j_1} \cdots x_{j_\xi}, \tag{A1}$$

where s is the space dimension, $f_\xi^{j_1,\ldots,j_\xi}$ are the polynomial coefficients and $x_1,\ldots,x_s$ are the coordinates of $\boldsymbol{x}$, then the mean $\bar{\mathcal{F}}$,

$$\bar{\mathcal{F}} = \int_{\mathbb{R}^s} \mathcal{F}(\boldsymbol{x})\, p(\boldsymbol{x})\, dx = \sum_{\xi=0}^{h} \sum_{\cancel{j_1,\ldots,j_\xi=1}\, j_1=1}^{s} \cdots \sum_{j_\xi=1}^{s} f_\xi^{j_1,\ldots,j_\xi} \int_{\mathbb{R}^s} x_{j_1} \cdots x_{j_\xi}\, p(\boldsymbol{x})\, d\boldsymbol{x}, \tag{A2}$$

depends only on the first h moments of the ~~probability density function~~ pdf p, which are represented by the last integral in equation (A2).

This proves that the mean of an operator can be computed exactly up to a certain order h by substituting the sampled probability distribution with any discrete finite probability distribution (~~i.e.,~~ such as a weighted ensemble) as long as their moments match (up to order h).

In order to build an ensemble with this property, the moment-matching equations are summarized ~~in~~ by the nonlinear system

$$\begin{cases} \vdots \\ \sum_{m=1}^{r+1} x_{j_1,m} \cdots x_{j_\xi,m} w_m = \mu_{j_1,\ldots,j_\xi}, \\ \vdots \end{cases} \tag{A3}$$

with one equation for each $\xi \in \{0,\ldots,h\}$ and for each $j_1,\ldots,j_\xi \in \{1,\ldots,s\}$.

In system (A3), ~~x_j^m is the j-th~~ $x_{j,m}$ is the j^{th} coordinate of the ~~m-th~~ m^{th} ensemble member, $r+1$ is the ensemble size, $w_m \geq 0$ are the weights and $\mu_{j_1,\ldots,j_\xi}$ are the statistical moments of p, i.e.,

$$\mu_{j_1,\ldots,j_\xi} = \int_{\mathbb{R}^s} x_{j_1} \cdots x_{j_\xi} p(\boldsymbol{x})\, d\boldsymbol{x}. \tag{A4}$$

The system is solved considering $\cancel{x_j^m}$ $x_{j,m}$ and w_m as the unknowns, while p can be taken uncorrelated, normalized and with 0 mean without loss of generality. In fact, the ensemble matrix $\cancel{\tilde{\mathbf{X}} = (x_j^m)}$ $\tilde{\mathbf{X}} = (x_{j,m})$ (which contains in its columns the ensemble members with 0 mean and covariance matrix equal to the identity matrix $\mathbf{I}_s$) can be transformed in the ensemble matrix X with arbitrary mean $\bar{\boldsymbol{x}}$ and covariance matrix LL^T through the projection

$$\mathbf{X} = \bar{\boldsymbol{x}} \mathbb{1}_{1\times(r+1)} + \mathbf{L}\tilde{\mathbf{X}}, \tag{A5}$$

where $\mathbb{1}_{1\times(r+1)}$ is a matrix (of the subscripted size) filled with ones.

In the special case of the standard normal distribution, i.e., $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}, 0, \mathbf{I}_s)$, the moments $\mu_{j_1,\ldots,j_\xi}$ are given by

$$\int_{\mathbb{R}^s} {x_1}^{k_1} \cdots {x_s}^{k_s} \mathcal{N}(\boldsymbol{x}; 0, \mathbf{I}_s)\, d\boldsymbol{x} = 0 \tag{A6}$$

if any exponent k_j is an odd number, or

$$\int_{\mathbb{R}^s} {x_1}^{k_1} \cdots {x_s}^{k_s} \mathcal{N}(\boldsymbol{x}; 0, \mathbf{I}_s)\, dx = \frac{k_1!}{\sqrt{2}^{k_1}\left(\frac{k_1}{2}!\right)} \cdots \frac{k_s!}{\sqrt{2}^{k_s}\left(\frac{k_s}{2}!\right)} \tag{A7}$$

otherwise. Further, the solutions $\cancel{x_j^m}$ $x_{j,m}$ and w_m are the nodes and weights of the Gauss-Hermite quadrature rule.

In general, by applying to system (A3) the substitution

$$\begin{cases} w_m = {v_m}^2, \\ x_{j,m} = \dfrac{u_{m,j}}{v_m}, \end{cases} \tag{A8}$$

the weights are naturally forced to be positive, and the system can be conveniently rewritten as

$$\begin{cases} \vdots \\ \displaystyle\sum_{m=1}^{r+1} u_{m,j_1} \cdots u_{m,j_\xi} {v_m}^{2-\xi} = \mu_{j_1,\ldots,j_\xi}, \\ \vdots \end{cases} \tag{A9}$$

which, in the case of order $h = 2$, can be expressed in matrix form as

$$\left(\begin{array}{c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h \end{array} \right)^T \left(\begin{array}{c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h \end{array} \right) = \mathbf{I}_s, \tag{A10}$$

where $\boldsymbol{v} \in \mathbb{R}^{r+1}$ is the column vector with coordinates v_m and $\boldsymbol{\Omega}^h$ is an $(r+1) \times s$ matrix with coordinates $\cancel{u_j^m}u_{m,j}$. Note that equation (A10) can be solved for any unit vector $\boldsymbol{v}$ and for any value of $s \leq r$. If constant weights are chosen and $s = r$, then this method is equivalent to the second-order exact sampling of the SEIK filter ~~(Pham (1996))~~(Pham, 2001), which leads to, in the formalism of equation (A5),

$$\mathbf{X} = \bar{\boldsymbol{x}} \mathbb{1}_{1 \times (r+1)} + \mathbf{L} {\boldsymbol{\Omega}^h}^T \mathbf{W}^{-\frac{1}{2}}, \tag{A11}$$

where $\mathbf{W} = \mathrm{diag}(\boldsymbol{w})$ is the $(r+1) \times (r+1)$ diagonal matrix of weights.

In general, if $h > 2$, s must be lower than r to obtain a solution to system (A9). The obtained s-dimensional ensemble can be extended to an r-dimensional one: its projection on the s-dimensional subspace defined by the first s coordinates will retain order h, while the remaining components are sampled with order 2. This is achieved by starting from $\boldsymbol{v}$ and $\boldsymbol{\Omega}^h$, given by a particular solution of system (A9). Then, exploiting symmetry and rotational invariance, randomness is added to avoid biases by multiplying $\boldsymbol{\Omega}^h$ by any random $s \times s$ orthogonal matrix $\boldsymbol{\Omega}^{rnd}$. Finally, the $(r+1) \times (r-s)$ matrix $\boldsymbol{\Omega}^\perp$ is chosen to fulfil

$$\left(\begin{array}{c|c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h \boldsymbol{\Omega}^{rnd} & \boldsymbol{\Omega}^\perp \end{array} \right)^T \left(\begin{array}{c|c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h \boldsymbol{\Omega}^{rnd} & \boldsymbol{\Omega}^\perp \end{array} \right) = \mathbf{I}_{r+1}. \tag{A12}$$

A procedure for randomly building or completing orthogonal matrices is described in Appendix B.

The sampling matrix $\boldsymbol{\Omega}$ is defined as

$$\boldsymbol{\Omega} = \left(\begin{array}{c|c} \boldsymbol{\Omega}^h \boldsymbol{\Omega}^{rnd} & \boldsymbol{\Omega}^\perp \end{array} \right). \tag{A13}$$

Now the columns of $\cancel{\boldsymbol{\Omega}^T W^{-\frac{1}{2}}}$ $\boldsymbol{\Omega}^T \mathbf{W}^{-\frac{1}{2}}$ represent an r-dimensional ensemble of order h in the first s coordinates and order 2 in the last $r-s$ coordinates. Equation (A11) needs to be modified to properly project the higher order subspace in the

~~most relevant~~ principal components of the covariance matrix $\mathbf{L}\mathbf{L}^T$. ~~It can be proved that such~~ Such components, ordered by relevance, are the column of the matrix $\mathbf{L}\mathbf{C}^T$, where $\mathbf{C}$ is a $r \times r$ orthogonal change-of-basis matrix such that

$$\mathbf{L}^T \boldsymbol{\Lambda}^{-1} \mathbf{L} = \mathbf{C}^T \mathbf{E} \mathbf{C}. \tag{A14}$$

In the last equation, the right-hand side is an eigenvalue decomposition of the left-hand side, with ~~E~~ $\mathbf{E}$ being an $r \times r$ diagonal matrix of eigenvalues in descending order and $\boldsymbol{\Lambda}$ being an appropriate $N \times N$ matrix. The purpose of $\boldsymbol{\Lambda}$ has already been discussed in Section 2.2.

~~After the~~ The eigenvalue decomposition is performed to produce a PCA of the uncertainty represented by the ensemble. In fact, the $\mathbf{C}$ matrix induces a change of basis ~~induced by C, the most relevant components~~ such that the principal component are on the left side of the matrix product ~~LC^T, which~~ $\mathbf{L}\mathbf{C}^T$. This is compliant with the ensemble stored in ~~$\boldsymbol{\Omega}^T W^{-\frac{1}{2}}$~~ $\boldsymbol{\Omega}^T \mathbf{W}^{-\frac{1}{2}}$, and the sampling is finally obtained by

$$\mathbf{X} = \bar{\boldsymbol{x}} \mathbb{1}_{1\times(r+1)} + \mathbf{L}\mathbf{C}^T {\boldsymbol{\Omega}^h}^T \mathbf{W}^{-\frac{1}{2}}. \tag{A15}$$

Appendix B: Orthogonal matrices

The GHOSH sampling algorithm requires to randomly generate orthogonal matrices, as in the case of $\boldsymbol{\Omega}^{rnd}$, or to complete a set of orthonormal vectors to form an orthogonal matrix, as in the case of $\boldsymbol{\Omega}^{\perp}$ (equations (4) and (A12)). Algorithms to extract such matrices have been discussed in literature (e.g., Pham, 2001; Nerger et al., 2012). Here we propose the algorithm that we have implemented in the provided python code.

Given a $m \times (m-s)$ matrix $\boldsymbol{\Omega}^{=}$, the columns of witch are a set of $(m-s)$ orthonormal vectors in $\mathbb{R}^m$, we want to randomly sample a $m \times s$ matrix $\boldsymbol{\Omega}^{\perp}$, such that the $m \times m$ matrix $\boldsymbol{\Omega}$ defined as

$$\boldsymbol{\Omega} := \left(\begin{array}{c|c} \quad \boldsymbol{\Omega}^{=} \quad & \quad \boldsymbol{\Omega}^{\perp} \quad \end{array} \right) \tag{B1}$$

is an orthogonal matrix.

To do so we proceed in an iterative way, building $\boldsymbol{\Omega}^{\perp}$ column by column. Thus, the problem reduces to producing a random vector $\boldsymbol{w} \in \mathbb{R}^m$ orthonormal to the columns of a given $\boldsymbol{\Omega}^{=}$. This can be done through the following steps:

1. extract a random unitary vector $\boldsymbol{u} \in \mathbb{R}^m$, i.e.:

   (a) generate a column vector $\boldsymbol{v} \in \mathbb{R}^m$ by extracting m random numbers from a standard normal distribution (i.e., 0-mean and unitary variance), one for each coordinate of $\boldsymbol{v}$;

   (b) normalize $\boldsymbol{v}$ to get $\boldsymbol{u} = \dfrac{\boldsymbol{v}}{\sqrt{\boldsymbol{v}^T \boldsymbol{v}}}$

2. decompose $\boldsymbol{u} = \boldsymbol{u}^{=} + \boldsymbol{u}^{\perp}$, with the first term laying in the subspace generated by the columns of $\boldsymbol{\Omega}^{=}$ and the last term in its orthogonal subspace, i.e.:

$$
\begin{aligned}
\boldsymbol{u}^{=} &= \boldsymbol{\Omega}^{=} \left(\boldsymbol{\Omega}^{=}\right)^{T} \boldsymbol{u}, \\
\boldsymbol{u}^{\perp} &= \boldsymbol{u} - \boldsymbol{u}^{=};
\end{aligned}
$$

3. normalize $\mathbf{u}^{\perp}$ to obtain $\boldsymbol{w} = \dfrac{\boldsymbol{u}^{\perp}}{\sqrt{\left(\boldsymbol{u}^{\perp}\right)^{T} \boldsymbol{u}^{\perp}}}$.

The vector $\mathbf{w}$ has unitary norm and it is orthogonal to $\boldsymbol{\Omega}^{=}$ as $\boldsymbol{u}^{\perp}$. The next columns of $\boldsymbol{\Omega}^{\perp}$ are computed in the same manner, after adding $\boldsymbol{w}$ as a new column of $\boldsymbol{\Omega}^{=}$.

This algorithm also covers how to generate a random orthogonal matrix from scratch. In fact, if $s = m$, then $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{\perp}$ and the first computed column of $\boldsymbol{\Omega}^{\perp}$ is $\boldsymbol{w} = \boldsymbol{u}^{\perp} = \boldsymbol{u}$.

Appendix C: Examples

The aim of this appendix is to help the reader understanding the concept of h^{th}-order of approximation by presenting some simple examples of ensembles with a certain order of approximation h. The examples are organized by the space dimension r. In all the examples the ensemble members, noted as $\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots$, are points of $\mathbb{R}^r$, and the target pdf, the moments of which are matched up to order h, is the standard normal distribution.

C1 Dimension 1 ($r = 1$)

The target pdf is the one-dimensional standard normal distribution $\mathcal{N}(x; 0, 1)$. According to equations (A6) and (A7), the moments characterizing the pdf are:

– the first moment (mean): $\mu_x = 0$,

– the second moment (variance): $\mu_{x^2} = 1$,

– the third moment (skewness): $\mu_{x^3} = 0$,

– the fourth moment (kurtosis): $\mu_{x^4} = 3$,

– the fifth moment: $\mu_{x^5} = 0$, ...

C1.1 Second-order ensemble ($h = 2$) with 2 members

$$
\begin{aligned}
\boldsymbol{P}_1 &: \quad x_1 = 1; \\
\boldsymbol{P}_2 &: \quad x_2 = -1.
\end{aligned}
\tag{C1}
$$

The ensemble first moment (mean) is:

$$\frac{1}{2}\left(x_1+x_2\right)=\frac{1}{2}\left(1+(-1)\right)=0. \tag{C2}$$

The ensemble second moment (variance) is:

$$\frac{1}{2}\left({x_1}^2+{x_2}^2\right)=\frac{1}{2}\left(1+1\right)=1. \tag{C3}$$

Thus, the ensemble $\boldsymbol{P}_1, \boldsymbol{P}_2$ matches the moments of the target pdf up to order 2.

C1.2 Third-order ensemble ($h = 3$) with 2 members

The present example uses ensemble (C1) of the previous example.

In one dimension, the ensemble of order 2 is also an ensemble of order 3. The same does not happen for higher dimensions.

The ensemble third moment (skewness) is:

$$\frac{1}{2}\left({x_1}^3+{x_2}^3\right)=\frac{1}{2}\left(1+(-1)\right)=0. \tag{C4}$$

Thus, the ensemble $\boldsymbol{P}_1, \boldsymbol{P}_2$ matches the moments of the target pdf up to order 3.

C1.3 Fifth-order weighted ensemble ($h = 5$) with 3 members

$$\begin{aligned}
&\boldsymbol{P}_1: \quad x_1=\sqrt{3}, \qquad w_1=\frac{1}{6};\\
&\boldsymbol{P}_2: \quad x_2=-\sqrt{3}, \quad w_2=\frac{1}{6};\\
&\boldsymbol{P}_3: \quad x_3=0, \qquad\;\; w_3=\frac{2}{3}.
\end{aligned} \tag{C5}$$

The ensemble first moment (mean) is:

$$w_1x_1+w_2x_2+w_3x_3=\frac{1}{6}\cdot\sqrt{3}+\frac{1}{6}\cdot\left(-\sqrt{3}\right)+\frac{2}{3}\cdot 0=0. \tag{C6}$$

The ensemble second moment (variance) is:

$$w_1{x_1}^2+w_2{x_2}^2+w_3{x_3}^2=\frac{1}{6}\cdot 3+\frac{1}{6}\cdot 3+\frac{2}{3}\cdot 0=1. \tag{C7}$$

The ensemble third moment (skewness) is:

$$w_1{x_1}^3+w_2{x_2}^3+w_3{x_3}^3=\frac{1}{6}\cdot\sqrt{3}^3+\frac{1}{6}\cdot\left(-\sqrt{3}\right)^3+\frac{2}{3}\cdot 0=0. \tag{C8}$$

The ensemble fourth moment (kurtosis) is:

$$w_1{x_1}^4+w_2{x_2}^4+w_3{x_3}^4=\frac{1}{6}\cdot 9+\frac{1}{6}\cdot 9+\frac{2}{3}\cdot 0=3. \tag{C9}$$

The ensemble fifth moment is:

$$w_1{x_1}^5 + w_2{x_2}^5 + w_3{x_3}^5 = \frac{1}{6}\cdot\sqrt{3}^5 + \frac{1}{6}\cdot\left(-\sqrt{3}\right)^5 + \frac{2}{3}\cdot 0 = 0. \tag{C10}$$

Thus, the ensemble $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3$ matches the moments of the target pdf up to order 5.

C2 Dimension 2 ($r = 2$)

The target pdf is the two-dimensional standard normal distribution $\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \mathbf{I}_2)$. According to equations (A6) and (A7), the moments characterizing the pdf are:

- first moments:
  - first moment associated to x (i.e., the mean along x) is: $\mu_x = 0$,
  - first moment associated to y (i.e., the mean along y) is: $\mu_y = 0$,
- second moments:
  - second moment associated to x^2 (i.e., the variance along x) is: $\mu_{x^2} = 1$,
  - second moment associated to y^2 (i.e., the variance along y) is: $\mu_{y^2} = 1$,
  - second moment associated to xy (i.e., the covariance between x and y) is: $\mu_{xy} = 0$,
- third moments:
  - third moment associated to x^3 is: $\mu_{x^3} = 0$,
  - third moment associated to x^2y is: $\mu_{x^2y} = 0$,
  - third moment associated to xy^2 is: $\mu_{xy^2} = 0$,
  - third moment associated to y^3 is: $\mu_{y^3} = 0$,
- fourth moments:
  - fourth moment associated to x^4 is: $\mu_{x^4} = 3$,
  - fourth moment associated to x^3y is: $\mu_{x^3y} = 0$,
  - fourth moment associated to x^2y^2 is: $\mu_{x^2y^2} = 1$,
  - fourth moment associated to xy^3 is: $\mu_{xy^3} = 0$,
  - fourth moment associated to y^4 is: $\mu_{y^4} = 3$,
- fifth moments:
  - fifth moment associated to x^5 is: $\mu_{x^5} = 0$,

- fifth moment associated to x^4y is: $\mu_{x^4y} = 0$,
- fifth moment associated to x^3y^2 is: $\mu_{x^3y^2} = 0$,
- fifth moment associated to x^2y^3 is: $\mu_{x^2y^3} = 0$,
- fifth moment associated to xy^4 is: $\mu_{xy^4} = 0$,
- fifth moment associated to y^5 is: $\mu_{y^5} = 0$, ...

C2.1 Second-order ensemble ($h = 2$) with 3 members

$$
\begin{aligned}
\boldsymbol{P}_1: &\quad (x_1, y_1) = \left(\sqrt{\frac{3}{2}}, -\frac{\sqrt{2}}{2}\right); \\
\boldsymbol{P}_2: &\quad (x_2, y_2) = \left(-\sqrt{\frac{3}{2}}, -\frac{\sqrt{2}}{2}\right); \\
\boldsymbol{P}_3: &\quad (x_3, y_3) = \left(0, \sqrt{2}\right).
\end{aligned}
\tag{C11}
$$

This ensemble has the shape of an equilateral triangle. It is one of the most common second-order ensembles sampled by deterministic sampling methods like SEIK's second-order-exact sampling (Pham, 2001).

The ensemble first moment associated to x (i.e., the mean along x) is:

$$
\frac{1}{3}(x_1 + x_2 + x_3) = \frac{1}{3}\left(\sqrt{\frac{3}{2}} + \left(-\sqrt{\frac{3}{2}}\right) + 0\right) = 0. \tag{C12}
$$

The ensemble first moment associated to y (i.e., the mean along y) is:

$$
\frac{1}{3}(y_1 + y_2 + y_3) = \frac{1}{3}\left(-\frac{\sqrt{2}}{2} + \left(-\frac{\sqrt{2}}{2}\right) + \sqrt{2}\right) = 0. \tag{C13}
$$

The ensemble second moment associated to x^2 (i.e., the variance along x) is:

$$
\frac{1}{3}\left({x_1}^2 + {x_2}^2 + {x_3}^2\right) = \frac{1}{3}\left(\frac{3}{2} + \frac{3}{2} + 0\right) = 1. \tag{C14}
$$

The ensemble second moment associated to y^2 (i.e., the variance along y) is:

$$
\frac{1}{3}\left({y_1}^2 + {y_2}^2 + {y_3}^2\right) = \frac{1}{3}\left(\frac{1}{2} + \frac{1}{2} + 2\right) = 1. \tag{C15}
$$

The ensemble second moment associated to xy (i.e., the covariance between x and y) is:

$$
\frac{1}{3}(x_1y_1 + x_2y_2 + x_3y_3) = \frac{1}{3}\left(\sqrt{\frac{3}{2}} \cdot \left(-\frac{\sqrt{2}}{2}\right) + \left(-\sqrt{\frac{3}{2}}\right) \cdot \left(-\frac{\sqrt{2}}{2}\right) + 0 \cdot \sqrt{2}\right) = 0. \tag{C16}
$$

Thus, the ensemble $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3$ matches the moments of the target pdf up to order 2.

Observe that any symmetry or rotation applied to this ensemble (i.e., applying an orthogonal transformation to the members) preserves its order.

C2.2 High-order sampling ensemble with 3 members, $h = 3$ along the first dimension

The present example uses ensemble (C11) of the previous example.
This particular second-order ensemble has order 3 along the x-axis. In general, this is no longer true if you apply any rotation (or symmetry) to the ensemble.
The ensemble third moment associated to x^3 is:

$$\frac{1}{3}\left({x_1}^3 + {x_2}^3 + {x_3}^3\right) = \frac{1}{3}\left(\left(\sqrt{\frac{3}{2}}\right)^3 + \left(-\sqrt{\frac{3}{2}}\right)^3 + 0\right) = 0. \tag{C17}$$

Thus, the ensemble $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3$ matches the moments of the target pdf up to order 3 along the x-axis, and up to order 2 elsewhere.

C2.3 High-order sampling ensemble with 3 weighted members, $h = 5$ along the first dimension

$$\begin{aligned}
\boldsymbol{P}_1 :&\quad (x_1, y_1) = \left(\sqrt{3}, -\sqrt{2}\right), && w_1 = \frac{1}{6};\\
\boldsymbol{P}_2 :&\quad (x_2, y_2) = \left(-\sqrt{3}, -\sqrt{2}\right), && w_2 = \frac{1}{6};\\
\boldsymbol{P}_3 :&\quad (x_3, y_3) = \left(0, \frac{\sqrt{2}}{2}\right), && w_3 = \frac{2}{3}.
\end{aligned} \tag{C18}$$

Compared to the previous example, this ensemble uses weighted members to achieve a higher order of approximation (5 insted of 3) along the x-axis. In fact, looking at the x-coordinate of the ensemble members (i.e., projecting on the x-axis), it results in the same ensemble as (C5). Thus, it only remains to verify the moments involving also the y direction.
The ensemble first moment associated to y (i.e., the mean along y) is:

$$w_1 y_1 + w_2 y_2 + w_3 y_3 = \frac{1}{6}\cdot\left(-\sqrt{2}\right) + \frac{1}{6}\cdot\left(-\sqrt{2}\right) + \frac{2}{3}\cdot\frac{\sqrt{2}}{2} = 0. \tag{C19}$$

The ensemble second moment associated to y^2 (i.e., the variance along y) is:

$$w_1 {y_1}^2 + w_2 {y_2}^2 + w_3 {y_3}^2 = \frac{1}{6}\cdot 2 + \frac{1}{6}\cdot 2 + \frac{2}{3}\cdot\frac{1}{2} = 1. \tag{C20}$$

The ensemble second moment associated to xy (i.e., the covariance between x and y) is:

$$w_1 x_1 y_1 + w_2 x_2 y_2 + w_3 x_3 y_3 = \frac{1}{6}\cdot\sqrt{3}\cdot\left(-\sqrt{2}\right) + \frac{1}{6}\cdot\left(-\sqrt{3}\right)\cdot\left(-\sqrt{2}\right) + \frac{2}{3}\cdot 0\cdot\frac{\sqrt{2}}{2} = 0. \tag{C21}$$

Thus, the ensemble $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3$ matches the moments of the target pdf up to order 5 along the x-axis, and up to order 2 elsewhere.

C2.4 Third-order ensemble ($h = 3$) with 4 members

$$
\begin{aligned}
&\boldsymbol{P}_1: \quad (x_1, y_1) = \left(\sqrt{2}, 0\right);\\
&\boldsymbol{P}_2: \quad (x_2, y_2) = \left(-\sqrt{2}, 0\right);\\
&\boldsymbol{P}_3: \quad (x_3, y_3) = \left(0, \sqrt{2}\right);\\
&\boldsymbol{P}_4: \quad (x_4, y_4) = \left(0, -\sqrt{2}\right).
\end{aligned}
\tag{C22}
$$

This square-shaped ensemble uses $2r$ members to achieve the third-order.

