

# A novel Gauss-Hermite High-Order Sampling Hybrid ensemble filter for computationally efficient data assimilation in geosciences - Part 1: PythonDA v1.2.1

Simone Spada<sup>1</sup>, Anna Teruzzi<sup>1</sup>, Stefano Maset<sup>2</sup>, Stefano Salon<sup>1</sup>, Cosimo Solidoro<sup>1</sup>, and Gianpiero Cossarini<sup>1</sup>

<sup>1</sup>National Institute of Oceanography and Applied Geophysics - OGS, 34010 Trieste, Italy

<sup>2</sup>University of Trieste, 34127 Trieste, Italy

**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 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. In the directions where the uncertainty is larger, the GHOSH 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 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. The twin-experiment results show that GHOSH outperforms SEIK in most of the assimilation settings up to a 70% reduction of the root mean square error on assimilated and 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 (among the others, see the methods and applications reviewed in Carrassi et al., 2018; Houtekamer and Zhang, 2016; Lahoz and Schneider, 2014; van Leeuwen et al., 2019; Martin et al., 2015; Roth et al., 2017; Vetra-Carvalho et al., 2018).

Thanks to the scalability of parallel implementations, the use of ensemble algorithms (see e.g., Vetra-Carvalho et al., 2018; Houtekamer and Zhang, 2016; Bannister, 2017) have been proposed to estimate uncertainty and improve assimilation skills in

Kalman filters and variational methodologies. On the other hand, some of the strong points of the ensemble and variational have been merged 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 (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 sampling methods (as opposed to stochastic EnKF methods, see e.g. Carrassi et al., 2018). 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  $h^{\text{th}}$ -order approximation, the ensemble has the property of providing a forecast mean with no error as long as the evolution function used for forecasting (i.e., the model) is a polynomial of 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).

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 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 approximation is more effective the closer the ensemble members are to each other (i.e., small uncertainty), thus the higher the uncertainty the worse will be the approximation error in the mean computation. Since 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 methods and consequently larger computational costs, given that a higher order of approximation implies a larger number of ensemble members to represent the same uncertainty subspace<sup>1</sup>. For instance, second-order methods use an ensemble of  $r + 1$  members to span an  $r$ -dimensional 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 (see e.g., Wan and Van Der Merwe, 2000; Ambadan and Tang, 2009; Luo and Moroz, 2009, for examples of third-order method with  $2r + 1$  ensemble members).

The approximation order can also be increased by using properly weighted ensemble members. A weighted ensemble based on multi-dimensional Gauss-Hermite quadrature rule (Mastroianni and Milovanovic, 2008, cap. 2) 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

---

<sup>1</sup>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.

exploited in particle filter methods (van Leeuwen et al., 2019), where weights are used to evolve both the ensemble members  
55 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.