The ensemble first moment associated to x (i.e., the mean along x) is:

$$\frac{1}{4}\left(x_1 + x_2 + x_3 + x_4\right) = \frac{1}{4}\left(\sqrt{2} + \left(-\sqrt{2}\right) + 0 + 0\right) = 0. \tag{C23}$$

The ensemble first moment associated to y (i.e., the mean along y) is:

$$\frac{1}{4}\left(y_1 + y_2 + y_3 + y_4\right) = \frac{1}{4}\left(0 + 0 + \sqrt{2} + \left(-\sqrt{2}\right)\right) = 0. \tag{C24}$$

The ensemble second moment associated to x^2 (i.e., the variance along x) is:

$$\frac{1}{4}\left({x_1}^2 + {x_2}^2 + {x_3}^2 + {x_4}^2\right) = \frac{1}{4}\left(2 + 2 + 0 + 0\right) = 1. \tag{C25}$$

The ensemble second moment associated to y^2 (i.e., the variance along y) is:

$$\frac{1}{4}\left({y_1}^2 + {y_2}^2 + {y_3}^2 + {y_4}^2\right) = \frac{1}{4}\left(0 + 0 + 2 + 2\right) = 1. \tag{C26}$$

The ensemble second moment associated to xy (i.e., the covariance between x and y) is:

$$\frac{1}{4}\left(x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4\right) = \frac{1}{4}\left(\sqrt{2}\cdot 0 + \left(-\sqrt{2}\right)\cdot 0 + 0\cdot\sqrt{2} + 0\cdot\left(-\sqrt{2}\right)\right) = 0. \tag{C27}$$

The ensemble third moment associated to x^3 is:

$$\frac{1}{4}\left({x_1}^3 + {x_2}^3 + {x_3}^3 + {x_4}^3\right) = \frac{1}{4}\left(\left(\sqrt{2}\right)^3 + \left(-\sqrt{2}\right)^3 + 0 + 0\right) = 0. \tag{C28}$$

The ensemble third moment associated to $x^2 y$ is:

$$\frac{1}{4}\left({x_1}^2 y_1 + {x_2}^2 y_2 + {x_3}^2 y_3 + {x_4}^2 y_4\right) = \frac{1}{4}\left(2\cdot 0 + 2\cdot 0 + 0\cdot\sqrt{2} + 0\cdot\left(-\sqrt{2}\right)\right) = 0. \tag{C29}$$

The ensemble third moment associated to xy^2 is:

$$\frac{1}{4}\left(x_1 {y_1}^2 + x_2 {y_2}^2 + x_3 {y_3}^2 + x_4 {y_4}^2\right) = \frac{1}{4}\left(\sqrt{2}\cdot 0 + \left(-\sqrt{2}\right)\cdot 0 + 0\cdot 2 + 0\cdot 2\right) = 0. \tag{C30}$$

The ensemble third moment associated to y^3 is:

$$\frac{1}{4}\left({y_1}^3+{y_2}^3+{y_3}^3+{y_4}^3\right)=\frac{1}{4}\left(0+0+\left(\sqrt{2}\right)^3+\left(-\sqrt{2}\right)^3\right)=0. \tag{C31}$$

Thus, the ensemble $\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3,\boldsymbol{P}_4$ matches the moments of the target pdf up to order 3.

C2.5 Third-order weighted ensemble with 4 members, fifth-order along the first dimension

$$\begin{aligned}
\boldsymbol{P}_1: &\quad (x_1,y_1)=\left(\sqrt{3},0\right), &\quad w_1=\frac{1}{6};\\
\boldsymbol{P}_2: &\quad (x_2,y_2)=\left(-\sqrt{3},0\right), &\quad w_2=\frac{1}{6};\\
\boldsymbol{P}_3: &\quad (x_3,y_3)=\left(0,\sqrt{\frac{3}{2}}\right), &\quad w_3=\frac{1}{3};\\
\boldsymbol{P}_3: &\quad (x_3,y_3)=\left(0,-\sqrt{\frac{3}{2}}\right), &\quad w_3=\frac{1}{3}.
\end{aligned} \tag{C32}$$

This ensemble extends ensemble (C5) to two dimensions. In fact, looking at the x-coordinates, it results in the same ensemble as (C5) after summing the weights of $\boldsymbol{P}_3$ and $\boldsymbol{P}_4$ collapsing in the same point on the x-axis. Differently from ensemble (C18), which also extend ensemble (C5), it achieves order 3 by using 4 members instead of 3. This is not considered a high-order sampling ensemble, since high-order sampling produces ensembles with $r+1$ members, as in the case of ensemble (C18). Since the moments along the x-axis are already checked for ensemble (C5), it only remains to verify the moments involving the y direction.

The ensemble first moment associated to y (i.e., the mean along y) is:

$$w_1y_1+w_2y_2+w_3y_3+w_4y_4=\frac{1}{6}\cdot 0+\frac{1}{6}\cdot 0+\frac{1}{3}\cdot\sqrt{\frac{3}{2}}+\frac{1}{3}\cdot\left(-\sqrt{\frac{3}{2}}\right)=0. \tag{C33}$$

The ensemble second moment associated to y^2 (i.e., the variance along y) is:

$$w_1{y_1}^2+w_2{y_2}^2+w_3{y_3}^2+w_4{y_4}^2=\frac{1}{6}\cdot 0+\frac{1}{6}\cdot 0+\frac{1}{3}\cdot\frac{3}{2}+\frac{1}{3}\cdot\frac{3}{2}=1. \tag{C34}$$

The ensemble second moment associated to xy (i.e., the covariance between x and y) is:

$$w_1x_1y_1+w_2x_2y_2+w_3x_3y_3+w_4x_4y_4=\frac{1}{6}\cdot\sqrt{3}\cdot 0+\frac{1}{6}\cdot\left(-\sqrt{3}\right)\cdot 0+\frac{1}{3}\cdot 0\cdot\sqrt{\frac{3}{2}}+\frac{1}{3}\cdot 0\cdot\left(-\sqrt{\frac{3}{2}}\right)=0. \tag{C35}$$

The ensemble third moment associated to x^2y is:

$$w_1{x_1}^2y_1+w_2{x_2}^2y_2+w_3{x_3}^2y_3+w_4{x_4}^2y_4=\frac{1}{6}\cdot 3\cdot 0+\frac{1}{6}\cdot 3\cdot 0+\frac{1}{3}\cdot 0\cdot\sqrt{\frac{3}{2}}+\frac{1}{3}\cdot 0\cdot\left(-\sqrt{\frac{3}{2}}\right)=0. \tag{C36}$$

The ensemble third moment associated to xy^2 is:

$$w_1x_1{y_1}^2+w_2x_2{y_2}^2+w_3x_3{y_3}^2+w_4x_4{y_4}^2=\frac{1}{6}\cdot\sqrt{3}\cdot 0+\frac{1}{6}\cdot\left(-\sqrt{3}\right)\cdot 0+\frac{1}{3}\cdot 0\cdot\frac{3}{2}+\frac{1}{3}\cdot 0\cdot\frac{3}{2}=0. \tag{C37}$$

The ensemble third moment associated to y^3 is:

$$w_1{y_1}^3+w_2{y_2}^3+w_3{y_3}^3+w_4{y_4}^3=\frac{1}{6}\cdot 0+\frac{1}{6}\cdot 0+\frac{1}{3}\cdot\left(\sqrt{\frac{3}{2}}\right)^3+\frac{1}{3}\cdot\left(-\sqrt{\frac{3}{2}}\right)^3=0. \tag{C38}$$

Thus, the ensemble $\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3,\boldsymbol{P}_4$ matches the moments of the target pdf up to order 5 along the x-axis, and up to order 3 elsewhere.

C2.6 Fifth-order weighted ensemble ($h=5$) with 7 members

$$\begin{aligned}
&\boldsymbol{P}_1: && (x_1,y_1)=\left(\sqrt{3},1\right), && w_1=\frac{1}{12};\\
&\boldsymbol{P}_2: && (x_2,y_2)=\left(\sqrt{3},-1\right), && w_2=\frac{1}{12};\\
&\boldsymbol{P}_3: && (x_3,y_3)=\left(-\sqrt{3},1\right), && w_3=\frac{1}{12};\\
&\boldsymbol{P}_4: && (x_4,y_4)=\left(-\sqrt{3},-1\right), && w_4=\frac{1}{12};\\
&\boldsymbol{P}_5: && (x_5,y_5)=(0,2), && w_5=\frac{1}{12};\\
&\boldsymbol{P}_6: && (x_6,y_6)=(0,-2), && w_6=\frac{1}{12};\\
&\boldsymbol{P}_7: && (x_7,y_7)=(0,0), && w_7=\frac{1}{2}.
\end{aligned} \tag{C39}$$

This hexagon-shaped ensemble uses its 7 weighted members to achieve the fifth-order. Looking at the x-coordinate of the ensemble members (i.e., projecting on the x-axis), it results in the same ensemble as (C5) by summing the weights of points with the same projection.

The ensemble first moment associated to x (i.e., the mean along x) is:

$$w_1x_1+w_2x_2+\ldots+w_7x_7=\frac{1}{12}\cdot\sqrt{3}+\frac{1}{12}\cdot\sqrt{3}+\frac{1}{12}\cdot\left(-\sqrt{3}\right)+\frac{1}{12}\cdot\left(-\sqrt{3}\right)+\frac{1}{12}\cdot 0+\frac{1}{12}\cdot 0+\frac{1}{2}\cdot 0=0. \tag{C40}$$

The ensemble first moment associated to y (i.e., the mean along y) is:

$$w_1y_1+w_2y_2+\ldots+w_7y_7=\frac{1}{12}\cdot 1+\frac{1}{12}\cdot(-1)+\frac{1}{12}\cdot 1+\frac{1}{12}\cdot(-1)+\frac{1}{12}\cdot 2+\frac{1}{12}\cdot(-2)+\frac{1}{2}\cdot 0=0. \tag{C41}$$

The ensemble second moment associated to x^2 (i.e., the variance along x) is:

$$w_1{x_1}^2+w_2{x_2}^2+\ldots+w_7{x_7}^2=\frac{1}{12}\cdot 3+\frac{1}{12}\cdot 3+\frac{1}{12}\cdot 3+\frac{1}{12}\cdot 3+\frac{1}{12}\cdot 0+\frac{1}{12}\cdot 0+\frac{1}{2}\cdot 0=1. \tag{C42}$$

The ensemble second moment associated to y^2 (i.e., the variance along y) is:

$$w_1{y_1}^2 + w_2{y_2}^2 + \ldots + w_7{y_7}^2 = \frac{1}{12}\cdot 1 + \frac{1}{12}\cdot 1 + \frac{1}{12}\cdot 1 + \frac{1}{12}\cdot 1 + \frac{1}{12}\cdot 4 + \frac{1}{12}\cdot 4 + \frac{1}{2}\cdot 0 = 1. \tag{C43}$$

The ensemble second moment associated to xy (i.e., the covariance between x and y) is:

$$\begin{aligned} &w_1x_1y_1 + w_2x_2y_2 + \ldots + w_7x_7y_7 = \\ &\quad = \frac{1}{12}\cdot\sqrt{3}\cdot 1 + \frac{1}{12}\cdot\sqrt{3}\cdot(-1) + \frac{1}{12}\cdot\left(-\sqrt{3}\right)\cdot 1 + \frac{1}{12}\cdot\left(-\sqrt{3}\right)\cdot(-1) + \frac{1}{12}\cdot 0\cdot 2 + \frac{1}{12}\cdot 0\cdot(-2) + \frac{1}{2}\cdot 0\cdot 0 = 0. \end{aligned} \tag{C44}$$

Given the comparably high number of moments to be matched, it could be useful here to introduce a result that relieves from checking some moments. In fact, if a zero-mean ensemble is symmetric (i.e., if $\boldsymbol{P}$ is a non-zero ensemble member then also $-\boldsymbol{P}$ is an ensemble member with the same weight), then all the odd moments are zero. It can be proven easily by noting that any couple of symmetric members adds zero to the moment calculation, since both of the members produce the same quantity but with opposite sign.

The ensemble (C39) is symmetric, then it remains only to prove that fourth-order moments matches the moments of the pdf.

The ensemble fourth moment associated to x^4 is:

$$w_1{x_1}^4 + w_2{x_2}^4 + \ldots + w_7{x_7}^4 = \frac{1}{12}\cdot 9 + \frac{1}{12}\cdot 9 + \frac{1}{12}\cdot 9 + \frac{1}{12}\cdot 9 + \frac{1}{12}\cdot 0 + \frac{1}{12}\cdot 0 + \frac{1}{2}\cdot 0 = 3. \tag{C45}$$

The ensemble fourth moment associated to x^3y is:

$$\begin{aligned} &w_1{x_1}^3y_1 + w_2{x_2}^3y_2 + \ldots + w_7{x_7}^3y_7 = \\ &= \frac{1}{12}\cdot\left(\sqrt{3}\right)^3\cdot 1 + \frac{1}{12}\cdot\left(\sqrt{3}\right)^3\cdot(-1) + \frac{1}{12}\cdot\left(-\sqrt{3}\right)^3\cdot 1 + \frac{1}{12}\cdot\left(-\sqrt{3}\right)^3\cdot(-1) + \frac{1}{12}\cdot 0\cdot 2 + \frac{1}{12}\cdot 0\cdot(-2) + \frac{1}{2}\cdot 0\cdot 0 = 0. \end{aligned} \tag{C46}$$

The ensemble fourth moment associated to x^2y^2 is:

$$w_1{x_1}^2{y_1}^2 + w_2{x_2}^2{y_2}^2 + \ldots + w_7{x_7}^2{y_7}^2 = \frac{1}{12}\cdot 3\cdot 1 + \frac{1}{12}\cdot 3\cdot 1 + \frac{1}{12}\cdot 3\cdot 1 + \frac{1}{12}\cdot 3\cdot 1 + \frac{1}{12}\cdot 0\cdot 4 + \frac{1}{12}\cdot 0\cdot 4 + \frac{1}{2}\cdot 0\cdot 0 = 1. \tag{C47}$$

The ensemble fourth moment associated to xy^3 is:

$$\begin{aligned} &w_1x_1{y_1}^3 + w_2x_2{y_2}^3 + \ldots + w_7x_7{y_7}^3 = \\ &\quad = \frac{1}{12}\cdot\sqrt{3}\cdot 1 + \frac{1}{12}\cdot\sqrt{3}\cdot(-1) + \frac{1}{12}\cdot\left(-\sqrt{3}\right)\cdot 1 + \frac{1}{12}\cdot\left(-\sqrt{3}\right)\cdot(-1) + \frac{1}{12}\cdot 0\cdot 8 + \frac{1}{12}\cdot 0\cdot(-8) + \frac{1}{2}\cdot 0\cdot 0 = 0. \end{aligned} \tag{C48}$$

The ensemble fourth moment associated to y^4 is:

$$w_1{y_1}^4 + w_2{y_2}^4 + \ldots + w_7{y_7}^4 = \frac{1}{12}\cdot 1 + \frac{1}{12}\cdot 1 + \frac{1}{12}\cdot 1 + \frac{1}{12}\cdot 1 + \frac{1}{12}\cdot 16 + \frac{1}{12}\cdot 16 + \frac{1}{2}\cdot 0 = 3. \tag{C49}$$

Thus, the ensemble $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4$ matches the moments of the target pdf up to order 5.

C3 Dimension 3 ($r = 3$)

The target pdf is the three-dimensional standard normal distribution $\mathcal{N}(\boldsymbol{x};\mathbf{0},\mathbf{I}_3)$. According to equations (A6) and (A7), the moments characterizing the pdf are:

- first moments:
  - first moment associated to x (i.e., the mean along x) is: $\mu_x = 0$,
  - first moment associated to y (i.e., the mean along y) is: $\mu_y = 0$,
  - first moment associated to z (i.e., the mean along z) is: $\mu_z = 0$,
- second moments:
  - second moment associated to x^2 (i.e., the variance along x) is: $\mu_{x^2} = 1$,
  - second moment associated to y^2 (i.e., the variance along y) is: $\mu_{y^2} = 1$,
  - second moment associated to z^2 (i.e., the variance along z) is: $\mu_{z^2} = 1$,
  - second moment associated to xy (i.e., the covariance between x and y) is: $\mu_{xy} = 0$,
  - second moment associated to xz (i.e., the covariance between x and z) is: $\mu_{xz} = 0$,
  - second moment associated to yz (i.e., the covariance between y and z) is: $\mu_{yz} = 0$, ...

C3.1 High-order sampling ensemble with 4 members, $h = 3$ along the xy-plane

$$
\begin{aligned}
\boldsymbol{P}_1 &: \quad (x_1,y_1,z_1) = \left(\sqrt{2},0,-\frac{1}{2}\right);\\
\boldsymbol{P}_2 &: \quad (x_2,y_2,z_2) = \left(-\sqrt{2},0,-\frac{1}{2}\right);\\
\boldsymbol{P}_3 &: \quad (x_3,y_3,z_3) = \left(0,\sqrt{2},\frac{1}{2}\right);\\
\boldsymbol{P}_4 &: \quad (x_4,y_4,z_4) = \left(0,-\sqrt{2},\frac{1}{2}\right).
\end{aligned}
\tag{C50}
$$

This tetrahedron-shaped ensemble is a second-order ensemble. It is one of the many possible outcomes of the SEIK's sampling (Pham, 2001) but, differently from other second-order ensembles, the specific orientation of this ensemble produce a square-shaped projection of its members in the xy-plane. In fact, this ensemble is an extension to three dimensions of the third-order ensemble (C22), which has members with the same x and y coordinates.

The moments in the xy-plane are already checked for ensemble (C22). Here it remains to check moments involving z. The ensemble first moment associated to z (i.e., the mean along z) is:

$$
\frac{1}{4}(z_1+z_2+z_3+z_4) = \frac{1}{4}\left(\left(-\frac{1}{2}\right)+\left(-\frac{1}{2}\right)+\frac{1}{2}+\frac{1}{2}\right) = 0.
\tag{C51}
$$

The ensemble second moment associated to z^2 (i.e., the variance along z) is:

$$\frac{1}{4}\left({z_1}^2+{z_2}^2+{z_3}^2+{z_4}^2\right)=\frac{1}{4}\left(\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4}\right)=1. \tag{C52}$$

The ensemble second moment associated to xz (i.e., the covariance between x and z) is:

$$\frac{1}{4}\left(x_1z_1+x_2z_2+x_3z_3+x_4z_4\right)=\frac{1}{4}\left(\sqrt{2}\cdot\left(-\frac{1}{2}\right)+\left(-\sqrt{2}\right)\cdot\left(-\frac{1}{2}\right)+0\cdot\frac{1}{2}+0\cdot\frac{1}{2}\right)=0. \tag{C53}$$

The ensemble second moment associated to yz (i.e., the covariance between y and z) is:

$$\frac{1}{4}\left(y_1z_1+y_2z_2+y_3z_3+y_4z_4\right)=\frac{1}{4}\left(0\cdot\left(-\frac{1}{2}\right)+0\cdot\left(-\frac{1}{2}\right)+\sqrt{2}\cdot\frac{1}{2}+\left(-\sqrt{2}\right)\cdot\frac{1}{2}\right)=0. \tag{C54}$$

Thus, the ensemble $\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3,\boldsymbol{P}_4$ matches the moments of the target pdf up to order 3 in the xy-plane, and up to order 2 elsewhere.

The present ensemble, is represented in the first two panels of Fig. 2.

C3.2 High-order sampling ensemble with 4 weighted members, $h=5$ along the x dimension and $h=3$ along the xy-plane

$$\begin{aligned}
\boldsymbol{P}_1:&\quad (x_1,y_1,z_1)=\left(\sqrt{3},0,-\sqrt{2}\right), && w_1=\frac{1}{6};\\
\boldsymbol{P}_2:&\quad (x_2,y_2,z_2)=\left(-\sqrt{3},0,-\sqrt{2}\right), && w_2=\frac{1}{6};\\
\boldsymbol{P}_3:&\quad (x_3,y_3,z_3)=\left(0,\sqrt{\frac{3}{2}},\frac{\sqrt{2}}{2}\right), && w_3=\frac{1}{3};\\
\boldsymbol{P}_3:&\quad (x_3,y_3,z_4)=\left(0,-\sqrt{\frac{3}{2}},\frac{\sqrt{2}}{2}\right), && w_3=\frac{1}{3}.
\end{aligned} \tag{C55}$$

This ensemble extends ensemble (C32) to three dimensions. In this way, it keeps the third-order in the xy-plane and the fifth-order along the x-axis. It only remains to prove the matching of moments involving z.

The ensemble first moment associated to z (i.e., the mean along z) is:

$$w_1z_1+w_2z_2+w_3z_3+w_4z_4=\frac{1}{6}\cdot\left(-\sqrt{2}\right)+\frac{1}{6}\cdot\left(-\sqrt{2}\right)+\frac{1}{3}\cdot\frac{\sqrt{2}}{2}+\frac{1}{3}\cdot\frac{\sqrt{2}}{2}=0. \tag{C56}$$

The ensemble second moment associated to z^2 (i.e., the variance along z) is:

$$w_1{z_1}^2+w_2{z_2}^2+w_3{z_3}^2+w_4{z_4}^2=\frac{1}{6}\cdot 2+\frac{1}{6}\cdot 2+\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{1}{2}=1. \tag{C57}$$

The ensemble second moment associated to xz (i.e., the covariance between x and z) is:

$$w_1x_1z_1+w_2x_2z_2+w_3x_3z_3+w_4x_4z_4=\frac{1}{6}\cdot\sqrt{3}\cdot\left(-\sqrt{2}\right)+\frac{1}{6}\cdot\left(-\sqrt{3}\right)\cdot\left(-\sqrt{2}\right)+\frac{1}{3}\cdot 0\cdot\frac{\sqrt{2}}{2}+\frac{1}{3}\cdot 0\cdot\frac{\sqrt{2}}{2}=0. \tag{C58}$$

The ensemble second moment associated to yz (i.e., the covariance between y and z) is:

$$w_1y_1z_1 + w_2y_2z_2 + w_3y_3z_3 + w_4y_4z_4 = \frac{1}{6}\cdot 0 \cdot \left(-\sqrt{2}\right) + \frac{1}{6}\cdot 0 \cdot \left(-\sqrt{2}\right) + \frac{1}{3}\cdot\sqrt{\frac{3}{2}}\cdot\frac{\sqrt{2}}{2} + \frac{1}{3}\cdot\left(-\sqrt{\frac{3}{2}}\right)\cdot\frac{\sqrt{2}}{2} = 0. \tag{C59}$$

Thus, the ensemble $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4$ matches the moments of the target pdf up to order 5 along the x-axis, up to order 3 in the xy-plane, and up to order 2 elsewhere.

The present ensemble, is represented (not in scale) in Fig. 2.

C4 Arbitrary number r of dimensions: third-order ensemble ($h = 3$) with $2r$ members

$$\begin{array}{llllllllllll}
\boldsymbol{P}_1: & (x_{1,1}, & x_{2,1}, & \ldots, & x_{r,1}) & = & (& \sqrt{r}, & 0, & 0, & \ldots, & 0); \\
\boldsymbol{P}_{-1}: & (x_{1,-1}, & x_{2,-1}, & \ldots, & x_{r,-1}) & = & (& -\sqrt{r}, & 0, & 0, & \ldots, & 0); \\
\boldsymbol{P}_2: & (x_{1,2}, & x_{2,2}, & \ldots, & x_{r,2}) & = & (& 0, & \sqrt{r}, & 0, & \ldots, & 0); \\
\boldsymbol{P}_{-2}: & (x_{1,-2}, & x_{2,-2}, & \ldots, & x_{r,-2}) & = & (& 0, & -\sqrt{r}, & 0, & \ldots, & 0); \\
\vdots & & & & & \vdots & & & & & & \\
\boldsymbol{P}_r: & (x_{1,r}, & x_{2,r}, & \ldots, & x_{r,r}) & = & (& 0, & 0, & \ldots, & 0, & \sqrt{r}); \\
\boldsymbol{P}_{-r}: & (x_{1,-r}, & x_{2,-r}, & \ldots, & x_{r,-r}) & = & (& 0, & 0, & \ldots, & 0, & -\sqrt{r}).
\end{array} \tag{C60}$$

This symmetric ensemble has $2r$ members. Thanks to the result presented in Section C2.6, the odd moments are zero. Also the covariances between different variables must be zero, since every ensemble member has only one non-zero coordinate. Finally, it only remains to prove that variances are equal to 1.

The ensemble second moment associated to the i^{th} coordinate (i.e., the variance along x_i) is:

$$\frac{1}{2r}\left({x_{i,1}}^2 + {x_{i,-1}}^2 + {x_{i,2}}^2 + {x_{i,-2}}^2 + \ldots + {x_{i,i}}^2 + {x_{i,-i}}^2 + \ldots + {x_{i,r}}^2 + {x_{i,-r}}^2\right) = \frac{1}{2r}(0 + r + r + 0) = 1. \tag{C61}$$

Thus, the ensemble $\boldsymbol{P}_1, \boldsymbol{P}_{-1}, \boldsymbol{P}_2, \boldsymbol{P}_{-2}, \ldots, \boldsymbol{P}_r, \boldsymbol{P}_{-r}$ matches the moments of the target pdf (i.e., the r-dimensional standard normal distribution $\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \mathbf{I}_r)$) up to order 3.

Author contributions. GC, S Salon and S Spada designed the study. S Spada developed the algorithms, supervised by SM. S Spada, AT and GC designed the experiments. S Spada implemented the code and performed the simulations. S Spada wrote the manuscript, with the contribution of AT, GC and CS. All the authors participated in acquiring the funding for the project.

Competing interests. The authors declare no competing interests.

Acknowledgements. We acknowledge the CINECA award under the ISCRA initiative and the HPC-TRES program, for the availability of ~~high-performance computing resources.This study has been conducted using E.U. Copernicus Marine Service data~~computing resources.
The research reported in this work was supported by OGS and by the SEAMLESS project (https://www.seamlessproject.org/).
This work was partly funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032 (project SEAMLESS).

References

Ambadan, J. T. and Tang, Y.: Sigma-Point Kalman Filter Data Assimilation Methods for Strongly Nonlinear Systems, Journal of the Atmospheric Sciences, 66, 261 – 285, https://doi.org/10.1175/2008JAS2681.1, 2009.

Anderson, J. L.: An adaptive covariance inflation error correction algorithm for ensemble filters, Tellus A: Dynamic Meteorology and Oceanography, https://doi.org/10.1111/j.1600-0870.2006.00216.x, 2007.

Bannister, R.: A review of operational methods of variational and ensemble-variational data assimilation, Quarterly Journal of the Royal Meteorological Society, 143, 607–633, https://doi.org/10.1002/qj.2982, 2017a.

Bannister, R. N.: A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics, Quarterly Journal of the Royal Meteorological Society, 134, 1971–1996, https://doi.org/10.1002/qj.340, _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.340, 2008.

Bannister, R. N.: A review of operational methods of variational and ensemble-variational data assimilation, Quarterly Journal of the Royal Meteorological Society, 143, 607–633, https://doi.org/https://doi.org/10.1002/qj.2982, 2017b.

Bishop, C. H., Etherton, B. J., and Majumdar, S. J.: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects, Monthly Weather Review, 129, 420 – 436, https://doi.org/https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2, 2001.

Bocquet, M., Pires, C., and Wu, L.: Beyond Gaussian Statistical Modeling in Geophysical Data Assimilation, Monthly Weather Review, 138, 2997–3023, https://doi.org/10.1175/2010MWR3164.1, 2010.

Bocquet, M., Raanes, P. N., and Hannart, A.: Expanding the validity of the ensemble Kalman filter without the intrinsic need for inflation, Nonlinear Processes in Geophysics, 22, 645–662, https://doi.org/10.5194/npg-22-645-2015, 2015.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L.: Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model, Journal of Computational Science, 44, 101 171, https://doi.org/https://doi.org/10.1016/j.jocs.2020.101171, 2020.

Buga, L., Sarbu, G., Fryberg, L., Magnus, W., Wesslander, K., Gatti, J., Leroy, D., Iona, S., Larsen, M., Koefoed Rømer, J., Østrem, A., Lipizer, M., and Giorgietti, A.: EMODnet Chemistry Eutrophication and Acidity aggregated datasets v2018, https://doi.org/https://doi.org/10.6092/EC8207EF-ED81-4EE5-BF48-E26FF16BF02E, 2018.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives, Wires climate change, 9, 2018.

Ciavatta, S., Kay, S., Saux-Picart, S., Butenschön, M., and Allen, J. I.: Decadal reanalysis of biogeochemical indicators and fluxes in the North West European shelf-sea ecosystem, Journal of Geophysical Research: Oceans, 121, 1824–1845, https://doi.org/https://doi.org/10.1002/2015JC011496, 2016.

Cossarini, G., Lazzari, P., and Solidoro, C.: Spatiotemporal variability of alkalinity in the Mediterranean Sea, Biogeosciences, 12(6), 1647–1658, 2015.

Cossarini, G., Mariotti, L., Feudale, L., Mignot, A., Salon, S., Taillandier, V., Teruzzi, A., and D'Ortenzio, F.: Towards operational 3D-Var assimilation of chlorophyll Biogeochemical-Argo float data into a biogeochemical model of the Mediterranean Sea, Ocean Modelling, 133, 112–128, https://doi.org/10.1016/j.ocemod.2018.11.005, 2019.

Di Biagio, V., Salon, S., Feudale, L., and Cossarini, G.: Subsurface oxygen maximum in oligotrophic marine ecosystems: mapping the interaction between physical and biogeochemical processes, Biogeosciences, 19, 5553–5574, https://doi.org/10.5194/bg-19-5553-2022, 2022.

Dovera, L. and Della Rossa, E.: Improved initial ensemble generation coupled with ensemble square root filters and inflation to estimate uncertainty, Computational Geosciences, 16, 437–453, https://doi.org/10.1007/s10596-011-9243-5, 2012.

Dowd, M., Jones, E., and Parslow, J.: A statistical overview and perspectives on data assimilation for marine biogeochemical models, Environmetrics, 25, 203–213, https://doi.org/https://doi.org/10.1002/env.2264, 2014.

Edwards, C. A., Moore, A. M., Hoteit, I., and Cornuelle, B. D.: Regional Ocean Data Assimilation, Annual Review of Marine Science, 7, 21–42, https://doi.org/10.1146/annurev-marine-010814-015821, pMID: 25103331, 2015.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, Journal of Geophysical Research: Oceans, 99, 10 143–10 162, https://doi.org/10.1029/94JC00572, 1994.

Fennel, K., Gehlen, M., Brasseur, P., Brown, C. W., Ciavatta, S., Cossarini, G., Crise, A., Edwards, C. A., Ford, D., Friedrichs, M. A. M., Gregoire, M., Jones, E., Kim, H.-C., Lamouroux, J., Murtugudde, R., Perruche, C., , t. G. O. M. E. A., and Team, P. T.: Advancing Marine Biogeochemical and Ecosystem Reanalyses and Forecasts as Tools for Monitoring and Managing Ecosystem Health, Frontiers in Marine Science, 6, https://doi.org/10.3389/fmars.2019.00089, 2019.

Fennel, K., Mattern, J. P., Doney, S. C., Bopp, L., Moore, A. M., Wang, B., and Yu, L.: Ocean biogeochemical modelling, Nature Reviews Methods Primers, 2, 76, 2022.

Fertig, E. J., Harlim, J., and Hunt, B. R.: A comparative study of 4D-VAR and a 4D Ensemble Kalman Filter: perfect model simulations with Lorenz-96, Tellus A: Dynamic Meteorology and Oceanography, https://doi.org/10.1111/j.1600-0870.2006.00205.x, 2007.

Ford, D. and Barciela, R.: Global marine biogeochemical reanalyses assimilating two different sets of merged ocean colour products, Remote Sensing of Environment, 203, https://doi.org/10.1016/j.rse.2017.03.040, 2017.

Ford, D. A.: Assessing the role and consistency of satellite observation products in global physical–biogeochemical ocean reanalysis, Ocean Science, 16, 875–893, https://doi.org/https://doi.org/10.5194/os-16-875-2020, publisher: Copernicus GmbH, 2020.

Foujols, M.-A., Lévy, M., Aumont, O., and Madec, G.: OPA 8.1 Tracer Model reference manual, Project report, Institut Pierre-Simon Laplace (IPSL), https://eprints.soton.ac.uk/64848/, 2000.

Gaspari, G. and Cohn, S. E.: Construction of correlation functions in two and three dimensions, Quarterly Journal of the Royal Meteorological Society, 125, 723–757, https://doi.org/https://doi.org/10.1002/qj.49712555417, 1999.

Gharamti, M. E.: Enhanced Adaptive Inflation Algorithm for Ensemble Filters, Monthly Weather Review, 146, 623 – 640, https://doi.org/https://doi.org/10.1175/MWR-D-17-0187.1, 2018.

Grooms, I.: A comparison of nonlinear extensions to the ensemble Kalman filter, Computational Geosciences, 26, 1–18, https://doi.org/10.1007/s10596-022-10141-x, 2022.

Grooms, I. and Robinson, G.: A hybrid particle-ensemble Kalman filter for problems with medium nonlinearity, PLOS ONE, 16, 1–20, https://doi.org/10.1371/journal.pone.0248266, 2021.

Hamill, T. M. and Snyder, C.: A Hybrid Ensemble Kalman Filter–3D Variational Analysis Scheme, Monthly Weather Review, 128, 2905 – 2919, https://doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2, 2000.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Hodyss, D.: Ensemble State Estimation for Nonlinear Systems Using Polynomial Expansions in the Innovation, Monthly Weather Review, 139, 3571 – 3588, https://doi.org/10.1175/2011MWR3558.1, 2011.

Hoteit, I., Pham, D.-T., Triantafyllou, G., and Korres, G.: Particle Kalman Filtering for Data Assimilation in Meteorology and Oceanography, in: 3rd WCRP International Conference on Reanalysis, pp. 1–6, Tokyo, Japan, https://hal.science/hal-00853919, 2008.

Houtekamer, P. L. and Zhang, F.: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation, Monthly Weather Review, 144, 4489 – 4532, https://doi.org/https://doi.org/10.1175/MWR-D-15-0440.1, 2016.

Hu, J., Fennel, K., Mattern, J. P., and Wilkin, J.: Data assimilation with a local Ensemble Kalman Filter applied to a three-dimensional biological model of the Middle Atlantic Bight, Journal of Marine Systems, 94, 145–156, https://doi.org/https://doi.org/10.1016/j.jmarsys.2011.11.016, 2012.

Ito, K. and Xiong, K.: Gaussian filters for nonlinear filtering problems, IEEE Transactions on Automatic Control, 45, 910–927, https://doi.org/10.1109/9.855552, 2000.

Janjic, T., Nerger, L., Albertella, A., Schroter, J., and Skachko, S.: On Domain Localization in Ensemble-Based Kalman Filter Algorithms, Monthly Weather Review, 139, 2046–2060, 2011.

Janjić, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., Losa, S. N., Nichols, N. K., Potthast, R., Waller, J. A., and Weston, P.: On the representation error in data assimilation, Quarterly Journal of the Royal Meteorological Society, 144, 1257–1278, https://doi.org/https://doi.org/10.1002/qj.3130, 2018.

Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems, Transaction of the ASME - Journal of Basic Engineering, pp. 35–45, 1960.

Kirchgessner, P., Nerger, L., and Bunse-Gerstner, A.: On the Choice of an Optimal Localization Radius in Ensemble Kalman Filter Methods, Monthly Weather Review, 142, 2165 – 2175, https://doi.org/10.1175/MWR-D-13-00246.1, 2014.

Lahoz, W. A. and Schneider, P.: Data assimilation: making sense of Earth Observation, Frontiers in Environmental Science, 2, https://doi.org/10.3389/fenvs.2014.00016, 2014.

Lazzari, P., Teruzzi, A., Salon, S., Campagna, S., Calonaci, C., Colella, S., Tonani, M., and Crise, A.: Pre-operational short-term forecasts for the Mediterranean Sea biogeochemistry, Ocean Science, 6, 25–39, 2010.

Lazzari, P., Solidoro, C., Ibello, V., Salon, S., Teruzzi, A., Béranger, K., Colella, S., and Crise, A.: Seasonal and inter-annual variability of plankton chlorophyll and primary production in the Mediterranean Sea: a modelling approach, Biogeosciences, 9, 217–233, 2012.

Lazzari, P., Solidoro, C., Salon, S., and Bolzon, G.: Spatial variability of phosphate and nitrate in the Mediterranean Sea: A modeling approach., Deep Sea Research Part I: Oceanographic Research, 108, 39 – 52, https://doi.org/https://doi.org/10.1016/j.dsr.2015.12.006, 2016.

Lei, J. and Bickel, P.: A Moment Matching Ensemble Filter for Nonlinear Non-Gaussian Data Assimilation, Monthly Weather Review, 139, 3964 – 3973, https://doi.org/10.1175/2011MWR3553.1, 2011.

Lorenz, E. N.: Predictability: A problem partly solved, in: Proc. Seminar on predictability, vol. 1, Reading, 1996.

Luo, X. and Moroz, I.: Ensemble Kalman filter with the unscented transform, Physica D: Nonlinear Phenomena, 238, 549–562, https://doi.org/https://doi.org/10.1016/j.physd.2008.12.003, 2009.

Martin, M., Balmaseda, M., Bertino, L., Brasseur, P., Brassington, G., Cummings, J., Fujii, Y., Lea, D., Lellouche, J.-M., Mogensen, K., Oke, P., Smith, G., Testut, C.-E., Waagbø, G., Waters, J., and A.T., W.: Status and future of data assimilation in operational oceanography, Journal of Operational Oceanography, 8, s28–s48, https://doi.org/10.1080/1755876X.2015.1022055, 2015.

Mastroianni, G. and Milovanovic, G. V.: Interpolation Processes: Basic Theory and Applications, Springer Publishing Company, Incorporated, 1 edn., 2008.

Mattern, J. P., Edwards, C. A., and Moore, A. M.: Improving Variational Data Assimilation through Background and Observation Error Adjustments, Monthly Weather Review, 146, 485–501, https://doi.org/10.1175/MWR-D-17-0263.1, 2017.

Melaku Canu, D., Ghermandi, A., Nunes, P., Lazzari, P., Cossarini, G., and Solidoro, C.: Estimating the value of carbon sequestration ecosystem services in the Mediterranean Sea: An ecological economics approach, Global Environmental Change, 32, 87–95, 2015.

Moore, A. M., Martin, M. J., Akella, S., Arango, H. G., Balmaseda, M., Bertino, L., Ciavatta, S., Cornuelle, B., Cummings, J., Frolov, S., Lermusiaux, P., Oddo, P., Oke, P. R., Storto, A., Teruzzi, A., Vidard, A., and Weaver, A. T.: Synthesis of Ocean Observations Using Data Assimilation for Operational, Real-Time and Reanalysis Systems: A More Complete Picture of the State of the Ocean, Frontiers in Marine Science, 6, https://doi.org/10.3389/fmars.2019.00090, 2019.

Nerger, L.: Data assimilation for nonlinear systems with a hybrid nonlinear Kalman ensemble transform filter, Quarterly Journal of the Royal Meteorological Society, 148, 620–640, https://doi.org/https://doi.org/10.1002/qj.4221, 2022.

Nerger, L. and Gregg, W.: Assimilation of SeaWiFS data into a global ocean-biogeochemical model using a local SEIK filter, Journal of Marine Systems, 68, 237–254, https://doi.org/10.1016/j.jmarsys.2006.11.009, 2007.

Nerger, L., Hiller, W., and Schröter, J.: A comparison of error subspace Kalman filters, Tellus A, 57, 715–735, https://doi.org/https://doi.org/10.1111/j.1600-0870.2005.00141.x, 2005.

Nerger, L., Janjic Pfander, T., Schröter, J., and Hiller, W.: A Unification of Ensemble Square Root Kalman Filters, Monthly Weather Review, 140, 2335–2345, https://doi.org/10.1175/MWR-D-11-00102.1, 2012.

Pham, D. T.: A Singular Evolutive Interpolated Kalman filter for data assimilation in Oceanography, 1996.

Pham, D. T.: Stochastic Methods for Sequential Data Assimilation in Strongly Nonlinear Systems, Monthly Weather Review, 129, 1194 – 1207, https://doi.org/10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2, 2001.

Pham, D. T., Verron, J., and Roubaud, M. C.: A Singular Evolutive Extended Kalman filter for data assimilation in Oceanography, Journal of Marine Systems, 16, 323–340, 1998.

Pradhan, H. K., Völker, C., Losa, S. N., Bracher, A., and Nerger, L.: Assimilation of Global Total Chlorophyll OC-CCI Data and Its Impact on Individual Phytoplankton Fields, Journal of Geophysical Research: Oceans, 124, 470–490, https://doi.org/10.1029/2018JC014329, 2019.

Raanes, P. N., Bocquet, M., and Carrassi, A.: Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures, Quarterly Journal of the Royal Meteorological Society, 145, 53–75, https://doi.org/https://doi.org/10.1002/qj.3386, 2019.

Rainwater, S. and Hunt, B. R.: Ensemble data assimilation with an adjusted forecast spread, Tellus A: Dynamic Meteorology and Oceanography, https://doi.org/10.3402/tellusa.v65i0.19929, 2013.

Roth, M., Hendeby, G., Fritsche, C., and Gustafsson, F.: The Ensemble Kalman Filter: A Signal Processing Perspective, EURASIP Journal on Advances in Signal Processing, 2017, 56, https://doi.org/10.1186/s13634-017-0492-x, 2017.

Sacher, W. and Bartello, P.: Sampling Errors in Ensemble Kalman Filtering. Part I: Theory, Monthly Weather Review, 136, 3035 – 3049, https://doi.org/https://doi.org/10.1175/2007MWR2323.1, 2008.

Salon, S., Cossarini, G., Bolzon, G., Feudale, L., Lazzari, P., Teruzzi, A., Solidoro, C., and Crise, A.: Novel metrics based on Biogeochemical Argo data to improve the model uncertainty evaluation of the CMEMS Mediterranean marine ecosystem forecasts, Ocean Science, 15, 997–1022, https://doi.org/https://doi.org/10.5194/os-15-997-2019, publisher: Copernicus GmbH, 2019.

Santana-Falcón, Y., Brasseur, P., Brankart, J. M., and Garnier, F.: Assimilation of chlorophyll data into a stochastic ensemble simulation for the North Atlantic Ocean, Ocean Science, 16, 1297–1315, https://doi.org/https://doi.org/10.5194/os-16-1297-2020, publisher: Copernicus GmbH, 2020.

Scheffler, G., Carrassi, A., Ruiz, J., and Pulido, M.: Dynamical effects of inflation in ensemble-based data assimilation under the presence of model error, Quarterly Journal of the Royal Meteorological Society, 148, 2368–2383, https://doi.org/https://doi.org/10.1002/qj.4307, 2022.

Song, H., Edwards, C. A., Moore, A. M., and Fiechter, J.: Data assimilation in a coupled physical-biogeochemical model of the California current system using an incremental lognormal 4-dimensional variational approach: Part 3—Assimilation in a realistic context using satellite and in situ observations, Ocean Modelling, 106, 159–172, https://doi.org/10.1016/j.ocemod.2016.06.005, 2016.

Spada, S., Teruzzi, A., Maset, S., Salon, S., Solidoro, C., and Cossarini, G.: GHOSH v1.0.0: a novel Gauss-Hermite High-Order Sampling Hybrid ensemble filter for computationally efficient data assimilation in geosciences - Part 2, https://doi.org/10.5194/gmd-2023-170, preprint version, 2024.

Stordal, A., Karlsen, H., Nævdal, G., Skaug, H., and Vallès, B.: Bridging the ensemble Kalman filter and particle filters: The adaptive Gaussian mixture filter, Computational Geosciences, 15, https://doi.org/10.1007/s10596-010-9207-1, 2011.

Talagrand, O. and Courtier, P.: Variational Assimilation of Meteorological Observations With the Adjoint Vorticity Equation. I: Theory, Quarterly Journal of the Royal Meteorological Society, 113, 1311–1328, https://doi.org/https://doi.org/10.1002/qj.49711347812, 1987.

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M., and Zhen, Y.: A Review of Innovation-Based Methods to Jointly Estimate Model and Observation Error Covariance Matrices in Ensemble Data Assimilation, Monthly Weather Review, 148, 3973–3994, https://doi.org/10.1175/mwr-d-19-0240.1, 2020.

Teruzzi, A., Dobricic, S., Solidoro, C., and Cossarini, G.: A 3-D variational assimilation scheme in coupled transport-biogeochemical models: Forecast of Mediterranean biogeochemical properties, J. Geophys. Res. Oceans, 119, 200–217, 2014.

Teruzzi, A., Bolzon, G., Salon, S., Lazzari, P., Solidoro, C., and Cossarini, G.: Assimilation of coastal and open sea biogeochemical data to improve phytoplankton simulation in the Mediterranean Sea, Ocean Modelling, 132, 46 – 60, https://doi.org/https://doi.org/10.1016/j.ocemod.2018.09.007, 2018a.

Teruzzi, A., Bolzon, G., Salon, S., Lazzari, P., Solidoro, C., and Cossarini, G.: Assimilation of coastal and open sea biogeochemical data to improve phytoplankton simulation in the Mediterranean Sea, Ocean Modelling, 132, 46–60, https://doi.org/10.1016/j.ocemod.2018.09.007, 2018b.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., and Whitaker, J. S.: Ensemble Square Root Filters, Monthly Weather Review, 131, 1485 – 1490, https://doi.org/10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2, 2003.

Triantafyllou, G., Hoteit, I., and Petihakisa, G.: A singular evolutive interpolated Kalman filter for efficient data assimilation in a 3-D complex physical-biogeochemical model of the Cretan Sea, Journal of Marine Systems, 40–41, 213–231, 2003.

Tsiaras, K., Hoteit, I., Kalaroni, S., Petihakis, G., and Triantafyllou, G.: A hybrid ensemble-OI Kalman filter for efficient data assimilation into a 3-D biogeochemical model of the Mediterranean, Ocean Dynamics, 67, 673–690, https://doi.org/10.1007/s10236-017-1050-7, 2017.

Tödter, J. and Ahrens, B.: A Second-Order Exact Ensemble Square Root Filter for Nonlinear Data Assimilation, Tödter and Ahrens, 140, 1347–1369, 2015.

van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., and Reich, S.: Particle filters for high-dimensional geoscience applications: A review, Quarterly Journal of the Royal Meteorological Society, 145, 2335–2365, https://doi.org/https://doi.org/10.1002/qj.3551, 2019.

Vetra-Carvalho, S., Van Leeuwen, P. J., Nerger, L., Barth, A., Altaf, M. U., Brasseur, P., Kirchgessner, P., and Beckers, J.-M.: State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems, Tellus A: Dynamic Meteorology and Oceanography, https://doi.org/10.1080/16000870.2018.1445364, 2018.

Volpe, G., Pitarch, J., Colella, S., Brando, V., Forneris, V., Bracaglia, M., and Benincasa, M.: OCEAN COLOUR PRODUCTION CENTRE, QUality Information Document of Ocean Colour Mediterranean and Black Sea Observation Product v1.5. Technical Report., 2017.

Wan, E. and Van Der Merwe, R.: The unscented Kalman filter for nonlinear estimation, in: Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373), pp. 153–158, https://doi.org/10.1109/ASSPCC.2000.882463, 2000.

GHOSH v1.0.0: a novel Gauss-Hermite High-Order Sampling Hybrid ensemble filter for computationally efficient data assimilation in geosciences - Part 2

Simone Spada[1], Anna Teruzzi[1], Stefano Maset[2], Stefano Salon[1], Cosimo Solidoro[1], and Gianpiero Cossarini[1]

[1]National Institute of Oceanography and Applied Geophysics - OGS
[2]University of Trieste

Correspondence: Simone Spada (sspada@ogs.it)

Abstract. Data assimilation ~~is used in a number of geophysical applications to optimally integrate observations and model knowledge. Providing an estimation of both state and uncertainty, ensemble algorithms are one of the most successful data assimilation approaches. Since the estimation quality depends on the ensemble, the sampling method is a crucial step in ensemble data assimilation. Among other options to improve the capability of generating an effective ensemble, a sampling method featuring a higher polynomial order of approximation represents a novelty. Indeed, the order of the most widespread ensemble algorithms is usually equal to or lower than 2. We propose a novel hybrid~~ (DA) methodologies play a crucial role in integrating observational data with numerical models to provide accurate estimates of geophysical systems. A novel ensemble algorithm, the Gauss-Hermite High-Order Sampling Hybrid (GHOSH) filter, version 1.0.0, ~~which we apply in a twin experiment (based on Lorenz96) and in a realistic geophysical application. In the most error components, the GHOSH sampling method can achieve a higher order of approximation~~ aims to improve DA performances by featuring an approximation order higher than in other ensemble ~~based filters . To evaluate the benefits of the higher approximation order, a set of thousands twin experiments of Lorenz96 simulations has been carried out using the GHOSH filter and a second order ensemble Kalman filter(SEIK; singular evolutive interpolated Kalman filter). The comparison between the GHOSH and the SEIK filter has been done by varying a number of data assimilation settings: ensemble size, inflation, assimilated observations, and initial conditions. The twin-experiment results show that GHOSH outperforms SEIK in most of the assimilation settings up to a 69% reduction of the root mean square error on assimilated and non-assimilated variables. A~~ DA filters like SEIK (Singular Evolutive Interpolated Kalman filter) or ETKF (Ensemble Transform Kalman Filter). Building upon the successful application in idealized twin experiments (Part 1), we propose a parallel implementation of the GHOSH filter which has been coupled with a realistic geophysical application: the assimilation of surface satellite chlorophyll in a physical-biogeochemical model of the Mediterranean Sea ~~with assimilation of surface satellite chlorophyll .~~ (OGSTM-BFM).

This application was selected due to its complexity and potential relevance to operational settings, such as those of the Copernicus Marine Service, where the uncertainty of prediction can also be crucial. By assimilating satellite chlorophyll data, GHOSH aims to improve the accuracy of the state estimation, not only for the assimilated variable but also for other non-assimilated variables of interest (nutrients), thereby enhancing our understanding of marine ecosystem dynamics.

The simulation results are validated using both semi-independent (satellite chlorophyll) and independent (nutrient concentrations from an in-situ climatology) observations. Results show ~~the feasibility of~~ that: i) the GHOSH implementation in a realistic three-dimensional application ~~. The~~ is just as computationally feasible as other ensemble assimilation filters, since the integration of a realistic model is by far more computationally expensive than the assimilation scheme; ii) the GHOSH assimilation algorithm ~~improves~~ improved the agreement between forecasts and observations without producing unrealistic effects on the non-assimilated variables. ~~Furthermore, the sensitivity analysis on GHOSH setup indicates that the use of a higher order of convergence substantially improves the performance of the assimilation with respect to nitrate (i.e., one of the non-assimilated variables). In view of potential implementation~~ Moreover, a sensitivity analysis revealed correlation between higher order and a key nutrient (nitrate) estimation improvement, highlighting the importance of the GHOSH ~~filter in operational applications, it should be noted that GHOSH and SEIK filters have not shown significant differences in terms of time to solution, since, as in all ensemble-like Kalman filters, the model integration is by far more computationally expensive than the assimilation scheme.~~ enhanced order of approximation in boosting DA performances on non-assimilated variables.

1 Introduction

~~Data assimilation (DA) methodologies provide a conceptual framework to integrate the information contents embedded in observations and numerical models, and plays a pivotal role in deriving accurate estimates of the state of earth systems components and reducing the uncertainties of geophysical systems (Bannister (2017a); Houtekamer and Zhang (2016); Edwards et al. (201). In fact, the use of DA has been steadily increasing in the last decade and it is now ubiquitous in earth sciences, also thanks to the unprecedented -and always increasing- availability of both data and computational capability.~~

~~The Kalman filters and the variational approaches are the two main classes of DA algorithms (Kalman (1960), Talagrand and Courtier (19)). Both approaches are based on Gaussian approximation and can be interpreted in a Bayesian framework, and include an estimate of the most probable state and its evolution, and possibly the associated variance. Both methods are strictly valid only for linear models, but modified implementations have been proposed for application to non-linear ones (Carrassi et al. (2018)).~~

~~Recently, thanks to the scalability of parallel implementations, the use of ensemble algorithms (e.g., EnKF Evensen (1994) and EnVar Bannister (2017b)) have been proposed to estimate uncertainty and improve assimilation skills in Kalman filters and variational methodologies. Furthermore, some of the strong points of the two approaches have been merged, leading to hybrid filters (e.g., Hamill and Snyder (2000)) .~~

~~However, the definition of the strategy for ensemble generation is not a trivial task (Moore et al. (2019)). Straightforward Montecarlo approaches are not usually a viable option in geoscience applications, because they would require a too large number of ensemble members, and consequently, computational effort. The number of ensemble members can be reduced by adopting deterministic (or semi-deterministic) sampling methods. These methods provide an ensemble mean of future state with no error after evolving the ensemble, as long as the evolution function (i.e., the model) is a polynomial of the same order~~

~~of the sampling method. Indeed, the use of deterministic or semi-deterministic ensembles, built using second-order sampling methods, are now quite common (as in SEIK, Pham (1996), or ETKF, Bishop et al. (2001)).~~

~~At the same time, most of the models used in geoscience applications are based on systems of differential equations that cannot be represented by a second order polynomial and in all of these cases the second order sampling methods provide only a second order approximation of the model . Furthermore, the second order approximation is more effective the closer the ensemble members are to each other, thus, the larger the ensemble spread the worse will be the approximation error in the mean computation. Since a relatively large ensemble spread is often required in data assimilation applications, this approximation error may be not negligible.~~

~~A potential strategy to reduce this error is the use of a higher order of approximation, but this would require a larger ensemble with respect to the second order case and consequently bigger computational costs, given that a higher order of approximation implies a larger number of ensemble members to represent the same error subspace. For instance, second order methods use an ensemble of $r+1$ members to span an r-dimensional error subspace (e.g., Pham (1996)), while it can be proven that $2r$ ensemble members are needed to achieve order 3 in the same subspace (an example of 3rd order methodwith $2r+1$ ensemble members is the unscented Kalman filter , Wan and Van Der Merwe (2000)).~~

~~The approximation order can be increased further by using properly weighted ensemble members. A weighted ensemble based on multi-dimensional~~ The Gauss-Hermite ~~quadrature rule (Mastroianni and Milovanovic (2008)) could arbitrarily increase the order of approximation, but always at the cost of a larger ensemble size (Ito and Xiong (2000)). Weighted ensembles are also exploited in particle filter methods (Bocquet et al. (2010)), where weights are used to evolve both the ensemble members and their probability~~High-Order Sampling Hybrid (GHOSH) filter, introduced in Part 1 (Spada et al., 2024), presents a promising advancement in data assimilation (DA) methodologies, particularly for addressing the challenges posed by non-linearities and model complexity in geophysical systems. Compared to a traditional second-order method, the GHOSH filter improves the accuracy of the state estimation in systems governed by complex dynamic.