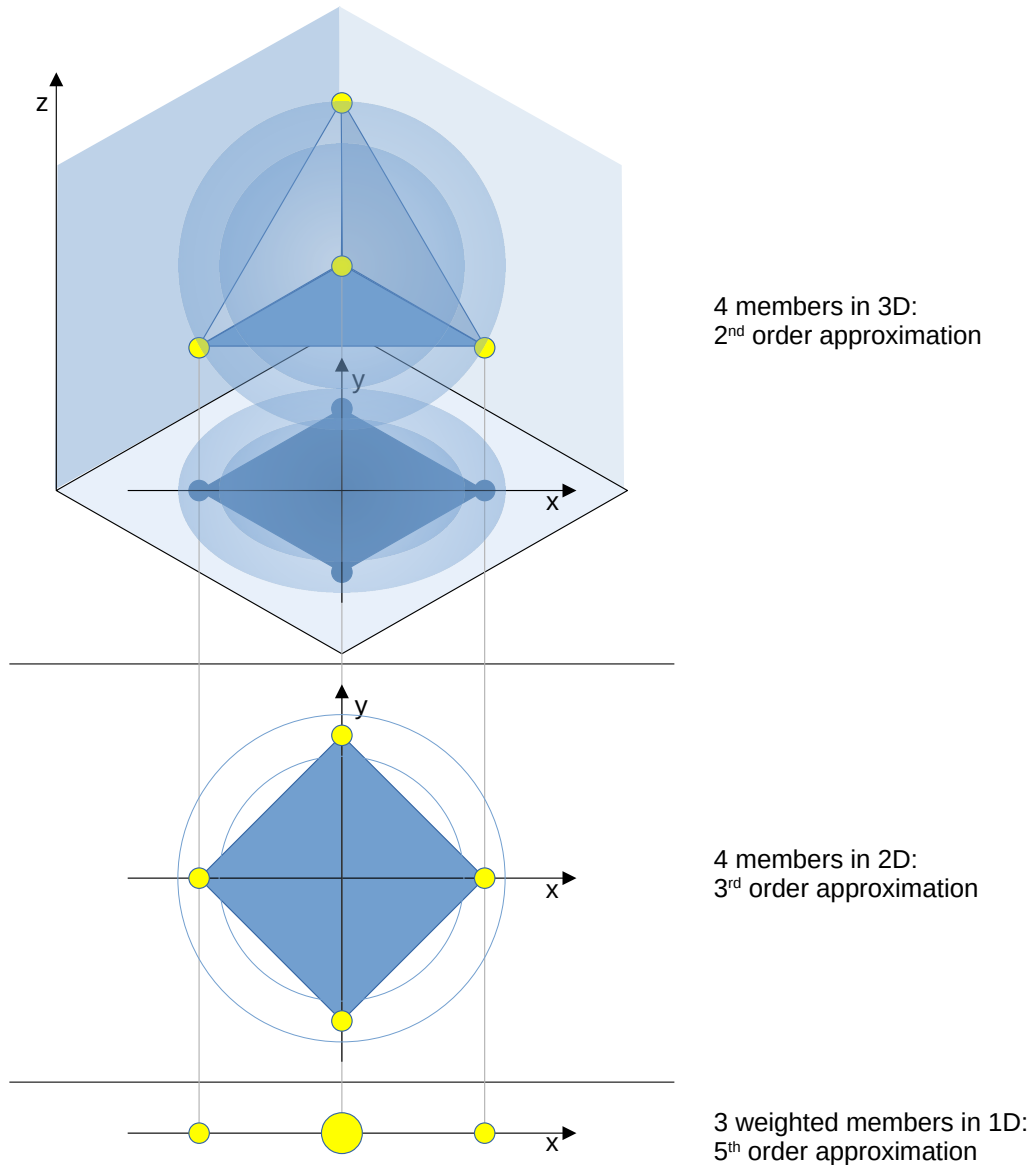
The proposed high-order filter (as its sampling method) exploits a Gauss-Hermite-like quadrature rule, along with a principal  
60 component analysis (PCA), and we named it Gauss-Hermite high-order sampling hybrid (GHOSH) filter. Members and related weights are chosen in such a way as to guarantee accuracy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

The reliability of the new filter is proven in a large set of twin experiments based on an idealized model commonly used to  
65 test DA methodologies (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), featuring assimilation of satellite observations.

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. The Lorenz96 numerical experiment is introduced in Section 4 and its results  
70 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 (see e.g., Carrassi et al., 2018), the novel strategy achieves a higher  
75 order of approximation without increasing the 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 principal components of the uncertainty (i.e., those directions where the ensemble spread is larger and the approximation is consequently worse). To better understand the GHOSH approach, Fig. 1 shows an uncertainty represented by a 3-dimensional standard normal distribution (spherical uncertainty isosurfaces are represented by shadowed areas). Using the second-order-exact sampling (Pham, 2001), 4 ensemble  
80 members (yellow points) are needed to represent this uncertainty probability density function. In a  $xyz$  Cartesian coordinate system, the 4 ensemble members are the vertices of a regular tetrahedron, and each possible second-order sampling corresponds to a rotation of this tetrahedron. With the appropriate rotation, the vertices draw (project) a square in the  $xy$  plane (Appendix C, Section C3.1), achieving a third-order approximation in the  $xy$  plane. Indeed, it can be proven that 4 square-shaped ensemble members provide a third-order approximation for a 2-dimensional normal distribution (Appendix C, Section C2.4). Similarly,  
85 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



**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 dimension (bottom) and by assigning proper weights, a fifth-order approximation is obtained.