In ~~the present work, we propose a novel weighted ensemble method based on a new high-order sampling, that provides ensemble mean estimates of order higher than 2 without increasing the number of ensemble members, and therefore the computational cost.~~

~~The proposed high-order filter (as its sampling method) exploits a Gauss-Hermite-like quadrature rule, along with a principal component analysis, and we named it Gauss-Hermite high-order sampling hybrid (GHOSH) filter. Members and related weights are chosen in such a way as to guarantee accuracy of order higher than 2 for a subset of relevant modes (a relevant subspace), and no less than 2 for the remaining ones. The GHOSH filter exploits a hybrid approach that considers the model uncertainty described by a constant and time-evolving part based on the evolution of an ensemble.~~

~~The~~ Part 1, the reliability of the new filter ~~is proven first~~ has been proven in a large set of twin experiments based on an idealized model commonly used to test DA methodologies ~~(Lorenz96, Lorenz (1996)) and then in~~ (Lorenz96 Lorenz, 1996). In the present work, we extend the application of the GHOSH filter to the much more complex and realistic marine biogeochemical application (the OGSTM-BFM) currently used in ~~Copernicus Marine CMEMS System (Salon et al. (2019))~~the Copernicus Marine Service (Salon et al., 2019), featuring assimilation of satellite observations.

~~The high non-linearity of the biogeochemical processes makes the chosen application a challenging framework suitable for a GHOSH test implementation. Indeed, biogeochemical DA~~Biogeochemical data assimilation (DA) is relatively recent both in terms of methods and assimilated variables ~~(Fennel et al. (2019))~~(Fennel et al., 2019). Surface chlorophyll concentration, which is estimated by satellite ocean colour observations, is the most commonly assimilated ~~biogeochemcial~~ biogeochemical variable since its relatively high coverage and its near real-time availability make it suitable also for operational applications ~~(e.g., Song et al. (2016); Tsiaras et al. (2017); Santana-Falcón et al. (2020); Pradhan et al. (2019); Teruzzi et al. (2018b); Ciavatta et al. (20~~ ~~).~~

~~Section **??** introduces the novel elements of the high-order sampling and Section **??** presents a synthetic description of the GHOSH algorithm and its localized version. Additional mathematical details are provided in Appendix **??**. The two sets of numerical experiments (a twin experiment of Lorenz96 model and a realistic 3D marine biogeochemical application) are introduced (Section **??**) and their results presented (Section **??**). The algorithm and the experimental results are discussed in Section 4.~~

2 ~~The high-order sampling~~

~~4 tetrahedron-shaped ensemble members (yellow dots) in a 3-dimensional space (top) represent a second-order approximation of a standard normal distribution (in shadow the spherical isosurfaces). The same 4 members form a square in 2 dimensions (middle), producing a third-order approximation. By projecting further into 1 direction (bottom), they become a fifth-order weighted ensemble.~~

~~In this work we propose a novel method to sample the ensemble members in data assimilation applications. Compared to deterministic square-root sampling methods (Carrassi et al. (2018)), the novel strategy achieves a higher order of approximation without increasing ensemble size. The strategy is based on the choice, among all the second order ensembles, of those that feature a higher order of approximation along the most relevant components of the error (i.e., those directions where the ensemble spread is larger). To better understand the idea, Fig. **??** shows an error represented by a 3-dimensional standard normal distribution (in shadow the spherical errors isosurface). Using the second order exact sampling (Pham (1996)), 4 ensemble members (yellow points) are needed to represent the uncertainty produced by this error distribution. In a xyz Cartesian coordinate system, the 4 ensemble members are the vertices of a regular tetrahedron, and all the possible second order samplings correspond to a rotation of this tetrahedron. With the appropriate rotation, the vertices draw (project) a square in the xy plane. In this way, we achieved a third order approximation in the xy plane, because it can be proved that 4 square-shaped ensemble members provide a third order approximation for a 2-dimensional normal distribution. Similarly, a further projection on a single axis returns 4 points or, as in the figure, only 3 points, one of which represents 2 of the original points. In the 1D case, the highest approximation order that can be obtained using 4 members is still limited to 3, unless proper weights are assigned to each ensemble member. In fact, 4 weighted members (or even only 3) are sufficient to reach an approximation order as high as 5 on the x axis. Finally, what we have obtained is a 3-dimensional second order weighted ensemble which is a third-order approximation in the xy plane and a fifth order approximation along the x axis.~~

~~This means that, if the z component of the distribution was less relevant than the x and y components (e.g., a non-standard normal distribution with smaller variance on the z coordinate), then we obtained a third order approximation in the dimensions where the spread is higher. And, if the variance is bigger on x than y, then we achieved a fifth order approximation where the spread is maximum.~~

~~The idea described above can be generalized considering that any probability distribution (under reasonable hypothesis) can be sampled with an arbitrary high order of approximation with a weighted ensemble that, in the special case of the Gaussian distribution, is represented by the nodes and weights of the Gauss-Hermit quadrature rule (Mastroianni and Milovanovic (2008)). Once the approximation order (h larger than 2) and the ensemble size ($r+1$) are fixed, the h-order weighted ensemble is computed for a relatively small s-dimensional subspace (s lower than r); then, a new $(r+1)$-sized second order weighted ensemble is computed such that its projections along the most relevant error directions coincide with the h-order ensemble. The resulting ensemble has a second order approximation in the r-dimensional subspace and a h-order approximation in the s-dimensional subspace of the most relevant error directions. By comparison, a second order exact sampling (e.g., SEIK) with $r+1$ ensemble members would provide a second order approximation for the whole r-dimensional subspace.~~

~~The high-order sampling is described in the following from an algorithmic point of view. Mathematical details, instead, can be found in Appendix ??.~~(e.g., Song et al., 2016; Tsiaras et al., 2017; Santana-Falcón et al., 2020; Pradhan et al., 2019; Teruzzi et al., 2018 . The ~~high-order sampling presents two phases: *initialization* and *sampling*. The former includes the steps that must be executed only once, while the latter presents the ensemble generation procedure.~~

1.1 ~~Initialization~~

~~First, 3 hyper-parameters r, h and s are chosen: r is the dimension of the error subspace, which implies an ensemble size of $r+1$ members; h is the higher approximation order reached in the most relevant directions; s represents the number of principal components that are approximated with order h and must be smaller than r.Then, the real numbers u_j^m and v_j, with $j \in \{1, \ldots, s\}$ and $m \in \{1, \ldots, r+1\}$, are calculated as a solution of the nonlinear system~~

$$\begin{cases} \vdots \\ \sum_{m=1}^{r+1} u_{j_1}^m \cdots u_{j_\xi}^m {v_m}^{2-\xi} = \mu_{j_1,\ldots,j_\xi}, \\ \vdots \end{cases}$$

~~with one equation for each $\xi \in \{0, \ldots, h\}$ and for each $j_1, \ldots, j_\xi \in \{1, \ldots, s\}$. $\mu_{j_1,\ldots,j_\xi}$ are the statistical moments of a s-dimensional probability distribution that approximates the assumed error distribution shape. Such probability distribution must be uncorrelated, normalized and have 0 mean. A typical choice is the Normal distribution $\mathcal{N}(\boldsymbol{z}; 0, \mathbf{I}_s)$, i.e.,~~

$$\mu_{j_1,\ldots,j_\xi} = \int_{\mathbb{R}^s} z_{j_1} \cdots z_{j_\xi} \mathcal{N}(\boldsymbol{z}; 0, \mathbf{I}_s)\, d\boldsymbol{z}.$$

~~In the Gaussian case, the moments $\mu_{j_1,\ldots,j_\xi}$, which are other hyper-parameters of the GHOSH algorithm, can easily be computed (i.e., without solving numerically the integral, see Appendix ??, equations (??) and (??)). Changing the value of the moments $\mu_{j_1,\ldots,j_\xi}$ allows to explore other non-Gaussian probability distributions.Note that, in order to actually have at least a solution of the system, the number of independent equations cannot be bigger than the number of variables, which implies that a bigger h requires a smaller s or a bigger r~~high non-linearity of the biogeochemical processes makes it difficult to properly estimate the cross-correlations between variables, thus degradation of non-assimilated variables is often a problematic issue (see e.g., Santana-Falcón et al., 2020; Ford, 2020; Fowler et al., 2023; Mamnun et al., 2022; Wang et al., 2020).

~~The system solution is used to define $\boldsymbol{\Omega}_h$, as the $(r+1)\times s$ matrix with coordinates u_j^m, and the ensemble weight vector $\boldsymbol{w}\in\mathbb{R}^{(r+1)}$, as the element-wise square of $\boldsymbol{v}$, i.e.,~~

$$\boldsymbol{w} = \left({v_1}^2, \ldots, {v_{r+1}}^2\right).$$

1.1 ~~Sampling~~

~~Given an N-dimensional state space, in the sampling phase a new ensemble is generated based on the ensemble mean $\boldsymbol{x}_i\in\mathbb{R}^N$ and the covariance matrix $\mathbf{L}_i\mathbf{L}_i^T$. Differently from non-varying quantities (like those defined in the *initialization* phase), i-indexed quantities are specific of each sampling. The columns of the $N\times r$ matrix $\mathbf{L}_i$ adopted to obtain a r-rank factorization of the covariance matrix represent the base of the error subspace spanned by the ensemble members.~~

~~A random $s\times s$ orthogonal matrix $\boldsymbol{\Omega}_i^{rnd}$ is drawn, then the $(r+1)\times(s+1)$ matrix $\left(\begin{array}{c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h\boldsymbol{\Omega}_i^{rnd}\end{array}\right)$ is completed to an $(r+1)\times(r+1)$ orthogonal matrix by the $(r+1)\times(r-s)$ matrix $\boldsymbol{\Omega}_i^\perp$, i.e.,~~

$$\left(\begin{array}{c|c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h\boldsymbol{\Omega}_i^{rnd} & \boldsymbol{\Omega}_i^\perp \end{array}\right)^T \left(\begin{array}{c|c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h\boldsymbol{\Omega}_i^{rnd} & \boldsymbol{\Omega}_i^\perp \end{array}\right) = \mathbf{I}_{r+1},$$

~~where $\boldsymbol{v}$ is the array of the square roots of the weights (as in equation (??)) and $\mathbf{I}_{r+1}$ is the identity matrix of rank $r+1$. Many procedures (e.g., Pham (1996)) are available to generate a random orthogonal matrix (or to randomly complete a given set of orthonormal vectors to an orthogonal matrix). Now, the $(r+1)\times r$ matrix $\boldsymbol{\Omega}_i$, defined as~~

$$\boldsymbol{\Omega}_i = \left(\begin{array}{c|c} \boldsymbol{\Omega}^h\boldsymbol{\Omega}_i^{rnd} & \boldsymbol{\Omega}_i^\perp \end{array}\right),$$

~~can be used to build the $N\times r$ ensemble matrix $\mathbf{X}_i$:~~

$$\mathbf{X}_i = \boldsymbol{x}_i\mathbb{1}_{1\times(r+1)} + \mathbf{L}_i{\mathbf{C}_i}^T{\boldsymbol{\Omega}_i}^T\mathbf{W}^{-\frac{1}{2}},$$

~~where $\mathbb{1}$ is a matrix (of the subscripted size) filled with ones, $\mathbf{W}=\mathrm{diag}(\boldsymbol{w})$ is the $(r+1)\times(r+1)$ diagonal matrix of weights and $\mathbf{C}_i$ is a $r\times r$ orthogonal change-of-basis matrix such that~~

$$\mathbf{L}_i{}^T \boldsymbol{\Lambda}_i{}^{-1} \mathbf{L}_i = \mathbf{C}_i^T \mathbf{E}_i \mathbf{C}_i.$$

~~In the last equation, the right-hand side is an eigenvalue decomposition of the left-hand side, with $\mathbf{E}_i$ being an $r \times r$ diagonal matrix of eigenvalues in descending order and $\boldsymbol{\Lambda}_i$ being an appropriate $N \times N$ matrix. The purpose of $\boldsymbol{\Lambda}_i$ is to weight the product between $\mathbf{L}_i$ and its transposed and therefore to define the criterion that decides the relevance of each PCA component. In the simplest cases, $\boldsymbol{\Lambda}_i$ can be the identical matrix, but in most scenarios, $\boldsymbol{\Lambda}_i$ needs to be properly designed (see sections **??** and 4).After $\mathbf{X}_i$ computation, the ensemble members can be retrieved from the column of the matrix, i.e.,~~

$$\mathbf{X}_i = \left(\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^{(r+1)}\right).$$

Hence, given its complexity and relevance in the operational framework, the OGSTM-BFM application serves as a critical testbed for evaluating the GHOSH filter's performance in a real-world scenario.

2 ~~The GHOSH filter~~

~~Similar to other data assimilation ensemble schemes, the GHOSH filter provides an estimate of the state of a system at some time t_i in terms of the state vector and the covariance matrix that represents the error estimate of the state vector. Hence, at time t_i, a forecast is produced in the form of the forecast state vector $\boldsymbol{x}_i^f \in \mathbb{R}^N$ (N is the dimension of the state space) and the forecast error covariance matrix $\mathbf{P}_i^f$. Moreover, if an observation vector $\boldsymbol{y}_i$ is available at time t_i, the information is assimilated in the analysis state vector $\boldsymbol{x}_i^a$ with error covariance matrix $\mathbf{P}_i^a$.Being an ensemble-based filter scheme, the GHOSH filter represents these quantities by an ensemble of state vectors~~

$$\mathbf{X}_i^a = \left(\boldsymbol{x}_i^{a,1}, \ldots, \boldsymbol{x}_i^{a,(r+1)}\right)$$

~~of $r+1$ model state realizations. However, differently from other ensemble filters that uniformly weight the ensemble members~~In this work, the GHOSH ~~filter assigns to $\mathbf{X}_i^a$ a vector of corresponding weights~~

$$\boldsymbol{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_{r+1} \end{pmatrix}.$$

~~The state estimate is given by the weighted mean~~

$$\boldsymbol{x}_i^a = \sum_{j=1}^{r+1} \boldsymbol{x}_i^{a,j} w_j = \mathbf{X}_i^a \boldsymbol{w},$$

~~while the error covariance is approximated by the ensemble covariance matrix~~

$$\mathbf{P}_i^a \approx \mathbf{X}_i^a \mathbf{W}\, \mathbf{X}_i^a{}^T - \boldsymbol{x}_i^a\, \boldsymbol{x}_i^a{}^T,$$

with $\mathbf{W} = \mathrm{diag}(\boldsymbol{w})$ being the diagonal matrix of weights.

The GHOSH algorithm consists of 5 phases, i.e., *initialization*, *forecast*, *forecast high-order sampling*, *analysis* and *analysis high-order sampling* (Fig. **??**). The *initialization* is performed once at the beginning of the filter application, the *forecast* and *forecast high-order sampling* phases are carried out at each time t_i, and *analysis* and *analysis high-order sampling* are executed only when an observation is available.

GHOSH filter flow chart. Subscripts represent the time steps of the numerical simulation.

The *analysis* equations are a novel weighted version of the ensemble Kalman filter equations (Evensen (1994)), while the *forecast* phase equations rely on a hybrid approach. In fact, the forecast uncertainty is obtained by combining the ensemble covariance (weighted by a forgetting factor) and a model error covariance matrix $\mathbf{Q}_i$, which is built from the existing knowledge of the system, as done for the background matrix in many variational schemes (Bannister (2017a, 2008)).

The method includes two resampling phases, after *analysis* and *forecast* (Fig. **??**), at which the whole ensemble is rebuilt using the high-order sampling method (Section **??**). The resampling produces a sample of state vectors that matches the moments of the error probability density function with better precision than in the other commonly used ensemble DA algorithms. In fact, projecting the ensemble in the subspace spanned by the most relevant components of the ensemble principal components analysis (PCA), the corresponding discrete probability distribution function matches the moments of the projection of the error uncertainty up to order $h \geq 2$ (see Appendix **??**). When the sampling is applied after the forecast, GHOSH allows to take further advantage of the high convergence order of the method, especially when the model error is not neglected.

1.1 The global GHOSH filter

1.0.1 Initialization

First, following the high-order sampling *initialization* phase (Section **??**), the three GHOSH hyper-parameters r (the dimension of the error subspace), h (the higher approximation order reached in the most relevant directions) and s (the number of principal components that are approximated with order h) are used to compute the sampling matrix $\mathbf{\Omega}^h$ and the weights vector $\boldsymbol{w}$.

A starting weighted ensemble must be provided, with the coordinates of $\boldsymbol{w}$ as weights. Such ensemble can come from any previous forecast/assimilation using the GHOSH filter, or it can be built from scratch with an ensemble generation method (e.g., a PCA on historical data followed by the GHOSH sampling method described in Section **??**). The ensemble members are stored in the columns of ensemble matrix $\mathbf{X}_0^f$, i.e.,

$$\mathbf{X}_0^f = \left(\boldsymbol{x}_0^{f,1}, \ldots, \boldsymbol{x}_0^{f,(r+1)}\right),$$

with the subscripted index 0 representing the first time step.The mean $\mathbf{x}_0^f$, can be computed as

$$\mathbf{x}_0^f = \mathbf{X}_0^f \mathbf{w}.$$

1.0.1 Analysis phase

~~If observations are available at time step i (Fig. ??), $\boldsymbol{y}_i \in \mathbb{R}^n$ represents the observation array and $\mathcal{H}_i$ is the observation operator, which incorporates all operations needed to obtain the measured quantities from the state vector. This operator is used on each ensemble member $\boldsymbol{x}_i^{f,k}$ to build the $n \times (r+1)$ matrix $\mathbf{Y}_i$, i.e.,~~

$$\boldsymbol{y}_i^k = \mathcal{H}_i\left(\boldsymbol{x}_i^{f,k}\right)$$

$$\mathbf{Y}_i = \left(\boldsymbol{y}_i^1, \ldots, \boldsymbol{y}_i^{(r+1)}\right),$$

~~with $k \in \{1, \ldots, r+1\}$.The base $\tilde{\mathbf{L}}_i^a$ of the error subspace spanned by the ensemble is computed by~~

$$\tilde{\mathbf{L}}_i^a = \mathbf{X}_i^f \mathbf{T},$$

~~where $\mathbf{T}$ is an $(r+1) \times r$ full-rank matrix with zero column sums. A reasonable choice is~~

$$\mathbf{T} = \left(\begin{array}{c} \mathbf{I}_r \\ \hline 0 \quad \cdots \quad 0 \end{array} \right) - \boldsymbol{w}\mathbb{1}_{1\times r},$$

~~with $\mathbf{I}_r$ being the identity matrix of size r and $\mathbb{1}$ is a matrix (of the subscripted size) filled with ones. Equation (??) implies the use of the ensemble anomalies as basis members, and it is a common choice in SEIK implementations (e.g., Triantafyllou et al. (2003)), but other $\mathbf{T}$ can be explored without affecting the algorithm (e.g., see Nerger et al. (2012)).The analysis error covariance matrix $\mathbf{P}_i^a$ is obtained by~~

$$\mathbf{P}_i^a = \tilde{\mathbf{L}}_i^a \mathbf{A}_i^a \, {\tilde{\mathbf{L}}_i^a}^T,$$

~~where~~

$${\mathbf{A}_i^a}^{-1} = \mathbf{T}^T W^{-1} \mathbf{T} + \mathbf{Z}_i^T \mathbf{R}_i^{-1} \mathbf{Z}_i,$$

~~and~~

$$\mathbf{Z}_i = \mathbf{Y}_i \mathbf{T},$$

~~$W = \mathrm{diag}(\boldsymbol{w})$ being the $(r+1)\times(r+1)$ diagonal matrix of weights and $\mathbf{R}_i$~~ skill is assessed qualitatively, looking at the ~~$n \times n$ error covariance matrix representing the uncertainty in the observation y_i.An opportune change in basis is used to obtain the simpler decomposition~~

$$\mathbf{P}_i^a = \mathbf{L}_i^a \mathbf{L}_i^{aT},$$

~~where~~

$$\mathbf{L}_i^a = \tilde{\mathbf{L}}_i^a \mathbf{S}_i^a$$

~~and $\mathbf{S}_i^a$ is the symmetric square root of $\mathbf{A}_i^a$, i.e.,~~

$$\mathbf{S}_i^a \mathbf{S}_i^a = \mathbf{A}_i^a.$$

~~Finally, the analysis state $\boldsymbol{x}_i^a$ is estimated by~~

$$\boldsymbol{x}_i^a = \boldsymbol{x}_i^f + \mathbf{L}_i^a \mathbf{S}_i^a \mathbf{Z}_i^T \mathbf{R}_i^{-1} \left(\boldsymbol{y}_i - \mathbf{Y}_i \boldsymbol{w} \right).$$

~~Equation (**??**) differs from the usual ensemble Kalman filter equations in the use of $\mathbf{Y}_i \boldsymbol{w}$ in the last term instead of $\mathcal{H}_i\left(\boldsymbol{x}_i^f\right)$. The two are the same if the observation operator $\mathcal{H}_i$ is linear. However, in the general case, $\mathbf{Y}_i \boldsymbol{w}$, which relies on the whole ensemble instead of the ensemble mean only, is a better estimator of the expected value of the observed quantities. Moreover, ensembles generated with GHOSH 's sampling method lead to a further advantage due to the higher order of approximation .~~

1.0.1 ~~Analysis high-order sampling~~

~~The $N \times (r+1)$ ensemble matrix $\mathbf{X}_i^a = \left(\boldsymbol{x}_i^{a,1}, \ldots, \boldsymbol{x}_i^{a,(r+1)}\right)$, whose columns are the ensemble members, is obtained by~~

$$\mathbf{X}_i^a = \boldsymbol{x}_i^a \mathbb{1}_{1\times(r+1)} + \mathbf{L}_i^a \mathbf{C}_i^{aT} \boldsymbol{\Omega}_i^{aT} W^{-\frac{1}{2}}.$$

~~The sampling procedure is described in Section **??**; therefore, $\mathbf{C}_i^a$ and $\boldsymbol{\Omega}_i^a$ are computed as $\mathbf{C}_i$~~ dynamics of chlorophyll, nutrients and their correlations, and ~~$\boldsymbol{\Omega}_i$ in equation (**??**).~~

1.0.1 ~~Forecast phase~~

~~In this phase, the time step index i is increased by 1 and the ensemble is evolved through the model operator $\mathcal{M}_i$ (which usually represents the numerical integration of a system of differential equations), i.e., for $k \in \{1, \ldots, r+1\}$,~~

$$\tilde{\boldsymbol{x}}_i^{f,k} = \mathcal{M}_i\left(\boldsymbol{x}_{i-1}^{a,k}\right),$$

$$\tilde{\mathbf{X}}_i^f = \left(\tilde{\boldsymbol{x}}_i^{f,1}, \ldots, \tilde{\boldsymbol{x}}_i^{f,(r+1)}\right).$$

~~The forecast state x_i^f is estimated as the weighted mean of the ensemble by~~

$$\boldsymbol{x}_i^f = \tilde{\mathbf{X}}_i^f \boldsymbol{w}.$$

~~To track the uncertainty of this estimation, the basis of the error subspace is obtained by~~

$$\tilde{\mathbf{L}}_i^f = \tilde{\mathbf{X}}_i^f \mathbf{T}.$$

~~The covariance matrix $\mathbf{P}_i^f$ is approximated by~~

$$\mathbf{P}_i^f = \tilde{\mathbf{L}}_i^f \left(\mathbf{T}^T \mathbf{W}^{-1} \mathbf{T}\right)^{-1} \tilde{\mathbf{L}}_i^{f\,T} + \mathbf{Q}_i \approx \tilde{\mathbf{L}}_i^f \mathbf{A}_i^f \tilde{\mathbf{L}}_i^{f\,T},$$

~~where $\mathbf{Q}_i$ is the $N \times N$ covariance matrix of the model error and $\mathbf{A}_i^f$ is the $r \times r$ matrix approximating the error covariance in the reduced basis expressed by $\tilde{\mathbf{L}}_i^f$. $\mathbf{A}_i^f$ can be computed in many ways, depending on the form of $\mathbf{Q}_i$ and on the chosen hybridization strategy. Here we consider the case of $\mathbf{Q}_i$ being a full rank matrix (e.g., diagonal), thus we suggest~~

$$\mathbf{A}_i^f = \left(\rho \mathbf{T}^T \mathbf{W}^{-1} \mathbf{T}\right)^{-1} + \left(\tilde{\mathbf{L}}_i^{f\,T} \mathbf{Q}_i^{-1} \tilde{\mathbf{L}}_i^f\right)^{-1},$$

~~where $\rho \leq 1$ is the forgetting factor, which is added to the equation to introduce inflation (Carrassi et al. (2018)). The last term in equation (**??**) is one of the novel element of the present work. It represents the projection of the model error on the error subspace defined by the ensemble. This non-orthogonal projection, induced by the scalar product defined by the matrix $\mathbf{Q}_i^{-1}$, is a form of hybridization which aims to focus on the effects of the model uncertainty in the ensemble error subspace.Since $\mathbf{Q}_i$ can have a very large size, it should be sparse or managed in a decomposed form. $\mathbf{Q}_i$ should be built accordingly to some knowledge of the system, as done, for example, for the background error covariance matrix in variational methods (Bannister (2017a, 2008)) . A new basis for the error subspace is obtained by~~

$$\mathbf{L}_i^f = \tilde{\mathbf{L}}_i^f \mathbf{S}_i^f,$$

~~where $\mathbf{S}_i^f$ is the symmetric square root of $\mathbf{A}_i^f$, i.e.,~~

$$\mathbf{S}_i^f \mathbf{S}_i^f = \mathbf{A}_i^f,$$

~~which can be computed by an eigenvalue decomposition of $\mathbf{A}_i^f$~~quantitatively, in terms of improved agreement with observations without degradation of non-assimilated variables. Moreover, a sensitivity analysis has been designed to highlight the importance of the GHOSH enhanced approximation order by computing its correlation with the root mean square deviation (RMSD) metric. Finally, ~~equation (??) can be rewritten as~~

$$\mathbf{P}_i^f \approx \mathbf{L}_i^f {\mathbf{L}_i^f}^T.$$

~~Writing $\mathbf{P}_i^f$ in this form is important for the consistency of the filter and is useful for the~~ the feasibility of a real implementation of the ~~localization (see Section ??).~~

1.0.1 ~~Forecast high-order sampling~~

~~The $N \times (r+1)$ ensemble matrix $\mathbf{X}_i^f = \left(\boldsymbol{x}_i^{f,1}, \dots, \boldsymbol{x}_i^{f,(r+1)}\right)$, whose columns are the ensemble members, is obtained by~~

$$\mathbf{X}_i^f = \boldsymbol{x}_i^f \mathbb{1}_{1\times(r+1)} + \mathbf{L}_i^f {\mathbf{C}_i^f}^T {\boldsymbol{\Omega}_i^f}^T W^{-\frac{1}{2}}.$$

~~The sampling procedure is similar to that in Section ??; therefore, $\mathbf{C}_i^f$ and $\boldsymbol{\Omega}_i^f$ are computed as $\mathbf{C}_i$ and $\boldsymbol{\Omega}_i$ in equation (??).~~

~~Compared to data assimilation schemes with resampling only after analysis, the GHOSH filter uses this sampling phase to produce a better representation of the uncertainty. In fact, it takes into account the model error effects in equations (??) and (??), which are otherwise neglected. If the model error is supposed to be not significant (i.e., $\mathbf{Q}_i = 0$) then the *forecast high-order sampling* phase can be widely simplified and the ensemble members can be obtained by~~

$$\boldsymbol{x}_i^{f,k} = \boldsymbol{x}_i^f + \frac{1}{\sqrt{\rho}}\left(\tilde{\boldsymbol{x}}_i^{f,k} - \boldsymbol{x}_i^f\right),$$

~~for $k \in \{1, \dots, r+1\}$~~GHOSH filter is supported also by an analysis of the computational resources needed for the filter solution with respect to the overall cost the application.

1.1 ~~The local GHOSH filter~~

~~Localization is a widely used procedure that aims to avoid spurious correlations induced by an overly small ensemble while reducing the degrees of freedom of the system (Janjic et al. (2011)).A localized version of the GHOSH algorithm is obtained by modifying some of the equations in Sections ?? and ??. Three operators $\mathcal{L}_p$~~Section 2 presents the experimental design and introduces the OGSTM-BFM model, ~~$\mathcal{L}_p^{\mathcal{H}}$ and $\mathcal{D}$ are used to improve readability. The localization operator $\mathcal{L}_p$ takes a global (array or) matrix with N rows and returns its localized counterpart with l rows with respect to the domain point p. The simplest example of the localization operator takes just the variable values of the cells in a small radius around p. $\mathcal{L}_p^{\mathcal{H}}$ works in the same way in the observation space, reducing the number of rows from n to l'. The delocalization operator $\mathcal{D}$ takes a set of local matrices (or arrays), ideally one for each point of the domain, and returns the global counterpart. Building the global matrix by taking the values at the central points of each local matrix is a common example of the delocalization operator.~~

1.0.1 ~~Local forecast~~

~~Instead of equations (??-??), the localized version of the forecast step of the algorithm computes, for each point~~ ~~p~~ ~~of the domain, the local covariance matrix~~ ~~$\mathbf{A}_i^{p,f}$~~ ~~by~~

$$\mathbf{A}_i^{p,f} = \left(\rho \mathbf{T}^T \mathbf{W}^{-1} \mathbf{T}\right)^{-1} + \left(\mathcal{L}_p\left(\tilde{\mathbf{L}}_i^f\right)^T \mathbf{Q}_i^{p-1} \mathcal{L}_p\left(\tilde{\mathbf{L}}_i^f\right)\right)^{-1},$$

~~where~~ ~~$\mathbf{Q}_i^p$~~ ~~is the~~ ~~$l \times l$~~ ~~model error covariance matrix at the point~~ ~~p.The global basis is built by changing the basis and delocalizing, i.e.,~~

$$\mathbf{L}_i^f = \mathcal{D}\left(\left\{\mathcal{L}_p\left(\tilde{\mathbf{L}}_i^f\right)\mathbf{S}_i^{p,f}\right\}_p\right),$$

~~where~~ ~~$\mathbf{S}_i^{p,f}$~~ ~~is the symmetric square root of~~ ~~$\mathbf{A}_i^{p,f}$. Note that the change in basis induced by~~ ~~$\mathbf{S}_i^{p,f}$~~ ~~is continuous (see, e.g., Nerger et al. (2012)); thus, the delocalization operator does not need to include averaging or other smoothing procedures to avoid discontinuities.~~

1.0.1 ~~Local analysis~~

~~The localized version of the analysis step substitutes equations (??) and (??) by calculating, for each point~~ ~~p~~ ~~of the domain, the local covariance matrix~~ ~~$\mathbf{A}_i^{p,a}$, i.e.,~~

$$\mathbf{A}_i^{p,a-1} = \mathbf{T}^T \mathbf{W}^{-1} \mathbf{T} + \mathcal{L}_p^{\mathcal{H}}\left(\mathbf{Z}_i\right)^T \mathbf{R}_i^{p-1} \mathcal{L}_p^{\mathcal{H}}\left(\mathbf{Z}_i\right),$$

~~where~~ ~~$\mathbf{R}_i^p$~~ ~~is the~~ ~~$l' \times l'$~~ ~~observation error covariance matrix at the point~~ ~~p. This matrix can be extracted by~~ ~~$\mathbf{R}_i^{-1}$, but it is convenient to increase the uncertainty at the points far from~~ ~~p~~ ~~(as in Nerger and Gregg (2007)). A reasonable choice is a polynomial weight function that goes smoothly to zero out of the localization radius, e.g.,~~

$$f(d) = \left(\left(\frac{d}{d_l}\right)^2 - 1\right)^2,$$

~~with~~ ~~d~~ ~~being the distance from~~ ~~p~~ the data used for assimilation and performance evaluation, and ~~d_l~~ ~~the localization radius.Similarly to equation (??) for local forecast, equation (??) becomes~~

$$\mathbf{L}_i^a = \mathcal{D}\left(\left\{\mathcal{L}_p\left(\tilde{\mathbf{L}}_i^a\right)\mathbf{S}_i^{p,a}\right\}_p\right),$$

~~where $\mathbf{S}_i^{p,a}$ is the symmetric square root of $\mathbf{A}_i^{p,a}$~~a summary of the relevant GHOSH key elements. Section 3.1 shows the results of the experiments, that are discussed in Section 4. Finally, Appendix A contains some specific details of the GHOSH implementation. ~~Finally, the analysis state $\boldsymbol{x}_i^a$ in equation (??) is instead estimated by~~

$$\boldsymbol{x}_i^a = \boldsymbol{x}_i^f + \mathcal{D}\left(\left\{\mathcal{L}_p\left(\tilde{\mathbf{L}}_i^a\right)\mathbf{A}_i^{p,a}\mathcal{L}_p^{\mathcal{H}}\left(\mathbf{Z}_i\right)^T\mathbf{R}_i^{p-1}\mathcal{L}_p^{\mathcal{H}}\left(\boldsymbol{y}_i - \mathbf{Y}_i\boldsymbol{w}\right)\right\}_p\right).$$

2 ~~Experimental setup~~

~~The GHOSH filter has been tested in two sets of numerical experiments of different complexity. First, in Section **??**, a very large set of twin experiment based on the Lorenz96 model (Lorenz (1996)) aims to evaluate the performance of GHOSH compared with a second order filter (SEIK). The Lorenz96 model, which has a chaotic behaviour that is comparable to the one of fluid dynamics equations at a largely lower computational cost , is a standard choice for testing new data assimilation schemes (e.g., Brajard et al. (2020); **?**); **?**); **?**); **?**).~~

~~The second experiment aims to assess the feasibility of the GHOSH implementation in a realistic geophysical application(Section **??**). In particular, the GHOSH filter has been coupled with the Mediterranean biogeochemical model system currently used in Copernicus Marine CMEMS System (OGSTM-BFM: Salon et al. (2019)) to assimilate ocean color satellite data.~~

1.1 ~~Twin Experiment~~

~~The set of twin experiments is based on the Lorenz96 model with random initial conditions. In each experiment, the *truth* is provided by integrating the model for a fixed time interval. Then, at regular time intervals, observations of a subset of the system variables have been extracted by the *truth*, adding a random error to each of them to represent the observation uncertainty. The same set of observations is then used for two assimilation experiments, one using the SEIK filter (as a second order approximation filter Nerger et al. (2005)), and one using the GHOSH filter with fifth approximation order (i.e., $h = 5$, see Section **??**). Both the filters are initialized from the same initial condition, chosen randomly around the *truth* initial condition. After an initial spin up, the skill of each filter is computed by averaging the root mean square error (RMSE) between the filter forecast and the *truth* before the assimilation step. RMSE evolution in time and a synthetic RMSE value for the whole simulation time has been evaluated for each filter. Then, in order to produce reliable statistics, the same procedure has been repeated 56000 times by changing the *truth* initial conditions, observations, filters initial conditions and filters hyper-parameters such as the ensemble size and the forgetting factor.~~

1.0.1 ~~The Lorenz96 model~~

~~The Lorenz96 model(Lorenz (1996)) is a dynamical system commonly used as a testing framework in data assimilation. The state vector $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathbb{R}^N$ is evolved according to the system of differential equations given by, for $1 \leq j \leq N$,~~

$$\cancel{\frac{dx_j}{dt} = (x_{j+1} - x_{j-2})\, x_{j-1} - x_j + F,}$$

~~with periodic conditions, i.e., $x_{-1} = x_{N-1}$, $x_0 = x_N$ and $x_{N+1} = x_1$. In our implementation, the size of the state array was $N = 62$ and the forcing term was $F = 8$, which is a common value used to generate chaotic behaviour.The equations have been implemented in python and solved numerically using SciPy's *solve_ivp* routine.At time $t = 0$, the state vector has been initialized by adding a small perturbation ($x_N = 0.01$) to the null solution $\boldsymbol{x} = 0$. The twin experiment truth has been obtained by integrating the model for 100 time units. In particular, after a 20 time units spin-up, the obtained solution in the time intervals $t \in [20, 40]$, $t \in [40, 60]$, $t \in [60, 80]$,$t \in [80, 100]$ has been referred as *truth_20*, *truth_40*, *truth_60*, *truth_80*.~~

1.0.1 ~~Observations~~

~~Observations are generated for each truth period (*truth_20*, *truth_40*, *truth_60*, *truth_80*) by extracting from the state vector, the values of the even indexed variables (i.e., $x_2, x_4, ..., x_{62}$) every Δt time units and adding to each observed variable a random number sampled from a standard normal distribution. Tests have been made for $\Delta t \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$.~~

1.0.1 ~~Filters setup~~

~~The same settings are used in both SEIK and GHOSH filters. The ensemble size was chosen among $7, 15, 31$ and 63 members. The initial condition of the ensemble mean is chosen randomly around each *truth* initial condition, adding a random Gaussian noise with 0 mean and a standard deviation of 5. Initial conditions of the ensemble members are sampled with each filter's own sampling method (second order for SEIK, $h = 5$ for GHOSH), setting the prior error covariance to a diagonal matrix representing a standard deviation of 5 in each variable.Both filters have the same forgetting factor, chosen among $0.7, 0.75, 0.8, 0.85, 0.9, 0.9$ and 1 (the last value describes a condition without inflation). The GHOSH filter is run without hybridization ($\mathbf{Q}_i = \mathbf{0}$) to keep GHOSH and SEIK as similar as possible. The order of the GHOSH filter was fixed at $h = 5$, with s (i.e. , the number of principal components approximated with order $h = 5$) as big as possible, depending on the ensemble size. The values adopted for s are $2, 3, 4$ and 5, respectively for $7, 15, 31$ and 63 ensemble members.~~

1.0.1 ~~Experiments design~~

~~100 twin experimentshave been launched for each of the 4 truth periods (*truth_20*, *truth_40*, *truth_60*, *truth_80*), 5 observation frequencies, 4 ensemble sizes and 7 forgetting factors. Each of the 100 twin experiments has its own set of random observations and initial conditions. Each experiment lasts 20 time units (as in the *truth* simulations), using the first half for spin-up. During the second half of the time interval, the root mean square error (RMSE) for assimilated and non-assimilated variables is separately measured before each assimilation.All the code was written in python and executed in a Linux computer equipped with an Intel(R) Core(TM) i7-11800H @2.30GHz.~~

1.1 ~~Three-dimensional marine biogeochemistry application~~

~~The GHOSH filter has been implemented in the OGSTM-BFM biogeochemical model system of the Mediterranean Sea at $\frac{1}{4}^{o}$ resolution. The GHOSH implementationfeatures a weekly assimilation of surface chlorophyll from satellite ocean color (as in Teruzzi et al. (2018a)).Consistently, an OGSTM-BFM model run without assimilation (control run) has been used as a reference to investigate the GHOSH filter sensitivity to some different parameterizations.~~

2 Material and methods

2.0.1 ~~The OGSTM-BFM model~~

2.1 The OGSTM-BFM model

The biogeochemical model used in this study is the coupled transport-biogeochemical model OGSTM-BFM, which is the biogeochemical component of the Mediterranean Copernicus Marine Service ~~(Salon et al. (2019))~~(Salon et al., 2019). The OGSTM transport model ~~is a modified version of the OPA 8.1 transport model (Foujols et al. (2000)), which~~ resolves the advection, the diffusion and the sinking terms of the biogeochemical variables ~~(Lazzari et al. (2010))~~(Lazzari et al., 2010). The biogeochemical BFM model describes the biogeochemical cycles of carbon, nitrogen, phosphorus and silicon through the dissolved inorganic, particulate living organic and non-living organic compartments, and includes nine plankton functional types, as well as dissolved oxygen, pH and the carbonate system (for a total of 51 tracers as state variables). The OGSTM-BFM model has been applied to study the chlorophyll and primary production ~~(Lazzari et al. (2012); Teruzzi et al. (2018a))~~(Cossarini et al., 2021; Teruzzi et al., 2021, 2018b; Lazzari et al., 2012), nutrient cycles ~~(Lazzari et al. (2016))~~(Lazzari et al., 2016), alkalinity, the carbon cycle and carbon fluxes ~~(Cossarini et al. (2015); Melaku Canu et al. (2015))~~ (Cossarini et al., 2015; Melaku Canu et and oxygen dynamics ~~(Di Biagio et al. (2022))~~(Di Biagio et al., 2022).

The OGSTM-BFM is forced by the output of an ocean general circulation model ~~(i.e., NEMO, from the Copernicus Marine Mediterranean system in the present application; Salon et al. (2019))~~ (i.e., NEMO, from the Copernicus Marine Mediterranean system in th that provides fields of physical quantities (i.e., velocity, temperature, salinity, diffusivity, and solar radiation and wind at the surface from a physical reanalysis) driving the transport processes, the kinetic rates of chemical reactions, and energy and matter exchanges at the air-sea interface.

The present setup (terrestrial and atmospheric inputs, boundary conditions) is described in ~~Salon et al. (2019)~~Salon et al. (2019) .

2.1.1 ~~Observations~~

2.2 Observations

The assimilated observations are satellite chlorophyll maps obtained from the CMEMS multisensor product OCEANCOLOUR_MED_CHL_L3_REP_OBSERVATIONS_009_073 ~~(Volpe et al. (2017))~~(Volpe et al., 2017). The original

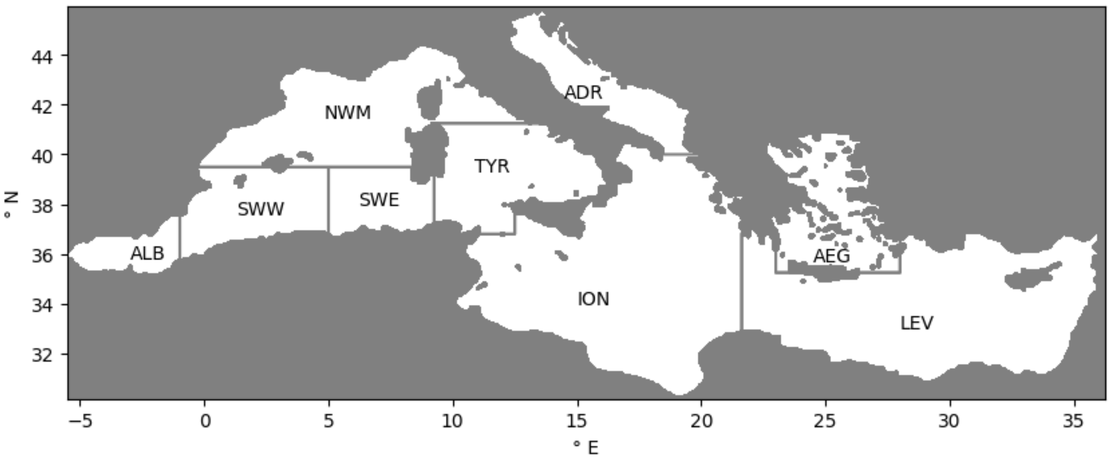


Figure 1. Sub-basins partition of the Mediterranean Sea.

data at 1 km resolution were checked for spikes, excluding values exceeding by two standard deviations the climatological mean and remapped on the model grid resolution.

Additionally, in situ climatological ~~data~~ datasets for nitrate and phosphate, to be used for validation, ~~is~~ are computed from a dataset that integrates the in situ EMODnet data collections ~~(Buga et al. (2018)) and the datasets listed in Lazzari et al. (2016)~~ (Buga et al., 2018) and the data sources listed in Lazzari et al. (2016). The climatology is calculated for the 0-30 m layer on 9 sub-basins ~~(Salon et al. , 2019)~~ represented in Fig. 1 (Salon et al., 2019) and for 4 different seasons (January-March; April-June; July-September; October-December).

2.2.1 ~~GHOSH setup~~

~~**Test name Order (h) Forgetting factor (ρ) Forecast frequency** *ctrl T_h2_rho1_d* 2 *T_h3_rho1_d* 3 *T_h5_rho1_d* 5 *T_h2_rho0.9_d* 2 *T_h3_rho0.9_d* 3 *T_h5_rho0.9_d* 5 *T_h2_rho0.8_d* 2 *T_h3_rho0.8_d* 3 *T_h5_rho0.8_d* 5 *T_h2_rho1_w* 2 *T_h3_rho1_w* 3 *T_h5_rho1_w* 5 *T_h2_rho0.7_w* 2 *T_h3_rho0.7_w* 3 *T_h5_rho0.7_w* 5 *T_h2_rho0.5_w* 2 *T_h3_rho0.5_w* 3 *T_h5_rho0.5_w* 5 Experimental setup.~~

~~In the 3D simulations, GHOSH assimilated surface chlorophyll and update the whole state vector (~~

2.3 The GHOSH filter

The GHOSH filter, described in details in (Spada et al., 2024), is an ensemble DA filter that uses $r+1$ ensemble members to represent the state estimation and its uncertainty. r is the dimension of the uncertainty subspace, i.e., ~~51 state variables) of the OGSTM-BFM model by using an ensemble of 16 members(i.e. , an error subspace of dimension $r = 15$). Tests were~~

~~carried out for each of three different maximum polynomial orders h (and corresponding numbers s of principal components approximated with such orders):~~

~~$h=2$, i.e., a second~~ the subspace of the state space spanned by the ensemble members. The GHOSH filter features an enhanced order of approximation ~~on the whole error subspace ($s=15$) ; $h=3$, i.e., third order of approximation on the 8 most relevant components ($s=8$) and second order on the remaining 7 principal directions; $h=5$,~~ (named h) in the s-dimensional subspace (included in the uncertainty subspace) where the approximation error may be maximum. This is realized by partially parameterizing (in the moment tensor μ) the prior uncertainty pdf (i.e., ~~fifth order of approximation on the top 3 principal components ($s=3$), and second order on the rest (12 components) .~~

~~The ensemble weights and $\mathbf{\Omega}^h$ were computed (usually during the *initialization* phase) by solving system (**??**) . In the case with $h=2$, the solution is easily obtained starting from the constant vector v of the square root of the weights and adding orthonormal column vectors in order to form an orthogonal matrix (see equation (**??**) in Appendix **??**)~~probability density function) and by using a principal component analysis (PCA) to chose the subspace of the best approximation order. Before the PCA computation, a scaling matrix $\mathbf{\Lambda}$ is used to rescale the state variables, i.e. to define which variables are of more interest. In ~~the case with $h=3$, a solution with constant weights ($v_m=\frac{1}{4}$ for each $m\in\{1,\dots,16\}$) was chosen, namely,~~

$$\cancel{\mathbf{\Omega}^{h=3}=\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}^T.}$$

~~In the case with $h=5$, the chosen solution was~~

$$\cancel{\boldsymbol{v}=\Big(\frac{1}{\sqrt{5}} \quad \frac{1}{\sqrt{5}} \quad \frac{1}{2\sqrt{5}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \cdots}$$

$$\cancel{\cdots \quad \frac{1}{2\sqrt{5}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \Big)^T}$$

~~and~~

$$\cancel{\mathbf{\Omega}^{h=5}=\left(\begin{matrix} 0 & 0 & \frac{1}{2} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \cdots \\ 0 & 0 & 0 & \frac{1}{\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & \cdots \\ 0 & 0 & 0 & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \cdots \end{matrix}\right.}$$

$$\cancel{\left.\begin{matrix} \cdots & -\frac{1}{2} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} \\ \cdots & 0 & \frac{1}{\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} \\ \cdots & 0 & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} \end{matrix}\right)^T.}$$

~~The scaling matrix $\mathbf{A}_i$ (equation (??)) is a diagonalmatrix representing the error~~ this implementation, μ describes (up to order h) a Gaussian distribution, and the scaling matrix $\mathbf{\Lambda}$ is diagonal. Being $\mathbf{P} = \mathbf{L}\mathbf{L}^T$ the uncertainty covariance matrix, $\mathbf{\Lambda}$ contains the uncertainty variance (i.e., the diagonal of ~~$\mathbf{P}_i^a$), and it~~ $\mathbf{P}$), which is computed as the squared Euclidean norms of the rows of ~~$\mathbf{L}_i^a$. The $\mathbf{A}_i$~~ $\mathbf{L}$. In the present application, the $\mathbf{\Lambda}$ matrix has been chosen valuing simplicity and generality over skill performance. The implemented scaling matrix is probably a suitable choice for many systems, since it implies that the PCA ~~measures~~ is computed on the Pearson correlation matrix and that it is not affected by the different amplitudes of the variable standard deviations.

During the forecast and assimilation steps, the uncertainty expressed by the ensemble member's covariance matrix is inflated by a factor $\frac{1}{\rho}$ (ρ is called forgetting factor, see e.g., Pham et al., 1998) and combined with a covariance matrix $\mathbf{Q}$ representing the parametric uncertainty that is non-dependent by the ensemble (as typically happens in hybrid filters, see e.g., Carrassi et al., 2018) .

The filter is implemented in its localized version, which takes into account only correlations within a certain maximum horizontal distance (the localization radius).

~~The simulations cover the period~~

2.4 Experimental design

Implemented in the OGSTM-BFM biogeochemical model system of the Mediterranean Sea at $\frac{1}{4}^o$ resolution, the local GHOSH filter uses 16 ensemble members (i.e., an uncertainty subspace of dimension $r = 15$) to perform a weekly assimilation of surface chlorophyll from satellite ocean color (as in Teruzzi et al., 2018a) during the period 1 January 2013 - 1 January ~~2014, with daily or weekly forecast frequency (1).The ensemble covariance was inflated as in equation (??). Tests were conducted with different forgetting factor values, namely, $\rho \in \{1, 0.9, 0.8\}$ for daily forecasts and $\rho \in \{1, 0.7\}$ for weekly forecasts.~~

~~The localized version of the filter used cylindrical and circular patches for $\mathcal{L}_p$ and $\mathcal{L}_p^{\mathcal{H}}$, using a 90 km radius (similarly to values adopted by other EnKF applications, e.g., Hu et al. (2012); Ciavatta et al. (2016)) centred at the point p. The delocalization operator $\mathcal{D}$ builds the global array by taking the values in the central cell of each localized array (Section ??). However, since the change in basis in the GHOSH algorithm was continuous (as noted in Section ??), more demanding localization strategies, such as averaging over the weighted values of each cell local array, would not introduce relevant improvements~~2014. Consistently, an OGSTM-BFM model run without assimilation (control run) is used as a reference to investigate the GHOSH filter performances in some different settings.

Given the non-Gaussianity of biogeochemical tracer concentrations, the GHOSH filter ~~was~~ is applied to the log-transformed biogeochemical variables ~~(Ciavatta et al. (2016); Pradhan et al. (2019); Song et al. (2016); Ford (2020))~~(Ciavatta et al., 2016; Pradhan et a . To avoid instabilities due to large negative values of logarithms of very small concentrations, a cut-off ~~was~~ is applied to values lower than the threshold of 10^{-4}. This solution can be seen as a localization that dynamically removes unwanted correlations, which can typically occur at deeper layers, where some tracers (such as, for example, phytoplankton biomass and chlorophyll) tend to very low and negligible concentrations.

The ~~mean initial condition x_0^f~~ control run initial condition and the starting ensemble ~~$\mathbf{X}_0^f$~~ were produced based on a 10-year hindcast simulation of the OGSTM-BFM model without assimilation. For each year, 31 days centred on the 1st of January were log-transformed ~~, after removing values smaller than~~ after applying a 10^{-4} cutoff on concentrations, for a total of 310 simulated multivariate states. The mean of the multivariate states was used as ~~x_0^f~~control initial condition and mean of the starting ensemble, while a PCA of the ensemble of ~~normalized~~ $\boldsymbol{\Lambda}$-scaled states provided the 15 (i.e., the dimension of the ~~error~~ uncertainty subspace r) ~~most relevant components , which are used to build the base matrix $\mathbf{L}_0^f$. Then, the GHOSH sampling method has been used to extract the 16 starting ensemblemembers, as in Section ??~~principal components used in the GHOSH filter's high-order sampling method (i.e., the GHOSH sampling method described in Spada et al., 2024) to sample the initial ensemble.

The ~~local model error covariance matrix $\mathbf{Q}_i^p$~~ covariance matrix of the parametric uncertainty $\mathbf{Q}$ has been assumed diagonal and invariant in time~~. The variances in $\mathbf{Q}_i^p$ are assumed~~, with variances inversely proportional to the height of the cells. Given that the resolution of the present model setup is not high enough to adequately resolve coastal processes, assimilating observations in coastal areas could lead to instabilities. Therefore, the ~~model error~~ $\mathbf{Q}$ uncertainty variance in the coastal areas was reduced by applying a ~~4th-degree~~ fourth-degree polynomial smoothing function and setting the covariance values at zero in the grid cells closest to the coast. More precisely, the uncertainty starts to decrease where sea depth is 200 metres and reaches zero in areas shallower than 50 metres.~~The local observation error covariance matrix $\mathbf{R}_i^p$ was~~ Also the covariance matrix $\mathbf{R}$ of the observation uncertainty has been assumed diagonal and constant in time. Using the same ~~4th-degree~~ fourth-degree polynomial smoothing applied ~~for $\mathbf{Q}_i^p$~~in the $\mathbf{Q}$ case, the inverse of the ~~$\mathbf{R}_i^p$ variances was~~ observation uncertainty variances is smoothed to zero in the coastal areas (i.e., the ~~$\mathbf{R}_i^p$~~ $\mathbf{R}$ variances increase in the coastal areas oppositely to ~~$\mathbf{Q}_i^p$). Moreover, the matrix $(\mathbf{R}_i^p)^{-1}$ was weighted with the polynomial function in equation (??) to apply~~ $\mathbf{Q}$). Furthermore, the fourth-degree polynomial smoothing was used to weight the local version of ~~the filter~~$\mathbf{R}^{-1}$ to gently remove assimilation of observations beyond the localization radius of 90 km.

~~A~~ In the assimilation experiment, the GHSOH filter is configured to achieve an approximation order $h = 5$ in a subspace of dimension $s = 3$. A forecast, which includes inflation ($\rho = 0.9$), hybridization with $\mathbf{Q}$ and resampling, is performed with daily frequency. Since a tuning of the ~~model and observation error variance~~ parametric uncertainty (represented in $\mathbf{Q}$) and observation uncertainty ($\mathbf{R}$) may be necessary to produce an effective assimilation when system ~~errors~~ uncertainties are not fully known ~~and parametrized (e.g., Cossarini et al. (2019); Ford and Barciela (2017); Janjić et al. (2018); Mattern et al. (2017); Tandeo et al. (2020)). The actual values of model and observations error variances were obtained testing different values and choosing those that provided~~ (e.g., Cossarini et al., 2019; Ford and Barciela, 2017; Janjić et al., 2018; Mattern et al., 2017; Tandeo et al., 2020), the forgetting factor and the uncertainty variances have been chosen (after many short tests not included in this description) in order to provide effective assimilation increments without introducing ~~biogeochemical variables degradation.~~ degradation on biogeochemical variables.