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. In summary, Fig. 1 represents a weighted

ensemble with approximation order 2 in the  $xyz$  3-dimensional space, approximation order 3 in the  $xy$  plane and 5 along the  $x$  axis (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 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 larger on  $x$  than  $y$ , then 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-Hermite quadrature rule (Mastroianni and Milovanovic, 2008, cap. 2). Once the approximation order ( $h$  larger than 2) and the ensemble size ( $r + 1$ ) are fixed, the  $h^{\text{th}}$ -order weighted ensemble is computed for a relatively small  $s$ -dimensional subspace ( $s$  lower than  $r$ ); then, a new  $(r + 1)$ -sized second-order weighted ensemble is computed such that its projections along the uncertainty principal components coincide with the  $h^{\text{th}}$ -order ensemble. The resulting ensemble has a second-order approximation in the  $r$ -dimensional subspace and a  $h^{\text{th}}$ -order approximation in the  $s$ -dimensional subspace of the principal uncertainty 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 two phases of the high-order sampling (*initialization* and *sampling*) are described in the following from an algorithmic point of view. The *initialization* includes the steps that must be executed only once, while the *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 initialization:  $r$  is the dimension of the 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  $u_{m,j}$  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

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

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

the equation of system (1) for  $\xi = 0$ , i.e.,

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

ensures that the weights will sum to 1; the  $s$  equations for  $\xi = 1$ , i.e.,

$$120 \quad \left\{ \begin{array}{l} \sum_{m=1}^{r+1} u_{m,1} v_m = 0 \\ \vdots \\ \sum_{m=1}^{r+1} u_{m,s} v_m = 0, \end{array} \right.$$

ensure that the ensemble has 0-mean; the  $s^2$  equations for  $\xi = 2$ , i.e.,

$$\left\{ \begin{array}{l} \sum_{m=1}^{r+1} u_{m,1} u_{m,1} = \delta_{1,1} = 1 \\ \vdots \\ \sum_{m=1}^{r+1} u_{m,1} u_{m,s} = \delta_{1,s} = 0 \\ \vdots \\ \sum_{m=1}^{r+1} u_{m,s} u_{m,1} = \delta_{s,1} = 0 \\ \vdots \\ \sum_{m=1}^{r+1} u_{m,s} u_{m,s} = \delta_{s,s} = 1, \end{array} \right.$$

force the ensemble covariance matrix to be equal to the identity matrix; and the same holds for higher moments up to  $h$  ( $2 < \xi \leq h$ ). The general equation reported in system (1) represents all the moment-matching equations up to order  $h$ , where

125  $\mu_{j_1, \dots, j_\xi}$  are the statistical moments of a  $s$ -dimensional probability distribution that approximates the assumed 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 standard normal distribution  $\mathcal{N}(\mathbf{z}; 0, \mathbf{I}_s)$ , i.e.,

$$\mu_{j_1, \dots, j_\xi} = \int_{\mathbb{R}^s} z_{j_1} \cdots z_{j_\xi} \mathcal{N}(\mathbf{z}; 0, \mathbf{I}_s) d\mathbf{z}. \quad (2)$$

130 In the Gaussian case, the moments  $\mu_{j_1, \dots, j_\xi}$ , 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  $\mu_{j_1, \dots, j_\xi}$  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 larger than the number of variables, which implies that given  $r$  (and consequently the ensemble size) a larger approximation order  $h$  implies a smaller  $s$  (i.e., a smaller number of components approximated with order  $h$ ).

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

$$\mathbf{w} = (v_1^2, \dots, v_{r+1}^2). \quad (3)$$

## 2.2 Sampling

Given an  $N$ -dimensional state space, in the sampling phase a new ensemble is generated based on the ensemble mean  $\mathbf{x}_i \in \mathbb{R}^N$  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 uncertainty subspace spanned by the ensemble members.

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

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

where  $\mathbf{v}$  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  $r+1$ . A procedure for randomly building or completing orthogonal matrices is described in Appendix B.

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

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

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

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

where  $\mathbb{1}$  is a matrix (of the subscripted size) filled with ones,  $\mathbf{W} = \text{diag}(\mathbf{w})$  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

$$\mathbf{L}_i^T \mathbf{\Lambda}_i^{-1} \mathbf{L}_i = \mathbf{C}_i^T \mathbf{E}_i \mathbf{C}_i. \quad (7)$$