~~In total, 18 assimilation experiments were run, in addition to the control run, as summarized in Tab. 1. Overall, the experimental design features a number of different forgetting factors ρ,~~

Test name	Order (h)	Forgetting factor (ρ)	Forecast frequency
ctrl	Free model		
T_h2_rho1_d	2	1	daily
T_h3_rho1_d	3 ~~values of the approximation~~		
T_h5_rho1_d	5		
T_h2_rho0.9_d	2	0.9	
T_h3_rho0.9_d	3		
T_h5_rho0.9_d	**5**		
T_h2_rho0.8_d	2	0.8	
T_h3_rho0.8_d	3		
T_h5_rho0.8_d	5		
T_h2_rho1_w	2	1	weekly
T_h3_rho1_w	3		
T_h5_rho1_w	5		
T_h2_rho0.7_w	2	0.7	
T_h3_rho0.7_w	3		
T_h5_rho0.7_w	5		
T_h2_rho0.5_w	2	0.5	
T_h3_rho0.5_w	3		
T_h5_rho0.5_w	5		

Table 1. Experimental setup: control run, assimilation experiment (in bold) and sensitivity analysis tests.

As a sensitivity analysis, 17 additional tests (summarized in tab. 1) are carried out varying the polynomial order h, the forgetting factor ρ and the forecast frequency. In particular, the polynomial order h (~~from 2 to 5) and 2 different time intervals between two *forecast* phases (1 and~~ with the corresponding number s of principal components approximated with order h) varies as follows:

– $h = 2$, i.e., a second order of approximation in the whole uncertainty subspace ($s = 15$);

– $h = 3$, i.e., third order of approximation in the top 8 principal components ($s = 8$) and second order in the remaining 7 ~~days)to test the sensitivity of the assimilation scheme~~components;

– $h = 5$, i.e., fifth order of approximation in the top 3 principal components ($s = 3$), and second order in the rest (12 components).

Appendix A reports additional technical details specific of the GHSOH filter settings.

~~The tests were~~

The tests are carried out in the GALILEO cluster at the Italian HPC centre CINECA, mounting 2 Intel Broadwell CPU (Intel(R) Xeon(TM) E5-2697 v4 @2.3GHz, 18 cores each) per each node. The parallel Fortran 2003 implementation of the GHOSH algorithm ran on 46 nodes (corresponding to 1656 cores).

3 Results

3.1 ~~Twin Experiment~~

~~Figure ?? (a and b) reports the evolution of the first 2 variables (one assimilated and one non-assimilated) of the Lorenz96 model for the two filters (GHOSH and SEIK). The selected simulation out of 56000 experiments shows that GHOSH filter evolution is closer to truth evolution (black line in Fig. ??) than SEIK. As expected, the uncertainty of variable 2 (shaded areas) of both filters is lower than in variable 1 because of the assimilation of observations in even variables. The RMSE computed at every time step over the two sets of variables (assimilated and non-assimilated, Fig. ??)shows that GHOSH outperforms SEIK after the spin-up time.~~

~~Results from one twin experiment example: *truth* (black line) is *truth_60*, time between observation (green dots) is 0.15, forgetting factor is 0.8, ensemble size is 31. Shady areas around blue (SEIK) and organge (GHOSH) lines represent the ensemble standard deviation. From top to bottom: time evolution of the first variable (non-assimilated); time evolution of the second variable (assimilated); RMSE along time based on all the assimilated variables; RMSE along time based on all the non-assimilated variables.~~

~~Result summary of twin experiments: each square in the color maps represents the aggregated results of 400 twin experiments, changing *truth*, observations and initial conditions. The results are summarized with color maps aggregated in six different rows, from top to bottom: SEIK RMSE of assimilated variables, SEIK RMSE of non-assimilated variables, GHOSH RMSE of assimilated variables, GHOSH RMSE of non-assimilated variables, the ratio of GHOSH RMSE over SEIK RMSE of assimilated variables, the ratio of GHOSH RMSE over SEIK RMSE of non-assimilated variables (red color implies that GHOSH is better than SEIK). Each column of color maps has different observation frequency, with the numbers on the top indicating the time elapsed between each observation/assimilation. Each color map shows different forgetting factors (Forget, along the y axis) and ensemble sizes (EnsSize, along the x axis).~~

~~In order to be statistically reliable, the RMSE of the 56000 experiments under different conditions have been computed and reported in aggregated form in Fig. ??. Each small square reports the RMSE value computed over a set of 400 experiments, which includes, for each of the four available truths, 100 experiments with different random observations and initial conditions.The first 2 lines of color maps refer to the SEIK filter observed and unobserved variables. In each color map, the time interval~~

~~between assimilated observations increases from left to right, while the different forgetting factor values decrease from top to bottom. The first line in each color map, labelled "*best*", represents the best result obtained between all the tested forgetting factors. The GHOSH experiments are summarized in the same way in~~ In the assimilation experiment, the GHOSH performances have been evaluated by considering several metrics: the forecast skill of the assimilated variable (chlorophyll, Section 3.1) and of the ~~3rd and 4th lines of Fig. **??**, while the last 2 lines show the ratio between GHOSH and SEIK RMSE values, with red color indicating RMSE reduction of GHOSH with respect to SEIK.~~

~~The column corresponding to the smallest ensemble size~~ non-assimilated variables (nutrients, Section 3.2), and the quality of the system state uncertainty estimation (Section 3.3). The results are discussed in detail through a qualitative and quantitative comparison with the control run (i.e., ~~7 members) is not shown, since in this case SEIK and GHOSH behave very similarly, showing very poor performances, due to lack of capability of describing the system complexity with an overly small ensemble size. In these conditions the GHOSH showed a very small improvement with respect to SEIK (0-3%). Looking at the last 2 rows of Fig. **??**, the GHOSH filter outperforms the SEIK filter in most conditions, up to a three times RMSE reduction. Furthermore GHOSH proves to be always at least as good as SEIK. The largest improvements (dark red, RMSE ratio 0.31) occur when: i) inflation is low (forgetting factor near to 1); ii) large amount of information is injected into the assimilation scheme (i.e., high frequency observations, left side of Fig. **??**); iii) the filter can take into account a high dimensional error subspace (i.e., the ensemble size is large).On the other hand, the comparison of each filter tuned with its best forgetting factor (the "*best*" top line in each color map) clearly shows that the GHOSH filter has considerably better performances than the SEIK filter (up to a RMSE ratio of 0.43) with a moderate number of ensemble members (31). In case of maximum ensemble size the improvement is still relevant but less intense (up to 0.9 RMSE ratio). The improvement is comparably small, up to 7%, in the case of small ensemble size (15 members).In general, it is evident that the GHOSH filter, compared to SEIK, is less dependent on inflation, reaching its best performance with a forgetting factor closer to 1~~a hindcast run with no assimilation). The RMSD between simulation (before assimilation) and observations (Section 2.2) is used as a quantitative performance indicator.

~~The skill of both the filters is closely related with the observations frequency: the shorter the time between observations, the higher the accuracy~~The approximation order impact on DA performances has been studied comparing the scores of 18 tests in different configurations (sensitivity analysis, Section 3.4).

Finally, ~~as expected, for both the filters the skill is better on assimilated variables than non-assimilated ones but, quite interestingly, the RMSE reductions of GHOSH with respect to SEIK are slightly larger in the unobserved variables (up to 0.06 RMSE ratio improvement, from 0.67 for assimilated to 0.61 for non-assimilated variables).~~

~~Example of a single twin experiment, with SEIK (blue line) diverging behaviour: *truth* (black line) is *truth_60*, time between observations (green dots) is 0.15, forgetting factor is 0.75, ensemble size is 31. Values over time of a non-assimilated (top) and an assimilated (bottom) variable.~~

~~From the computational point of view, the whole experiment set needed around 7 computational hours. Since the computational cost is dominated by~~ the ~~integration of the model, the time to solution of the two tested filters should be nearly the same, however, unexpectedly, the average time to solution (averaged every 100 twin experiments) of the SEIK filter sometimes is longer than the time to solution of the GHOSH filter, up to twice the time, depending on some settings and random factors.~~

~~This difference in computational time is more evident and occurs more often when the forgetting factor is lower (i.e., when the inflation is more pronounced). A possible explanation comes from the model integration algorithm used in the SciPy's *solve_ivp* routine: the integration time step is not constant and depends on the stability of the equations (stiffness). Which means that, if the filter is less accurate, the output can diverge from the expected trajectory in areas where the equations are stiffer, and the integration needs more time steps to converge rising the time to solution. This is supported by the fact that the SEIK filter can show a diverging behaviour when the forgetting factor is too low (Fig. **??** shows an example). Remarkably, we did not observe any diverging behaviour in the GHOSH filter runs.~~

3.1 ~~Three-dimensional realistic complexity marine biogeochemistry application~~

GHOSH applicability in a real complex marine biogeochemical application with assimilation of ocean color observation has been assessed by measuring the impact of the assimilation routines on the total computational cost.

~~The GHOSH capability to reproduce expected dynamics in a complex and real geophysical application has been evaluated by looking at: the skill performance of forecast of the assimilated variable (chlorophyll) and of the non-assimilated variables (nutrients) and the quality of the system state uncertainty estimation.~~

~~Furthermore, the GHOSH performances have been evaluated varying the order of approximation h. Table 2 shows that the performance of the assimilated variable is substantially unaffected by the GHOSH polynomial order, whilst h demonstrates a relevant impact on a non-assimilated variable, namely the nitrate. In particular, when the order h is 5, nitrate root mean square difference (RMSD) between GHOSH and independent observations improves up to 81% with respect to order $h = 2$ (scored in the set of tests with $\rho = 0.9$ and daily forecasts). Every $h = 5$ test has a smaller RMSD than every corresponding $h = 2$ and, in general, data clearly show that increasing h implies an improvement in RMSD. In fact, the nitrate's RMSD values, centred and normalized over their standard deviation, present a strong negative correlation with h (Pearson's correlation coefficient is -0.80).~~

~~Among the 18 assimilation experiments, *T_h5_rho0.9_d* proved to have the best GHOSH configuration (Tab. 2). The results from this run are discussed in detail in the following sections, through a qualitative and quantitative comparison with the control run (i.e., a hindcast run with no assimilation).~~

3.0.1 ~~Surface chlorophyll~~

3.1 Surface chlorophyll

The surface chlorophyll estimated by the ~~free~~ control run and the GHOSH filter has been compared with satellite data from January to December 2013. A detailed analysis of a specific relevant event (winter bloom, ~~19 March~~ end of March - beginning of April 2013) is shown in Fig. 2~~)~~). The RMSD time series is shown in Fig. 3.

The ~~March map~~ sequence of ocean color maps (**a** row of Fig. 2) describes a typical winter bloom in the western Mediterranean, with the northernmost part still under the convective phase (i.e., blue hole surrounded by high chlorophyll concentration patches during the first phase (maps of 5 and 19 March, Fig. 2~~a~~**a**), the start of the surface bloom in the whole area of the northern

Test name	Order (h)	Chlorophyll RMSD (mg/m^3)	Phosphate RMSD ($mmol/m^3$)	Nitrate RMSD ($mmol/m^3$)
ctrl	no	0.084	0.036	0.99
T_h2_rho1_d	2	0.068	0.035	0.93
T_h3_rho1_d	3	0.068	0.035	0.85
T_h5_rho1_d	5	0.068	0.035	0.83
T_h2_rho0.9_d	2	0.059	0.033	1.50
T_h3_rho0.9_d	3	0.060	0.033	1.14
T_h5_rho0.9_d	**5**	**0.059**	**0.033**	**0.83**
T_h2_rho0.8_d	2	0.059	0.033	1.96
T_h3_rho0.8_d	3	0.058	0.032	2.15
T_h5_rho0.8_d	5	0.058	0.032	1.76
T_h2_rho1_w	2	0.071	0.035	0.94
T_h3_rho1_w	3	0.071	0.035	0.89
T_h5_rho1_w	5	0.071	0.035	0.91
T_h2_rho0.7_w	2	0.063	0.034	0.94
T_h3_rho0.7_w	3	0.063	0.035	0.93
T_h5_rho0.7_w	5	0.063	0.034	0.91
T_h2_rho0.5_w	2	0.060	0.033	1.35
T_h3_rho0.5_w	3	0.060	0.032	1.02
T_h5_rho0.5_w	5	0.060	0.033	0.91

~~Results: summary~~

Table 2. Summary of the root mean square ~~difference~~ deviation (RMSD) indicators~~for each test~~. ~~Best GHOSH configuration~~ Results for control run, assimilation experiment (in bold) and sensitivity analysis tests.

western Mediterranean (map of 26 March) and its continuation in the following weeks (maps of 2 and 9 April in Fig. 2). Additionally, high chlorophyll levels are present along the Algerian Current from the beginning of March as a result of the active mesoscale dynamics~~, whereas~~. The high concentration of chlorophyll along the Algerian Current decreases in the second half of March. On the other hand, with the exception of few coastal river-influenced areas, the eastern Mediterranean Sea has a much lower chlorophyll concentration due to the lower intensity of the winter vertical mixing (Fig. 2~~a~~**a**).

~~While the free run overestimates the surface chlorophyll in the whole Mediterranean Sea~~ The control run anticipates the start of the bloom in the north-western region by 2-3 weeks, with the peak of chlorophyll concentration being reached in the map of 19 March. The evolution of chlorophyll in the Algerian current is quite well reproduced in the winter period, even if it is somewhat overestimated in the first weeks. Finally, the control run tends to slightly overestimate the chlorophyll concentration in the entire eastern basin in March (Fig. 2~~b),~~ **b**). The maps of the pre-assimilation ~~picture of GHOSH run shows a more~~

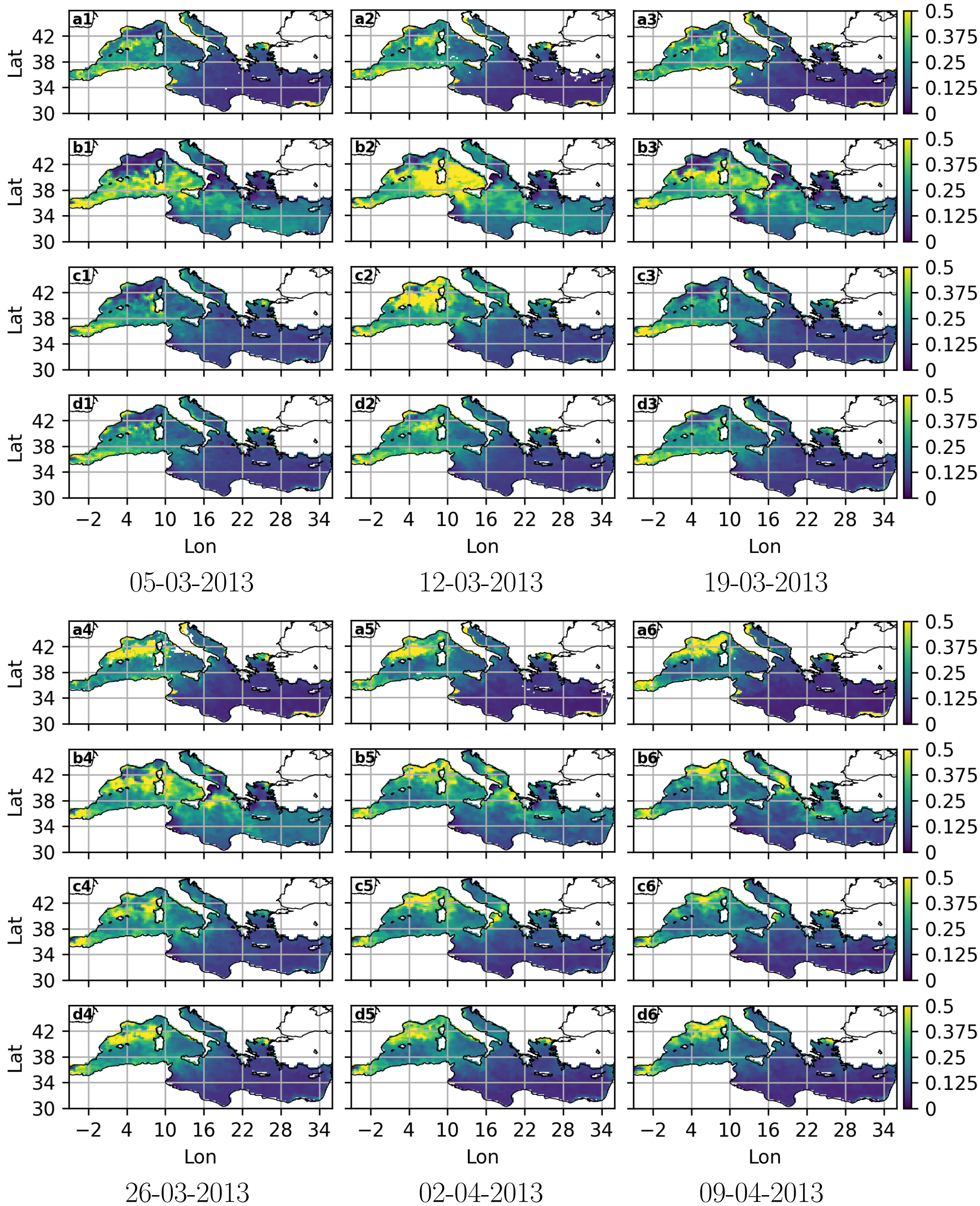


Figure 2. Surface chlorophyll (mg/m^3) comparison: satellite (**a**), control run (**b**) and data assimilation (before and after assimilation, **c** and **d** respectively) ~~on~~ during a winter bloom event (19.03.2013 - 09.04.2013).

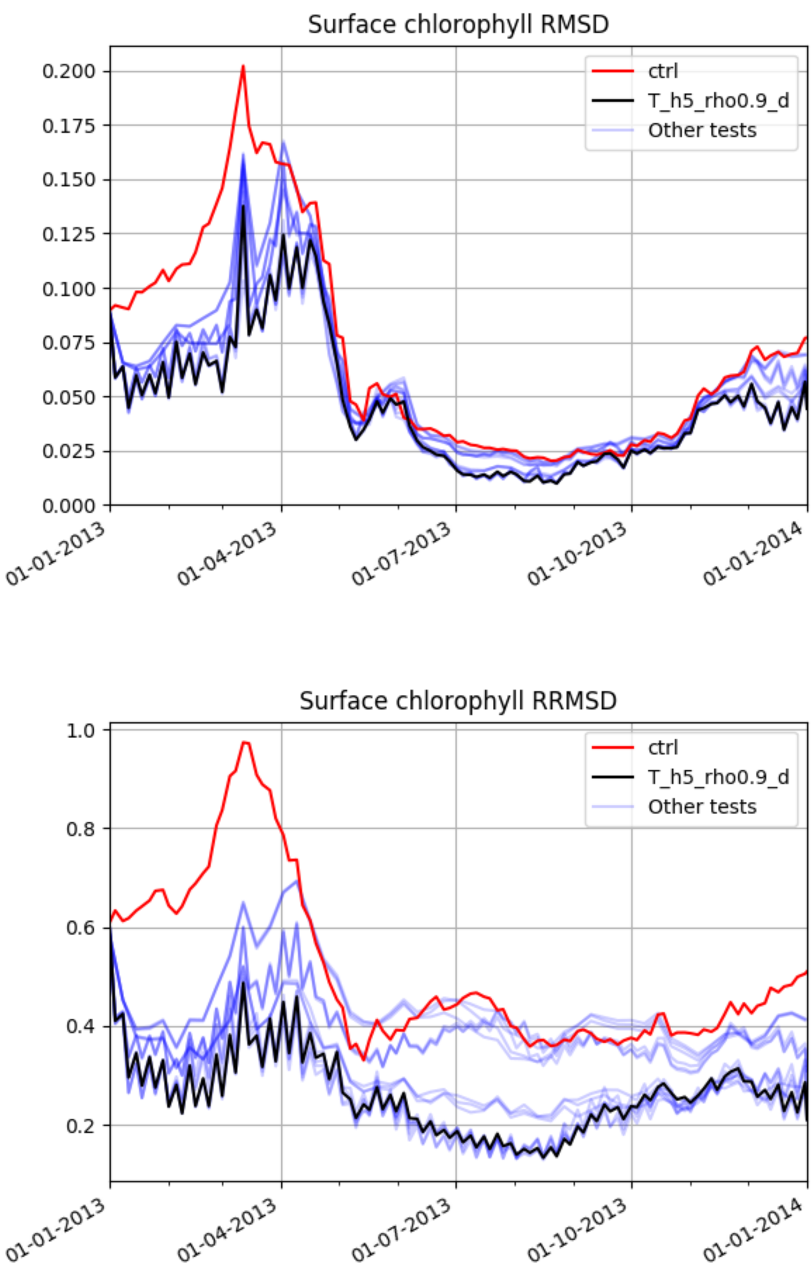


Figure 3. Surface chlorophyll RMSD (mg/m^3) and RRMSD over time.

~~consistent condition~~ for the March 5 and 12 dates of the GHOSH run (Fig. 2~~c), which is further improved by the assimilation~~ , panels **c1** and **c2**) show that the BFM model tends to predict surface bloom formation in the northwestern region as in the control run (i.e. too early bloom start). However, the assimilation on March 12 corrects this tendency, allowing the model to maintain a low concentration in the following forecast periods until the correct onset of bloom on March 26 (Fig. 2~~d).~~ , panels **d2** to **d4**). In fact, the assimilation on March 12 has changed the nutrient conditions in the surface layer so that the condition remains unfavorable for rapid phytoplankton growth and no major correction takes place in the following week (e.g. small differences between the map before and after the assimilation on March 19). During the bloom period (after March 26), assimilation corrects the some spatial mismatch of the bloom pattern, but essentially the phenology of the event is well captured. While the corrections to chlorophyll dynamics along the Algerian current appear marginal (i.e. difference between before and after maps), the GHOSH corrects the state of phytoplankton and nutrients in the euphotic zone in the eastern Mediterranean as early as the first week of March and no further significant corrections are necessary thereafter. In fact, the before assimilation maps show correctly low concentrations for the entire period.

Time series of RMSD computed between satellite and model before the assimilation are plotted in Fig. 3. Since the ~~amount~~ values of surface chlorophyll can vary considerably during the year, the relative root mean square ~~difference (RRMSD) is also provided (~~deviation (RRMSD, i.e., ~~RMSD over concentration)~~ the RMSD divided by tracer concentration) is also provided, as a normalized proxy of the skill. Figure 3 shows that the GHOSH assimilation produces an improvement in the assimilated variable (i.e., surface chlorophyll) with respect to the control run. The RMSD of the control run follows a seasonal cycle, with maximum values in late winter-early spring, corresponding to the bloom period, and minimum values during summer when the surface layer is mostly nutrient depleted and chlorophyll is at its lowest level.

The winter-spring RMSD is significantly reduced by the GHOSH filter whereas the RMSD reduction results fairly small during summer, due to a general smaller ~~amount of~~ values of modelled surface chlorophyll in that season. In fact, when the RRMSD is considered, the summer improvement due to assimilation is more detectable.

As a proxy of the overall performance of each 1-year experiment, the RMSD between the simulation and observation, averaged in space and time, is summarized in Tab. 2: the chlorophyll RMSD for *ctrl* (control run) is 0.084, while it is 0.055 for the GHOSH test, resulting in a 53% improvement.

3.1.1 ~~Nutrients~~

3.2 Nutrients

The effect of DA on non-assimilated variables has been assessed using independent observations to validate phosphate and nitrate (Tab. 2), which can both act as limiting nutrient for phytoplankton growth in the Mediterranean Sea ~~(Lazzari et al. (2016))~~(Lazzari et al., 2016).

The Hovmöller diagrams of phosphate and nitrate in the DA and the control runs show the effects of the assimilation on the two nutrient compartments (Fig. 4). Both nitrate and phosphate show lower concentrations in the 150–400 m layer from April in the simulation with satellite assimilation, with highest impact at the beginning of April. At the surface (0–100 m layer), the assimilation slightly enhances the phosphate concentrations, particularly between April and July, while it reduces the nitrate concentrations. However, these effects do not degrade the typical vertical distribution of nutrients in the Mediterranean Sea,

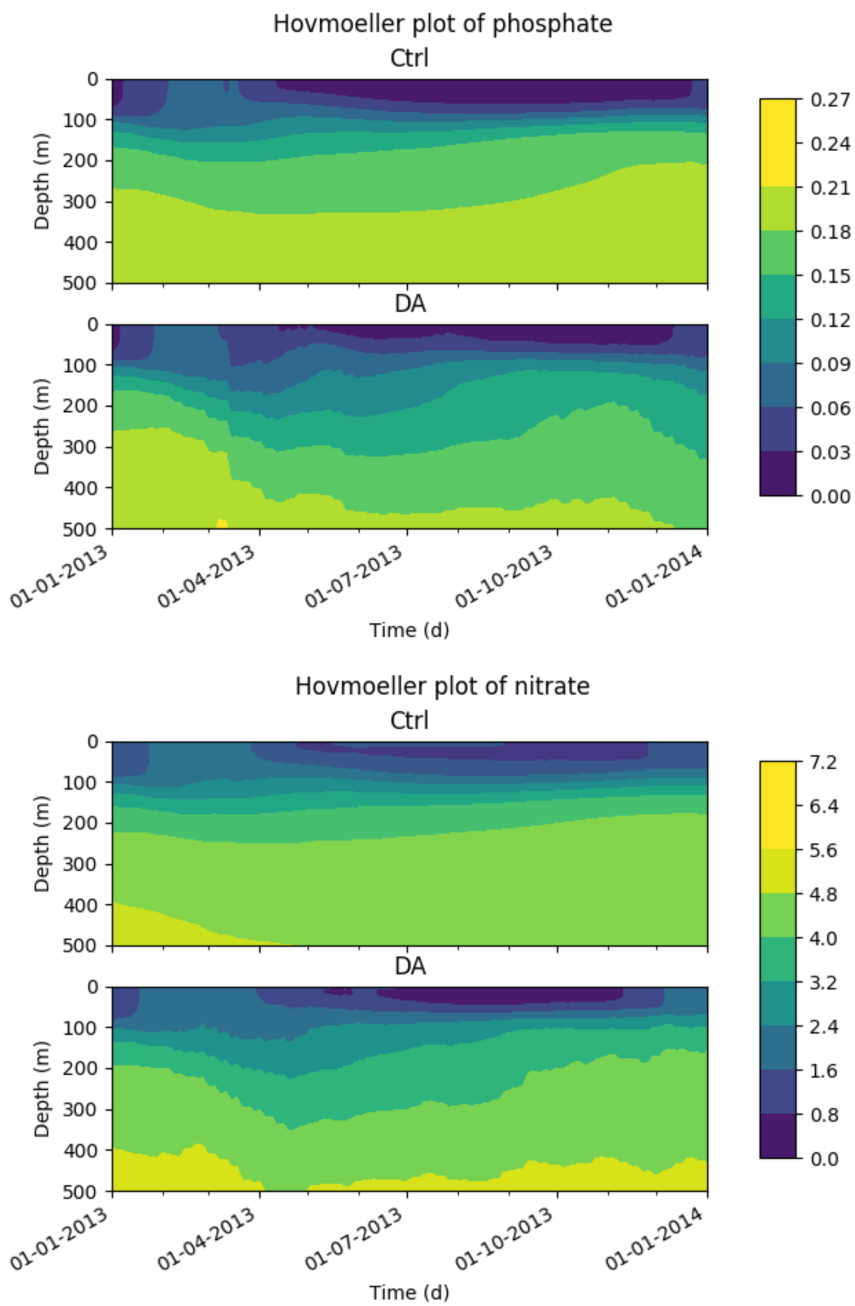


Figure 4. Hovmöller diagrams of nutrients (phosphate and nitrate, $mmol/m^3$): comparison between the control run *ctrl* (up) and ~~data~~ the assimilation ~~test *T_h5_rho0.9_d*~~ experiment (bottom)

and the seasonal sequence of late-spring/winter mixed conditions and of summer stratification is preserved.

Interestingly, the Hovmöller diagrams of Fig. 4 show that the DA is able to correct the profiles of the nutrients down to 300–400 m by changing the shape of the nutricline (i.e., deepening and creating a gentler shape) during the month of the deep chlorophyll maximum formation (from April to June) and by re-setting the nutricline to a shallower level after summer.

3.2.1 ~~Uncertainty estimation~~

3.3 Uncertainty estimation

In the GHOSH filter, a weighted ensemble is used to estimate the uncertainty of the forecast and analysis states~~(equation (??)). The weighted ensemble standard deviation~~ . The standard deviation of the weighted ensemble is used to quantify the uncertainty, whereas a comparison of the standard deviations before and after assimilation is calculated to evaluate the uncertainty reduction produced by the assimilation. Moreover, the ensemble correlation matrix is analyzed to investigate the relationships between variables. The standard deviation and correlation are shown for the assimilated variable and one of the nutrients during a bloom event in winter.

During the winter ~~event (March~~ bloom (March 26), the spatial patterns of chlorophyll and phosphate standard deviations at the surface (second row of Fig. 5) reflect those of their concentration maps (first row of Fig. 5): areas with high surface concentrations in the western Mediterranean Sea are associated with the highest values of the ensemble standard deviation given the high variability of the surface bloom dynamics. ~~The map also shows an area with low chlorophyll and high phosphate concentration in the northwestern Mediterranean where winter convection is still active and standard deviation is low.~~

As an effect of the assimilation, the areas with the highest standard deviation for both variables in the western Mediterranean Sea decrease, proving the effectiveness of the surface assimilation in constraining the surface phytoplankton bloom dynamics. Moreover, a relevant reduction in the standard deviation, in relative terms, is also observed in the less productive eastern Mediterranean Sea.

On average, the assimilation reduces the standard deviation at the surface by 33% and 22% for chlorophyll and phosphate, respectively.

A few spots of high standard deviations remain only in some coastal areas in the western Mediterranean Sea for both variables. The system has been tuned to focus on open sea and the assimilation in shallow waters is less effective due to imposed higher observation ~~errors~~ uncertainties in the coastal areas (see Section ~~??~~ 2.4 for details).

~~Maps of chlorophyll and phosphate (a and b) and maps of their uncertainty (ensemble standard deviation) before assimilation (c and d) and after assimilation (e and f) for the day 19.03.2013.~~ A multivariate correlation matrix (Figure 6) of the north western region shows that the assimilation of satellite chlorophyll influences mostly the chlorophyll in the upper 50 m layer and can impact the phosphate profile to a depth up to 400 m, with correlation values between the surface chlorophyll and phosphate decreasing with depth. In the more oligotrophic conditions of the Eastern Mediterranean Sea, chlorophyll and phosphate are less correlated than in the western Mediterranean, whereas the impact of surface assimilation on chlorophyll is deeper (i.e., correlations with surface chlorophyll is high down to nearly 400 m).

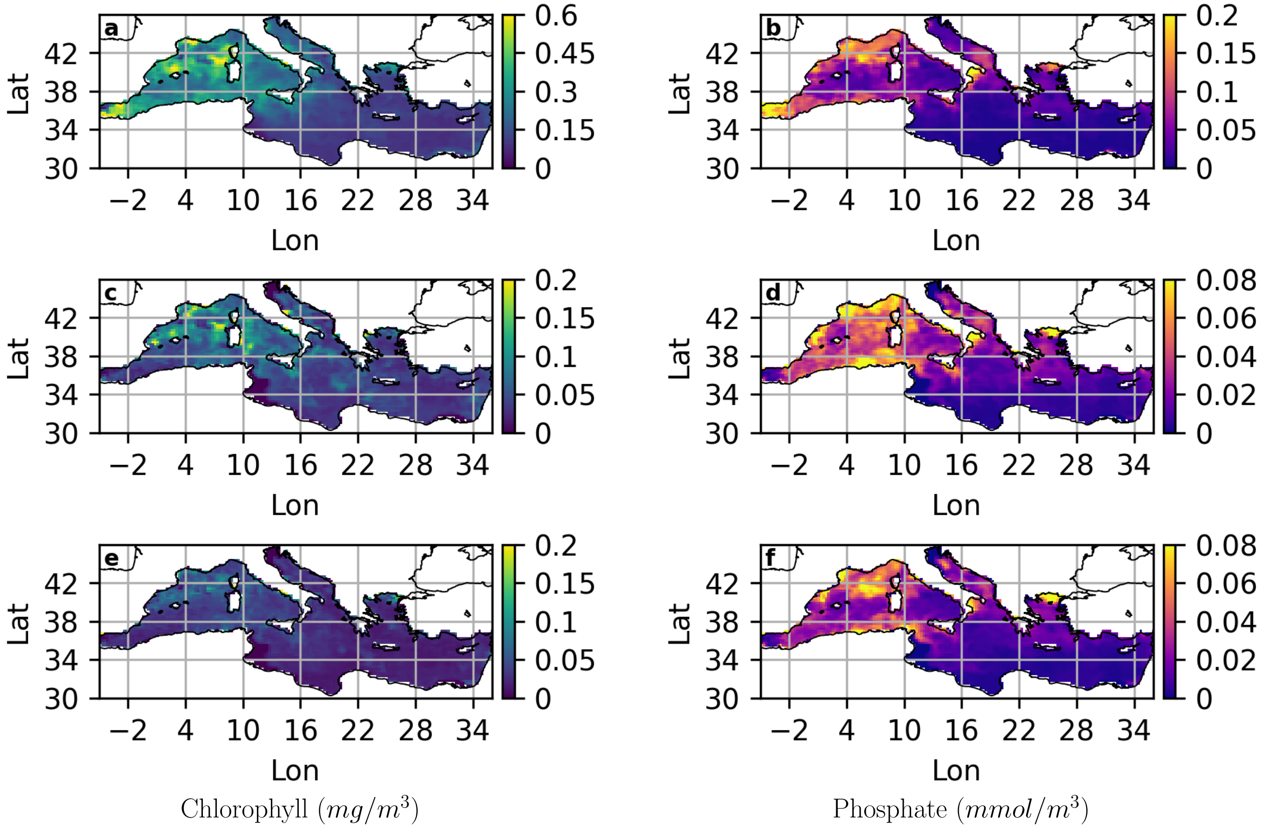
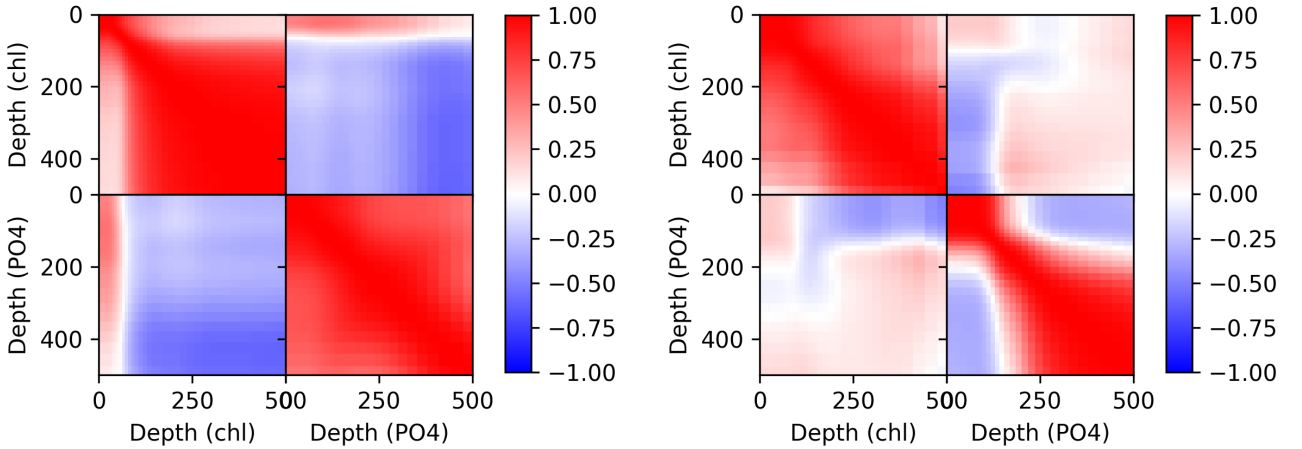


Figure 5. Maps of chlorophyll and phosphate (a and b) and maps of their uncertainty (ensemble standard deviation) before assimilation (c and d) and after assimilation (e and f) for the day 26.03.2013.

3.4 Sensitivity analysis

Performance sensitivity to the approximation order h has been evaluated by looking at RMSD scores in the tested configurations. Table 2 shows that the performance of the assimilated variable is substantially unaffected by the approximation order, whilst h demonstrates a relevant impact on a non-assimilated variable, namely the nitrate. In particular, when the order h is 5, nitrate RMSD between GHOSH and independent observations improves up to 81% with respect to order $h = 2$ (scored in the set of tests with $\rho = 0.9$ and daily forecasts). Every $h = 5$ test has a smaller RMSD than every corresponding $h = 2$ and, in general, data clearly show that increasing h implies an improvement in RMSD. In fact, the nitrate's RMSD values, centred



(a) Western Mediterranean

(b) Eastern Mediterranean

Figure 6. Post-assimilation multivariate (chlorophyll and phosphate) correlation during winter (19.03.2013). Horizontally averaged vertical profiles in the western (left) and eastern (right) Mediterranean Sea.

and normalized over their standard deviation, present a strong negative correlation with h (Pearson's correlation coefficient is -0.80).

3.4.1 ~~Three-dimensional Feasibility and computational cost~~

~~The feasibility of the GHOSH scheme in a realistic 3D operational system has been confirmed by the experiments: each~~

3.5 Computational cost

Each 1-year assimilation test, consisting of 16 ensemble members, needed approximately 4 hours to be computed in the CINECA's GALILEO super computer, running in parallel on 46 computational nodes (1656 cores).
The computational cost of the GHOSH filter itself is ~~negligible~~ marginal compared with the model integration time and the I/O operations, being ~~around~~ less than 1% of the total computational time.

4 Discussion

~~The name "Gauss-Hermite high-order sampling hybrid filter" was chosen to indicate the two main features of this novel filter. First, it is a hybrid filter, in the sense that the error covariance matrix is obtained by averaging the ensemble covariance and a background error covariance matrix, as described in Carrassi et al. (2018). Second, it exploits a novel sampling method based~~

~~on a Gauss-Hermite-like quadrature rule to reach an arbitrary high (depending on the ensemble size) polynomial order of approximation.~~

~~The latter feature introduces, to the best of our knowledge, a completely new class of DA algorithms with an order higher than 2. The advantage of moving to a higher order of approximation has previously been documented in the literature, in particular Nerger et al. (2005) compares order 1 algorithms (e.g., SEEK Pham et al. (1998)) and order 2 algorithms (e.g., SEIK Pham (1996)). Results of Section **??** confirm the expected benefits of the high order sampling adopted in the GHOSH filter. In particular, in the Lorenz96 idealized and controlled conditions, the novel method outperforms (up to 3 times better RMSE) a typical second order method like SEIK and features higher stability and better performances.At the same time, in a realistic 3D high dimensional~~ In the realistic 3D high-dimensional parallel implementation with assimilation of real satellite data, the GHOSH filter proved (in Section 3.1) to be a reliable data assimilation scheme with 30% RMSD reduction for the assimilated variable compared to the control run. Moreover, the GHOSH assimilation ~~does~~ did not degrade non-assimilated variables, actually improving their RMSDs with respect to independent observations (~~figures 3 and~~ Fig. 3 and Fig. 4). Consistently, the GHOSH assimilation results allowed catching some relevant system dynamics (e.g. formation of the winter bloom in the North Western region) and spatial and inter-variable covariances (Fig. 5 ~~). Further, the sensitivity analysis performed in the 3D application~~ and Fig. 6). Further, a sensitivity analysis underlines the role of the ~~high order~~ high-order sampling in improving performances in a non-assimilated variable (~~Tab. 2).~~ Section 3.4).

~~Thanks to the large number of settings tested in the Lorenz96 twin experiment, it is possible to suggest the conditions where the GHOSH filter considerably increases the assimilation skill and reliability. Concerning the ensemble size, the larger improvements occur when the dimension of the error subspace spanned by the ensemble is big enough to be representative. In other words, if the number of ensemble members is too small, the assimilation simply fails independently from the assimilation algorithm. Moreover, it has been shown that the GHOSH increased accuracy reduces instabilities and numerical divergence (Fig. **??**). This aspect is consistent with the fact that a higher order of approximation helps to manage non-linearities (Nerger et al. (2005)). GHOSH filter showed a lower need for inflation with respect to SEIK, removing one of the instability causes observed in the twin experiment (Section **??**). Together with the GHOSH's lower need of inflation, the hybridization adopted in the 3D application prevented the "collapse into the mean" effect (not shown), which often affects ensemble algorithms (Anderson (2007)). Indeed, the hybridization adopted in each *forecast* phase in the GHOSH 3D application has the beneficial effect of maintaining a reasonable spread, even without the forgetting factor (i.e.,~~ $\cancel{\rho = 1}$~~).~~

The results obtained coupling the GHOSH filter to the OGSTM-BFM model show a major improvement in the nitrate performance when increasing the approximation order h. However, chlorophyll and phosphate RMSDs are not substantially affected by h. Consistently with the twin experiments results ~~(Section **??**)~~in Part 1 (Spada et al., 2024), this behaviour may depend on the relatively low number of ensemble members (16, $r = 15$), which imply that only a limited number of principal components ($s = 3$) takes advantage of the higher order $h = 5$. Indeed, ~~the negligible~~ with the normalization rules chosen in this experiment (i.e., $\mathbf{\Lambda}$ in Section 2.3), the marginal effects of the higher approximation order on chlorophyll and phosphate could be motivated by the fact that ~~, with the normalization rules chosen in this experiment (i.e.,~~ $\cancel{\Lambda_i}$~~), they possibly contribute little to the few most relevant~~ they possibly slightly contribute to the first 3 principal components of the state uncertainty. A larger number of ensemble members (i.e., bigger r) would allow a larger s-subspace, probably extending the high-order effects on further state variables. This view is consistent with the generally observed benefits of a larger ensemble in ~~EnKF-based data assimilation (Edwards et al. (2015); Nerger et al. (2005); Sacher and Bartello (2008))~~ data assimilation based on EnKF (ensemble kalman filter) (see e.g., Edwards et al., 2015; Nerger et al., 2005; Sacher and Bartello, 2008).

The scaling matrix ~~$\mathbf{A}_i$~~ $\mathbf{\Lambda}$ used in the sampling procedure plays ~~a relevant~~ an important role in the filter, since ~~$\mathbf{A}_i$ defines how variables are compared to each others~~ it defines which variables are of more interest in the computation of the ~~state error~~ most relevant components ~~(Equation (??)). In the Lorenz96 case, it is straightforward to set $\mathbf{A}_i$ equal to the identity matrix, since each variable is equivalent to any other. In~~ of the state uncertainty. When implementing GHOSH in the realistic OGSTM-BFM application, we ~~adopted~~ used a normalization matrix to represent ~~$\mathbf{A}_i$ (Section ??)~~ $\mathbf{\Lambda}$ (Section 2.3) in order to properly scale the different biogeochemical variables. However, this strategy implies that the ~~most relevant~~ principal components are composed by variables that have a high correlation with as many other variables as possible, including processes that can be less relevant for an analysis that is more interested in the photic layer (e.g., ~~deep layer processes~~processes/dynamics occurring in deep layers). Alternative options can be investigated ~~and, for~~ to optimize the assimilation among the model variables. For instance, similarly to scaling strategies applied in sampling methods proposed in literature ~~(Dovera and Della Rossa (2012)), $\mathbf{A}_i$~~ (Dovera and Della Rossa, 2012), $\mathbf{\Lambda}$ could be designed to focus on the most relevant processes. In particular, ~~$\mathbf{A}_i$~~ $\mathbf{\Lambda}$ can be a sparse matrix with non-zero coordinates corresponding to variables and regions of interest (as the photic layer) based on information derived by a climatology or a multi-annual simulation.

As for ~~$\mathbf{A}_i$~~$\mathbf{\Lambda}$, a preliminary approach is applied for the definition of the ~~model error covariance matrix $\mathbf{Q}_i$ (with its local counterpart $\mathbf{Q}_i^p$) used in the 3D realistic GHOSH implementation (Section ??~~parametric uncertainty covariance matrix $\mathbf{Q}$ (Section 2.3). The diagonal matrix ~~$\mathbf{Q}_i$ is not very representative of the model error covariances of~~ $\mathbf{Q}$ can not be very representative in the relatively complex ~~OGSTm-BFM model~~OGSTM-BFM application. A more advanced parameterization of ~~$\mathbf{Q}_i^p$~~ $\mathbf{Q}$ deserves to be further investigated for future applications ~~(e.g., using a non-diagonal low rank matrix as in 3D-variational methods, Teruzzi et al. (2014)).~~

~~Compared to other ensemble filters, the key feature of the GHOSH filter is the higher polynomial order of its sampling method. The advantage of a higher order is not limited to a better estimation of the mean state but it extends to the covariance matrix of the error probability distribution. In fact, the polynomial order in the estimation of the covariance matrix is half the order of the mean~~(e.g., using a non-diagonal low rank matrix as in 3D-variational methods, Teruzzi et al., 2014). However, even in its basic implementation, the adopted hybrid approach (together with the GHOSH's lower need of inflation proved in Spa prevented the "collapse into the mean" effect (not shown)~~. Since the covariance matrix is involved in the state estimation, the more accurate GHOSH results shown in Section ?? can be partly related to an improvement in the error covariance matrix order of approximation.~~, which can affect ensemble algorithms (e.g., Anderson, 2007). Indeed, the hybridization adopted in each *forecast* phase has the beneficial effect of maintaining a reasonable spread, even in tests without forgetting factor (i.e., $\rho = 1$).

~~In addition to an higher polynomial order, the GHOSH filter features a resampling taking place twice per step. Ensemble Kalman-like filters usually adopt an interpolation strategy by applying the observation operator directly to the $\tilde{\mathbf{X}}_i^f$ ensemble~~

~~of the states after the model evolution (see equation (??))~~ The implemented statistical moments μ (Section 2.3) describe a Gaussian uncertainty pdf (see also Discussion in Spada et al., 2024). However, ~~this strategy might lead to inaccurate estimations because the model error covariance matrix $\mathbf{Q}_i$ is not taken into account in the interpolation. For this reason, GHOSH applies the observation operator to a new ensemble, resampled between the forecast and analysis phases (Fig. ??). In this way, $\mathbf{Q}_i$ is taken into account and, especially in the case of a nonlinear observation operator, the high order of~~ the ~~sampling method is exploited to improve the accuracy of the projection into the observation space.~~

~~Similarly to other ensemble filters, a weighted ensemble is used in the GHOSH filter. However, the GHOSH filter differs substantially from the particle filter (Bocquet et al. (2010)) and other types of weighted ensemble filters, since the latters evolve the value of the weights, whilst GHOSH keeps them fixed and resamples the ensemble to find new members that are~~ log-transformation of the tracers before applying the assimilation filter implies a log-Gaussian uncertainty pdf on concentrations, which is well-suited ~~for each (not evolving) weight. Adopting this approach, the problem of some particle weights quickly tending to zero, that usually affects particle filters (van Leeuwen et al. (2019)) can be avoided.~~ for non-negative biogechemical variables (see e.g., Ciavatta et al., 2016; Pradhan et al., 2019; Song et al., 2016; Ford, 2020).
On the other hand, the transformation has the side effect of changing also the observation operator (that is represented by the sum of the chlorophyll components of different phytoplankton species), which becomes non-linear. In this context, the resampling after forecast featured by the GHOSH filter is beneficial, since it exploits the high-order sampling's enhanced capability of managing non-linearity (showed in Part 1) not only with the model operator but also with the non-linear observation operator.

~~Finally, as in most ensemble methods, the computational cost of~~ From a computational point of view, the ~~GHOSH filter mainly depends on~~ negligible (less than 1%) amount of time spent on DA operations in our tests, is compliant with literature results (see e.g., Nerger et al., 2005). Indeed, the main source of computational cost is the integration of the model scaled by the ensemble size~~(Nerger et al. (2005)). The performance of GHOSH and SEIK implemented in Lorenz96 and the profiling of the GHOSH implemented in OGSTM showed that the cost of the GHOSH is no more a concern than in other order 2 data assimilation schemes (??). The benefits of the higher order offered by the GHOSH filter do not depend on a bigger computational effort, instead they rely on a mathematical advantage. In fact, the GHOSH allows the exploitation of the degrees of freedom in the randomness of the sampling process, choosing ensembles that have a better consistency with the error distribution~~. This behaviour is also in line with the GHOSH asymptotic computational complexity that is the same as in EnKF second-order schemes (see Part 1 Spada et al., 2024).

5 Conclusions

~~This work presented a novel ensemble hybrid DA filter (GHOSH). The GHOSH filter features a higher order of approximation compared to other ensemble filters, which resulted in better data assimilation performances.The orderof an ensemble method is one of the most relevant proxies of a filter skill. In fact, the order 2 schemes (e.g., SEIK, ETKF, ESTKF) are peferred nowadays~~

~~to older order 1 methods (e.g., SEEK). This study represents a natural further step in this direction, opening to a new class of data assimilation schemes with order greater than 2.~~

~~The outstanding improvements of the high order GHOSH filter with respect to the SEIK filter are demonstrated in an extensive set of twin experiments based on the Lorenz96 model. Higher RMSE reductions (GHOSH error up to three times smaller than SEIK) occurred in case of small inflation, large number of observations or large ensemble size. Considering each filter tuned with its optimal forgetting factor, the GHOSH filter showed improved capacity to take into account non-linearities by a lesser need of inflation with respect to SEIK, and achieved the best RMSE reductions (GHOSH error lower than half of the SEIK error)in cases of moderate ensemble size~~ GHOSH is a novel DA ensemble filter that was introduced in Part 1, and features an enhanced approximation order.

~~The feasibility of the GHOSH filter in a realistic geophysical application has also been assessed by testing a~~ In this work, it has been tested in a 3D ~~parallel implementation of a Mediterranean physical-biogeochemical model . The~~ parallel implementation of a state-of-the-art Mediterranean physical-biogeochemical model with assimilation of ocean color observations. Testing a GHOSH implementation in a computationally demanding realistic application was a mandatory step to assess the suitability of the novel filter in the geosciences context. Moreover, the challenge of tracking inter-variables covariances in the complex processes characterizing biogeochemical applications represents an interesting benchmark.

Our results show that, from a computational point of view, the cost of the GHOSH filter is comparable to other second-order deterministic filters (e.g., SEIK, ETKF), since the computational cost chiefly depends on the ensemble size that proportionally expands the model integration cost.

Furthermore, GHOSH improved the agreement between the forecast and observations without producing unrealistic effects on the non-assimilated variables. Considering relevant marine ecosystem dynamics, GHOSH filter successfully constrained the uncertainty of the phytoplankton winter bloom events while assimilating ocean color. ~~Furthermore~~Moreover, a sensitivity analysis indicates that the use of a higher order has positive impacts on the performance of a non-assimilated variable (e.g., nutrients).

~~From the computational point of view, the implementation of the GHOSH filter proved to be as demanding as the SEIK, given that the computational cost is dominated by the integration of the model and by the ensemble size.~~

Code availability. A GHOSH python implementation is available from the GitHub page: https://github.com/Sword-Code/PythonDA under the licence GNU GPLv3. The FORTRAN implementation of the GHOSH is based on the OGSTM model (available at https://github.com/inogs/ogstm) coupled with the BFM model (available at https://bfm-community.github.io/www.bfm-community.eu/). The exact version of the OGSTM-BFM-GHOSH system used to produce results and plots described in Section 3.1 is archived on Zenodo (https://doi.org/10.5281/zenodo.12819521).

Appendix A: ~~High-order sampling~~

~~Sampling a limited number of ensemble members that effectively represent the uncertainty of the state estimation is a crucial part of any ensemble algorithm. The GHOSH filter uses multi-dimensional Gauss-Hermite-like quadrature rules to choose the ensemble members, achieving a high polynomial order of convergence.~~

~~The core idea behind the GHOSH sampling can be proved by taking two independent random variables with the same moments up to a certain~~

Appendix A: High-order sampling solutions

Given a certain maximum approximation order h ~~. Then, they have the same mean after applying any polynomial operator of order h. In fact, if $\mathcal{F}:\mathbb{R}^s \longrightarrow \mathbb{R}$ is a polynomial operator such that~~

$$\mathcal{F}(\boldsymbol{x}) = \sum_{\xi=0}^{h} \sum_{j_1,\ldots,j_\xi=1}^{s} f_\xi^{j_1,\ldots,j_\xi} x_{j_1} \cdots x_{j_\xi},$$

~~where s is the space dimension, $f_\xi^{j_1,\ldots,j_\xi}$ are the polynomial coefficients and $x_1,\ldots,x_s$ are the coordinates of $\boldsymbol{x}$, then the mean $\bar{\mathcal{F}}$,~~

$$\bar{\mathcal{F}} = \int_{\mathbb{R}^s} \mathcal{F}(\boldsymbol{x})\, p(\boldsymbol{x})\, dx = \sum_{\xi=0}^{h} \sum_{j_1,\ldots,j_\xi=1}^{s} f_\xi^{j_1,\ldots,j_\xi} \int_{\mathbb{R}^s} x_{j_1} \cdots x_{j_\xi} p(\boldsymbol{x})\, d\boldsymbol{x},$$

~~depends only on the first h moments of the probability density function p, which are represented by the last integral in equation (**??**) .~~

~~This proves that the mean of an operator can be computed exactly up to a certain order h by substituting the sampled probability distribution with any discrete finite probability distribution (i.e., a weighted ensemble) as long as their moments match (up to order h) .In order to build an ensemble with this property, the moment-matching equations are summarized in the nonlinear system~~

$$\left\{ \begin{array}{l} \vdots \\ \displaystyle\sum_{m=1}^{r+1} x_{j_1}^m \cdots x_{j_\xi}^m w_m = \mu_{j_1,\ldots,j_\xi}, \\ \vdots \end{array} \right.$$

~~with one equation for each $\xi \in \{0,\ldots,h\}$ and for each $j_1,\ldots,j_\xi \in \{1,\ldots,s\}$. In system (**??**), x_j^m is the j-th coordinate of the m-th ensemble member, $r+1$ is the ensemble size, $w_m \geq 0$ are the weights~~ and ~~$\mu_{j_1,\ldots,j_\xi}$ are the statistical moments of p, i.e. ,~~

$$\mu_{j_1,\ldots,j_\xi} = \int_{\mathbb{R}^s} x_{j_1} \cdots x_{j_\xi} p(\boldsymbol{x})\, d\boldsymbol{x}.$$

~~The system is solved considering x_j^m~~ the dimension s of the subspace in which the h^{th}-order approximation is achieved, the GHOSH filter (during its initialization) needs a solution of a specific non-linear system to encode the moments information (μ) in the GHOSH sampling matrix $\boldsymbol{\Omega}^h$ and in the vector $\boldsymbol{v}$ of the square roots of the ensemble weights (Spada et al., 2024). Most of the time, the non-linear system admits more than one solution, hence different $\boldsymbol{\Omega}^h$ matrices (and ~~w_m as the unknowns, while p can be taken uncorrelated, normalized and with 0 mean without loss of generality. In fact, the ensemble matrix $\tilde{\mathbf{X}} = (x_j^m)$ (which contains in its columns the ensemble members with 0 mean and covariance matrix equal to the identity matrix $\mathbf{I}_s$) can be transformed in the ensemble matrix X with arbitrary mean $\bar{\boldsymbol{x}}$ and covariance matrix LL^T through the projection~~

$$\mathbf{X} = \bar{\boldsymbol{x}}\mathbb{1}_{1\times(r+1)} + \mathbf{L}\tilde{\mathbf{X}},$$

~~where $\mathbb{1}_{1\times(r+1)}$ is a matrix (of the subscripted size)filled with ones~~corresponding $\boldsymbol{v}$ vectors) are possible. Since $\boldsymbol{\Omega}^h$ and $\boldsymbol{v}$ have a central role in the GHOSH sampling strategy, this appendix reports their values used in the GHOSH configurations explored in this work (Section 2.4).