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  $\mathbf{\Lambda}_i$  being an appropriate  $N \times N$  matrix. The purpose of  $\mathbf{\Lambda}_i$  is to weight the

product between  $\mathbf{L}_i$  and its transposed; therefore,  $\mathbf{\Lambda}_i$  defines the criterion that decides the relative importance of different parts of the state vector, affecting the uncertainty principal components. In relatively simple applications,  $\mathbf{\Lambda}_i$  can be the identity matrix, but in most scenarios,  $\mathbf{\Lambda}_i$  needs to be properly designed (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.,

$$\mathbf{X}_i = (\mathbf{x}_i^1, \dots, \mathbf{x}_i^{(r+1)}). \quad (8)$$

### 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 pre-fixed times  $t_i$  in terms of the state vector and the covariance matrix representing the estimation uncertainty. Hence, at time  $t_i$ , a forecast is composed by the forecast state vector  $\mathbf{x}_i^f \in \mathbb{R}^N$  ( $N$  is the dimension of the state space) and by the forecast uncertainty covariance matrix  $\mathbf{P}_i^f$ . If an observation vector  $\mathbf{y}_i$  is available at time  $t_i$ , the information is assimilated in the analysis state vector  $\mathbf{x}_i^a$  with 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.,

$$\mathbf{X}_i^f = (\mathbf{x}_i^{f,1}, \dots, \mathbf{x}_i^{f,(r+1)}) \quad (9)$$

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

$$\mathbf{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_{r+1} \end{pmatrix}. \quad (10)$$

The state estimate is given by the weighted mean

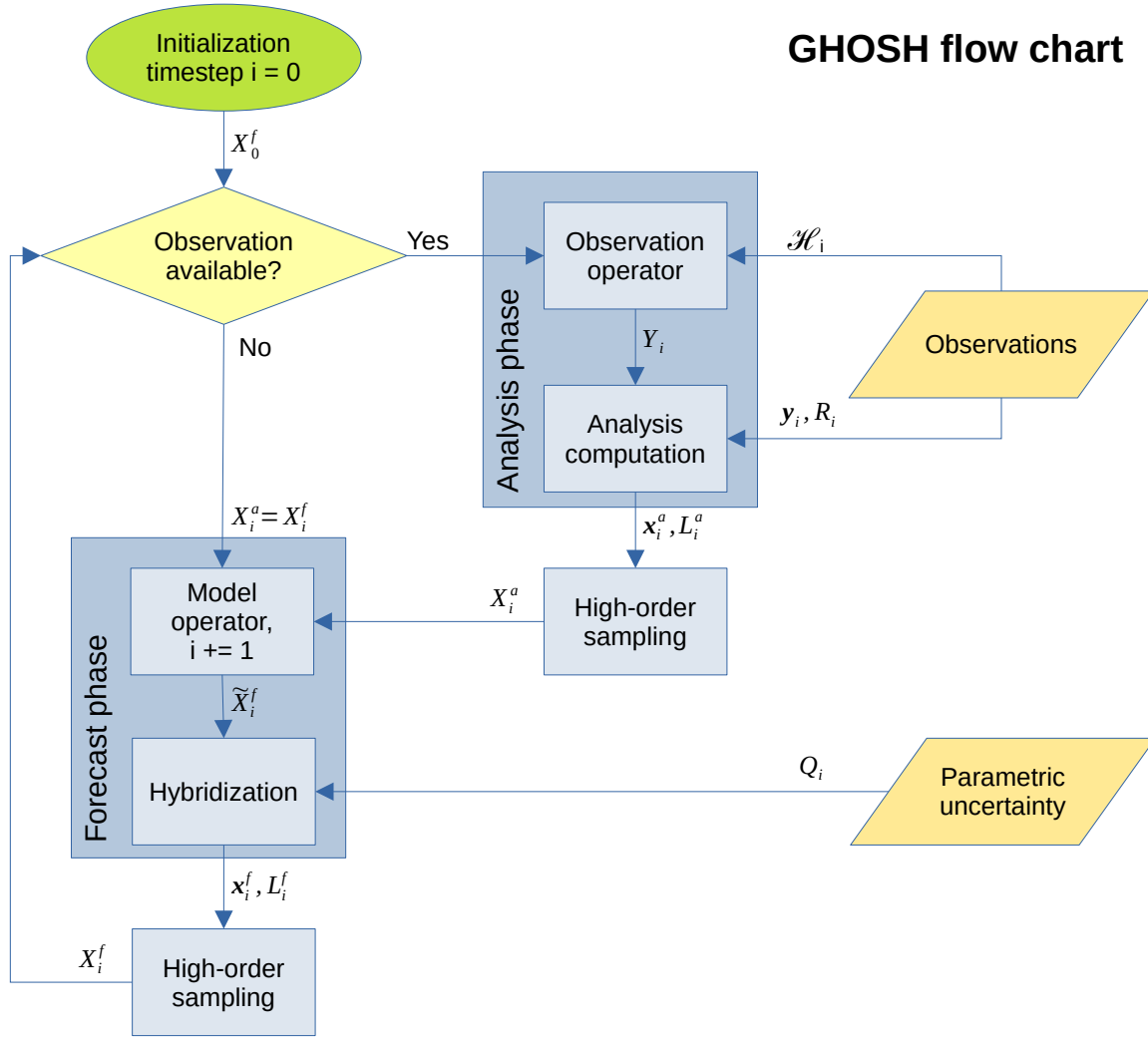
$$\mathbf{x}_i^f = \sum_{j=1}^{r+1} \mathbf{x}_i^{f,j} w_j = \mathbf{X}_i^f \mathbf{w}, \quad (11)$$