In the ~~special case of the standard normal distribution, i.e., $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}, 0, \mathbf{I}_s)$, the moments $\mu_{j_1,\ldots,j_\xi}$ are given by~~

$$\int\limits_{\mathbb{R}^s} {x_1}^{k_1}\cdots {x_s}^{k_s}\mathcal{N}(\boldsymbol{x};0,\mathbf{I}_s)\,d\boldsymbol{x} = 0$$

~~if any exponent k_j is an odd number, or~~

$$\int\limits_{\mathbb{R}^s} {x_1}^{k_1}\cdots {x_s}^{k_s}\mathcal{N}(\boldsymbol{x};0,\mathbf{I}_s)\,dx = \frac{k_1!}{\sqrt{2}^{k_1}\left(\frac{k_1}{2}!\right)}\cdots\frac{k_s!}{\sqrt{2}^{k_s}\left(\frac{k_s}{2}!\right)}$$

~~otherwise. Further, the solutions x_j^m and w_m are the nodes and weights of the Gauss-Hermite quadrature rule.In general, by applying to system (**??**) the substitution~~

$$\begin{cases} w_m = {v_m}^2, \\ x_j^m = \dfrac{u_j^m}{v_m}, \end{cases}$$

~~the weights are naturally forced to be positive, and the system can be conveniently rewritten as~~

$$\begin{cases} \vdots \\ \displaystyle\sum_{m=1}^{r+1} u_{j_1}^m \cdots u_{j_\xi}^m {v_m}^{2-\xi} = \mu_{j_1,\ldots,j_\xi}, \\ \vdots \end{cases}$$

~~which, in the case of order $h = 2$, can be expressed in matrix form as~~

$$\left(\begin{array}{c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h \end{array}\right)^T \left(\begin{array}{c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h \end{array}\right) = \mathbf{I}_s,$$

~~where $\boldsymbol{v} \in \mathbb{R}^{r+1}$ is the column vector with coordinates v_m and $\boldsymbol{\Omega}^h$ is an $(r+1)\times s$ matrix with coordinates u_j^m. Note that equation (**??**) can be solved for any unit~~ case with $h=2$ and $s=15$, the high-order sampling method reduces to the traditional second-order-exact sampling (Pham, 2001). The solution is easily obtained starting from a constant vector $\boldsymbol{v}$ and ~~for any value of $s \leq r$. If constant weights are chosen and $s=r$, then this method is equivalent to the second-order exact sampling of the SEIK filter (Pham (1996)), which leads to, in the formalism of equation (**??**),~~

$$\mathbf{X} = \bar{\boldsymbol{x}}\mathbb{1}_{1\times(r+1)} + \mathbf{L}\boldsymbol{\Omega}^{h^T}\mathbf{W}^{-\frac{1}{2}},$$

~~where $\mathbf{W} = \mathrm{diag}(\boldsymbol{w})$ is the $(r+1)\times(r+1)$ diagonal matrixof weights.~~

~~In general, if $h>2$~~adding orthonormal columns in order to form an orthogonal matrix.

In the case with $h=3$ and $s=8$, a solution with constant weights ($v_m = \frac{1}{4}$ for each $m \in \{1,\ldots,16\}$) is chosen, ~~s must be lower than r to obtain a solution to system (**??**). The obtained s-dimensional ensemble can be extended to an r-dimensional one: its projection on the s-dimensional subspace defined by the first s coordinates will retain order h, while the remaining components are sampled with order 2. This is achieved by starting from $\boldsymbol{v}$ and $\boldsymbol{\Omega}^h$, given by a particular solution of system (**??**). Then, exploiting symmetry and rotational invariance, randomness is added to avoid biases by multiplying $\boldsymbol{\Omega}^h$ by any random $s\times s$ orthogonal matrix $\boldsymbol{\Omega}^{rnd}$. Finally, the $(r+1)\times(r-s)$ matrix $\boldsymbol{\Omega}^\perp$ is chosento fulfil~~

$$\left(\begin{array}{c|c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h\boldsymbol{\Omega}^{rnd} & \boldsymbol{\Omega}^\perp \end{array}\right)^T \left(\begin{array}{c|c|c} \boldsymbol{v} & \boldsymbol{\Omega}^h\boldsymbol{\Omega}^{rnd} & \boldsymbol{\Omega}^\perp \end{array}\right) = \mathbf{I}_{r+1}.$$

~~The sampling matrix $\boldsymbol{\Omega}$ is defined as~~

$$\boldsymbol{\Omega} = \left(\begin{array}{c|c} \boldsymbol{\Omega}^h\boldsymbol{\Omega}^{rnd} & \boldsymbol{\Omega}^\perp \end{array}\right).$$

~~Now the columns of $\boldsymbol{\Omega}^T W^{-\frac{1}{2}}$ represent an r-dimensional ensemble of order h in the first s coordinates and order 2 in the last $r-s$ coordinates. Equation (**??**) needs to be modified to properly project the higher order subspace in the most relevant components of the covariance matrix $\mathbf{L}\mathbf{L}^T$. It can be proved that such components, ordered by relevance, are the column of the matrix $\mathbf{L}\mathbf{C}^T$, where $\mathbf{C}$ is a $r\times r$ orthogonal change-of-basis matrix such that~~ namely,

$$\mathbf{L}^T\boldsymbol{\Lambda}^{-1}\mathbf{L}\boldsymbol{\Omega}^{h=3} = \mathbf{C}\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}^T \mathbf{E}\mathbf{C}. \tag{A1}$$

In the ~~last equation, the right-hand side is an eigenvalue decomposition of the left-hand side, with E being an $r \times r$ diagonal matrix of eigenvalues in descending order and $\mathbf{\Lambda}$ being an appropriate $N \times N$ matrix. The purpose of $\mathbf{\Lambda}$ has already been discussed in Section ??.After the change of basis induced by C, the most relevant components are on the left side of the matrix product LC^T, which is compliant with the ensemble stored in $\mathbf{\Omega}^T W^{-\frac{1}{2}}$, and the sampling is finally obtained by~~

$$\mathbf{X} = \bar{\boldsymbol{x}}\mathbb{1}_{1\times(r+1)} + \mathbf{L}\mathbf{C}^T {\mathbf{\Omega}^h}^T \mathbf{W}^{-\frac{1}{2}}.$$

case with $h = 5$ and $s = 3$, the chosen solution is

$$\boldsymbol{v} = \begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{1}{2\sqrt{5}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \cdots \\ \cdots & \frac{1}{2\sqrt{5}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} \end{pmatrix}^T \tag{A2}$$

and

$$\mathbf{\Omega}^{h=5} = \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \cdots \\ 0 & 0 & 0 & \frac{1}{\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & \cdots \\ 0 & 0 & 0 & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \cdots \end{pmatrix.$$
$$\begin{matrix} \cdots & -\frac{1}{2} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} \\ \cdots & 0 & \frac{1}{\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} \\ \cdots & 0 & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} \end{matrix}\Bigg)^T. \tag{A3}$$

Author contributions. GC, S Salon and S Spada designed the study. S Spada developed the algorithms, supervised by SM. S Spada, AT and GC designed the experiments. S Spada implemented the code and performed the simulations collaborating with AT. S Spada wrote the manuscript, with the contribution of AT, GC and CS. All the authors participated in acquiring the funding for the project.

Competing interests. The authors declare no competing interests.

Acknowledgements. We acknowledge the CINECA award under the ISCRA initiative and the HPC-TRES program, for the availability of high-performance computing resources.
This study has been conducted using E.U. Copernicus Marine Service data.
The research reported in this work was supported by OGS and by the SEAMLESS project (https://www.seamlessproject.org/).

This work was partly funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004032 (project SEAMLESS).

References

Anderson, J. L.: An adaptive covariance inflation error correction algorithm for ensemble filters, Tellus A: Dynamic Meteorology and Oceanography, https://doi.org/10.1111/j.1600-0870.2006.00216.x, 2007.

Bannister, R.: A review of operational methods of variational and ensemble-variational data assimilation, Quarterly Journal of the Royal Meteorological Society, 143, 607–633, https://doi.org/10.1002/qj.2982, 2017a.

Bannister, R. N.: A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics, Quarterly Journal of the Royal Meteorological Society, 134, 1971–1996, https://doi.org/10.1002/qj.340, _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.340, 2008.

Bannister, R. N.: A review of operational methods of variational and ensemble-variational data assimilation, Quarterly Journal of the Royal Meteorological Society, 143, 607–633, https://doi.org/https://doi.org/10.1002/qj.2982, 2017b.

Bishop, C. H., Etherton, B. J., and Majumdar, S. J.: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects, Monthly Weather Review, 129, 420 – 436, https://doi.org/https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2, 2001.

Bocquet, M., Pires, C., and Wu, L.: Beyond Gaussian Statistical Modeling in Geophysical Data Assimilation, Monthly Weather Review, 138, 2997–3023, https://doi.org/10.1175/2010MWR3164.1, 2010.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L.: Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model, Journal of Computational Science, 44, 101 171, https://doi.org/https://doi.org/10.1016/j.jocs.2020.101171, 2020.

Buga, L., Sarbu, G., Fryberg, L., Magnus, W., Wesslander, K., Gatti, J., Leroy, D., Iona, S., Larsen, M., Koefoed Rømer, J., Østrem, A., Lipizer, M., and Giorgietti, A.: EMODnet Chemistry Eutrophication and Acidity aggregated datasets v2018, https://doi.org/https://doi.org/10.6092/EC8207EF-ED81-4EE5-BF48-E26FF16BF02E, 2018.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives, Wires climate change, 9, 2018.

Ciavatta, S., Kay, S., Saux-Picart, S., Butenschön, M., and Allen, J. I.: Decadal reanalysis of biogeochemical indicators and fluxes in the North West European shelf-sea ecosystem, Journal of Geophysical Research: Oceans, 121, 1824–1845, https://doi.org/https://doi.org/10.1002/2015JC011496, 2016.

Cossarini, G., Lazzari, P., and Solidoro, C.: Spatiotemporal variability of alkalinity in the Mediterranean Sea, Biogeosciences, 12(6), 1647–1658, 2015.

Cossarini, G., Mariotti, L., Feudale, L., Mignot, A., Salon, S., Taillandier, V., Teruzzi, A., and D'Ortenzio, F.: Towards operational 3D-Var assimilation of chlorophyll Biogeochemical-Argo float data into a biogeochemical model of the Mediterranean Sea, Ocean Modelling, 133, 112–128, https://doi.org/10.1016/j.ocemod.2018.11.005, 2019.

Cossarini, G., Feudale, L., Teruzzi, A., Bolzon, G., Coidessa, G., Solidoro, C., Di Biagio, V., Amadio, C., Lazzari, P., Brosich, A., and Salon, S.: High-Resolution Reanalysis of the Mediterranean Sea Biogeochemistry (1999–2019), Frontiers in Marine Science, 8, https://doi.org/10.3389/fmars.2021.741486, 2021.

Di Biagio, V., Salon, S., Feudale, L., and Cossarini, G.: Subsurface oxygen maximum in oligotrophic marine ecosystems: mapping the interaction between physical and biogeochemical processes, Biogeosciences, 19, 5553–5574, https://doi.org/10.5194/bg-19-5553-2022, 2022.

Dovera, L. and Della Rossa, E.: Improved initial ensemble generation coupled with ensemble square root filters and inflation to estimate uncertainty, Computational Geosciences, 16, 437–453, https://doi.org/10.1007/s10596-011-9243-5, 2012.

Dowd, M., Jones, E., and Parslow, J.: A statistical overview and perspectives on data assimilation for marine biogeochemical models, Environmetrics, 25, 203–213, https://doi.org/https://doi.org/10.1002/env.2264, 2014.

Edwards, C. A., Moore, A. M., Hoteit, I., and Cornuelle, B. D.: Regional Ocean Data Assimilation, Annual Review of Marine Science, 7, 21–42, https://doi.org/10.1146/annurev-marine-010814-015821, pMID: 25103331, 2015.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, Journal of Geophysical Research: Oceans, 99, 10 143–10 162, https://doi.org/10.1029/94JC00572, 1994.

Fennel, K., Gehlen, M., Brasseur, P., Brown, C. W., Ciavatta, S., Cossarini, G., Crise, A., Edwards, C. A., Ford, D., Friedrichs, M. A. M., Gregoire, M., Jones, E., Kim, H.-C., Lamouroux, J., Murtugudde, R., Perruche, C., , t. G. O. M. E. A., and Team, P. T.: Advancing Marine Biogeochemical and Ecosystem Reanalyses and Forecasts as Tools for Monitoring and Managing Ecosystem Health, Frontiers in Marine Science, 6, https://doi.org/10.3389/fmars.2019.00089, 2019.

Fennel, K., Mattern, J. P., Doney, S. C., Bopp, L., Moore, A. M., Wang, B., and Yu, L.: Ocean biogeochemical modelling, Nature Reviews Methods Primers, 2, 76, 2022.

Ford, D. and Barciela, R.: Global marine biogeochemical reanalyses assimilating two different sets of merged ocean colour products, Remote Sensing of Environment, 203, https://doi.org/10.1016/j.rse.2017.03.040, 2017.

Ford, D. A.: Assessing the role and consistency of satellite observation products in global physical–biogeochemical ocean reanalysis, Ocean Science, 16, 875–893, https://doi.org/https://doi.org/10.5194/os-16-875-2020, publisher: Copernicus GmbH, 2020.

Foujols, M.-A., Lévy, M., Aumont, O., and Madec, G.: OPA 8.1 Tracer Model reference manual, Project report, Institut Pierre-Simon Laplace (IPSL), https://eprints.soton.ac.uk/64848/, 2000.

Fowler, A. M., Skákala, J., and Ford, D.: Validating and improving the uncertainty assumptions for the assimilation of ocean-colour-derived chlorophyll a into a marine biogeochemistry model of the Northwest European Shelf Seas, Quarterly Journal of the Royal Meteorological Society, 149, 300–324, https://doi.org/https://doi.org/10.1002/qj.4408, 2023.

Hamill, T. M. and Snyder, C.: A Hybrid Ensemble Kalman Filter–3D Variational Analysis Scheme, Monthly Weather Review, 128, 2905 – 2919, https://doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2, 2000.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Houtekamer, P. L. and Zhang, F.: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation, Monthly Weather Review, 144, 4489 – 4532, https://doi.org/https://doi.org/10.1175/MWR-D-15-0440.1, 2016.

Hu, J., Fennel, K., Mattern, J. P., and Wilkin, J.: Data assimilation with a local Ensemble Kalman Filter applied to a three-dimensional biological model of the Middle Atlantic Bight, Journal of Marine Systems, 94, 145–156, https://doi.org/https://doi.org/10.1016/j.jmarsys.2011.11.016, 2012.

Ito, K. and Xiong, K.: Gaussian filters for nonlinear filtering problems, IEEE Transactions on Automatic Control, 45, 910–927, https://doi.org/10.1109/9.855552, 2000.

Janjic, T., Nerger, L., Albertella, A., Schroter, J., and Skachko, S.: On Domain Localization in Ensemble-Based Kalman Filter Algorithms, Monthly Weather Review, 139, 2046–2060, 2011.

Janjić, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., Losa, S. N., Nichols, N. K., Potthast, R., Waller, J. A., and Weston, P.: On the representation error in data assimilation, Quarterly Journal of the Royal Meteorological Society, 144, 1257–1278, https://doi.org/https://doi.org/10.1002/qj.3130, 2018.

Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems, Transaction of the ASME - Journal of Basic Engineering, pp. 35–45, 1960.

Lahoz, W. A. and Schneider, P.: Data assimilation: making sense of Earth Observation, Frontiers in Environmental Science, 2, https://doi.org/10.3389/fenvs.2014.00016, 2014.

Lazzari, P., Teruzzi, A., Salon, S., Campagna, S., Calonaci, C., Colella, S., Tonani, M., and Crise, A.: Pre-operational short-term forecasts for the Mediterranean Sea biogeochemistry, Ocean Science, 6, 25–39, 2010.

Lazzari, P., Solidoro, C., Ibello, V., Salon, S., Teruzzi, A., Béranger, K., Colella, S., and Crise, A.: Seasonal and inter-annual variability of plankton chlorophyll and primary production in the Mediterranean Sea: a modelling approach, Biogeosciences, 9, 217–233, 2012.

Lazzari, P., Solidoro, C., Salon, S., and Bolzon, G.: Spatial variability of phosphate and nitrate in the Mediterranean Sea: A modeling approach., Deep Sea Research Part I: Oceanographic Research, 108, 39 – 52, https://doi.org/https://doi.org/10.1016/j.dsr.2015.12.006, 2016.

Lorenz, E. N.: Predictability: A problem partly solved, in: Proc. Seminar on predictability, vol. 1, Reading, 1996.

Mamnun, N., Völker, C., Vrekoussis, M., and Nerger, L.: Uncertainties in ocean biogeochemical simulations: Application of ensemble data assimilation to a one-dimensional model, Frontiers in Marine Science, 9, https://doi.org/10.3389/fmars.2022.984236, 2022.

Mastroianni, G. and Milovanovic, G. V.: Interpolation Processes: Basic Theory and Applications, Springer Publishing Company, Incorporated, 1 edn., 2008.

Mattern, J. P., Edwards, C. A., and Moore, A. M.: Improving Variational Data Assimilation through Background and Observation Error Adjustments, Monthly Weather Review, 146, 485–501, https://doi.org/10.1175/MWR-D-17-0263.1, 2017.

Melaku Canu, D., Ghermandi, A., Nunes, P., Lazzari, P., Cossarini, G., and Solidoro, C.: Estimating the value of carbon sequestration ecosystem services in the Mediterranean Sea: An ecological economics approach, Global Environmental Change, 32, 87–95, 2015.

Moore, A. M., Martin, M. J., Akella, S., Arango, H. G., Balmaseda, M., Bertino, L., Ciavatta, S., Cornuelle, B., Cummings, J., Frolov, S., Lermusiaux, P., Oddo, P., Oke, P. R., Storto, A., Teruzzi, A., Vidard, A., and Weaver, A. T.: Synthesis of Ocean Observations Using Data Assimilation for Operational, Real-Time and Reanalysis Systems: A More Complete Picture of the State of the Ocean, Frontiers in Marine Science, 6, https://doi.org/10.3389/fmars.2019.00090, 2019.

Nerger, L. and Gregg, W.: Assimilation of SeaWiFS data into a global ocean-biogeochemical model using a local SEIK filter, Journal of Marine Systems, 68, 237–254, https://doi.org/10.1016/j.jmarsys.2006.11.009, 2007.

Nerger, L., Hiller, W., and Schröter, J.: A comparison of error subspace Kalman filters, Tellus A, 57, 715–735, https://doi.org/https://doi.org/10.1111/j.1600-0870.2005.00141.x, 2005.

Nerger, L., Janjic Pfander, T., Schröter, J., and Hiller, W.: A Unification of Ensemble Square Root Kalman Filters, Monthly Weather Review, 140, 2335–2345, https://doi.org/10.1175/MWR-D-11-00102.1, 2012.

Pham, D. T.: A Singular Evolutive Interpolated Kalman filter for data assimilation in Oceanography, 1996.

Pham, D. T.: Stochastic Methods for Sequential Data Assimilation in Strongly Nonlinear Systems, Monthly Weather Review, 129, 1194 – 1207, https://doi.org/10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2, 2001.

Pham, D. T., Verron, J., and Roubaud, M. C.: A Singular Evolutive Extended Kalman filter for data assimilation in Oceanography, Journal of Marine Systems, 16, 323–340, 1998.

Pradhan, H. K., Völker, C., Losa, S. N., Bracher, A., and Nerger, L.: Assimilation of Global Total Chlorophyll OC-CCI Data and Its Impact on Individual Phytoplankton Fields, Journal of Geophysical Research: Oceans, 124, 470–490, https://doi.org/10.1029/2018JC014329, 2019.

Sacher, W. and Bartello, P.: Sampling Errors in Ensemble Kalman Filtering. Part I: Theory, Monthly Weather Review, 136, 3035 – 3049, https://doi.org/https://doi.org/10.1175/2007MWR2323.1, 2008.

Salon, S., Cossarini, G., Bolzon, G., Feudale, L., Lazzari, P., Teruzzi, A., Solidoro, C., and Crise, A.: Novel metrics based on Biogeochemical Argo data to improve the model uncertainty evaluation of the CMEMS Mediterranean marine ecosystem forecasts, Ocean Science, 15, 997–1022, https://doi.org/https://doi.org/10.5194/os-15-997-2019, publisher: Copernicus GmbH, 2019.

Santana-Falcón, Y., Brasseur, P., Brankart, J. M., and Garnier, F.: Assimilation of chlorophyll data into a stochastic ensemble simulation for the North Atlantic Ocean, Ocean Science, 16, 1297–1315, https://doi.org/https://doi.org/10.5194/os-16-1297-2020, publisher: Copernicus GmbH, 2020.

Song, H., Edwards, C. A., Moore, A. M., and Fiechter, J.: Data assimilation in a coupled physical-biogeochemical model of the California current system using an incremental lognormal 4-dimensional variational approach: Part 3—Assimilation in a realistic context using satellite and in situ observations, Ocean Modelling, 106, 159–172, https://doi.org/10.1016/j.ocemod.2016.06.005, 2016.

Spada, S., Teruzzi, A., Maset, S., Salon, S., Solidoro, C., and Cossarini, G.: GHOSH v1.0.0: a novel Gauss-Hermite High-Order Sampling Hybrid ensemble filter for computationally efficient data assimilation in geosciences - Part 1, https://doi.org/10.5194/gmd-2023-170, preprint version, 2024.

Talagrand, O. and Courtier, P.: Variational Assimilation of Meteorological Observations With the Adjoint Vorticity Equation. I: Theory, Quarterly Journal of the Royal Meteorological Society, 113, 1311–1328, https://doi.org/https://doi.org/10.1002/qj.49711347812, 1987.

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M., and Zhen, Y.: A Review of Innovation-Based Methods to Jointly Estimate Model and Observation Error Covariance Matrices in Ensemble Data Assimilation, Monthly Weather Review, 148, 3973–3994, https://doi.org/10.1175/mwr-d-19-0240.1, 2020.

Teruzzi, A., Dobricic, S., Solidoro, C., and Cossarini, G.: A 3-D variational assimilation scheme in coupled transport-biogeochemical models: Forecast of Mediterranean biogeochemical properties, J. Geophys. Res. Oceans, 119, 200–217, 2014.

Teruzzi, A., Bolzon, G., Salon, S., Lazzari, P., Solidoro, C., and Cossarini, G.: Assimilation of coastal and open sea biogeochemical data to improve phytoplankton simulation in the Mediterranean Sea, Ocean Modelling, 132, 46 – 60, https://doi.org/https://doi.org/10.1016/j.ocemod.2018.09.007, 2018a.

Teruzzi, A., Bolzon, G., Salon, S., Lazzari, P., Solidoro, C., and Cossarini, G.: Assimilation of coastal and open sea biogeochemical data to improve phytoplankton simulation in the Mediterranean Sea, Ocean Modelling, 132, 46–60, https://doi.org/10.1016/j.ocemod.2018.09.007, 2018b.

Teruzzi, A., Bolzon, G., Feudale, L., and Cossarini, G.: Deep chlorophyll maximum and nutricline in the Mediterranean Sea: emerging properties from a multi-platform assimilated biogeochemical model experiment, Biogeosciences Discussions, 2021, 1–29, https://doi.org/10.5194/bg-2021-97, 2021.

Triantafyllou, G., Hoteit, I., and Petihakisa, G.: A singular evolutive interpolated Kalman filter for efficient data assimilation in a 3-D complex physical-biogeochemical model of the Cretan Sea, Journal of Marine Systems, 40–41, 213–231, 2003.

Tsiaras, K., Hoteit, I., Kalaroni, S., Petihakis, G., and Triantafyllou, G.: A hybrid ensemble-OI Kalman filter for efficient data assimilation into a 3-D biogeochemical model of the Mediterranean, Ocean Dynamics, 67, 673–690, https://doi.org/10.1007/s10236-017-1050-7, 2017.

van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., and Reich, S.: Particle filters for high-dimensional geoscience applications: A review, Quarterly Journal of the Royal Meteorological Society, 145, 2335–2365, https://doi.org/https://doi.org/10.1002/qj.3551, 2019.

Volpe, G., Pitarch, J., Colella, S., Brando, V., Forneris, V., Bracaglia, M., and Benincasa, M.: OCEAN COLOUR PRODUCTION CENTRE, QUality Information Document of Ocean Colour Mediterranean and Black Sea Observation Product v1.5. Technical Report., 2017.

Wan, E. and Van Der Merwe, R.: The unscented Kalman filter for nonlinear estimation, in: Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373), pp. 153–158, https://doi.org/10.1109/ASSPCC.2000.882463, 2000.

Wang, B., Fennel, K., Yu, L., and Gordon, C.: Assessing the value of biogeochemical Argo profiles versus ocean color observations for biogeochemical model optimization in the Gulf of Mexico, Biogeosciences, 17, 4059–4074, https://doi.org/10.5194/bg-17-4059-2020, 2020.