while the uncertainty covariance is approximated by the ensemble covariance matrix

$$\mathbf{P}_i^f \approx \mathbf{X}_i^f \mathbf{W} \mathbf{X}_i^{f,T} - \mathbf{x}_i^f \mathbf{x}_i^{f,T}, \quad (12)$$

with  $\mathbf{W} = \text{diag}(\mathbf{w})$  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 observations are available.



**Figure 2.** 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 parametric covariance matrix  $Q_i$ , which is built from the existing knowledge of the system, as done, for instance, for the background matrix in many variational schemes (Bannister, 2017, 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 uncertainty probability density function with better precision than in the other commonly used ensemble DA

algorithms, resulting in an order of approximation greater than 2 (see Appendix A) in the principal uncertainty modes. In this way the GHOSH filter takes advantage of the high approximation order of the sampling method twice: before applying the model operator and before applying the observation operator.

### 3.1 The global GHOSH filter

#### 195 3.1.1 Initialization

First, as the high-order sampling *initialization* phase (Section 2.1), the three GHOSH hyper-parameters  $r$  (the dimension of the 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  $\Omega^h$  and the weights vector  $w$ .

200 A starting weighted ensemble must be provided, with the coordinates of  $w$  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.,

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

205 with the subscripted index 0 representing the first time step.

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

$$\mathbf{x}_0^f = \mathbf{X}_0^f \mathbf{w}. \quad (14)$$

#### 3.1.2 Analysis phase

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

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

$$\mathbf{Y}_i = \left( \mathbf{y}_i^1, \dots, \mathbf{y}_i^{(r+1)} \right), \quad (16)$$

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

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

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

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

$$\mathbf{T} = \begin{pmatrix} & & \mathbf{I}_r \\ \hline 0 & \dots & 0 \end{pmatrix} - \mathbf{w} \mathbb{1}_{1 \times r}, \quad (18)$$

220 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), but other  $\mathbf{T}$ s can be explored without affecting the main structure of the algorithm (see e.g., Nerger et al., 2012). The analysis uncertainty covariance matrix  $\mathbf{P}_i^a$  is obtained by

$$\mathbf{P}_i^a = \tilde{\mathbf{L}}_i^a \mathbf{A}_i^a \tilde{\mathbf{L}}_i^{aT}, \quad (19)$$

225 where

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

and

$$\mathbf{Z}_i = \mathbf{Y}_i \mathbf{T}, \quad (21)$$

230  $\mathbf{W} = \text{diag}(\mathbf{w})$  being the  $(r+1) \times (r+1)$  diagonal matrix of weights and  $\mathbf{R}_i$  the  $n \times n$  covariance matrix representing the uncertainty in the observation  $y_i$ .

An opportune change in basis is used to obtain the simpler decomposition

$$\mathbf{P}_i^a = \mathbf{L}_i^a \mathbf{L}_i^{aT}, \quad (22)$$

where

$$\mathbf{L}_i^a = \tilde{\mathbf{L}}_i^a \mathbf{S}_i^a \quad (23)$$

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

$$\mathbf{S}_i^a \mathbf{S}_i^a = \mathbf{A}_i^a. \quad (24)$$

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

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

Equation (25) differs from the usual ensemble Kalman filter equations in the use of  $\mathbf{Y}_i \mathbf{w}$  in the last term instead of  $\mathcal{H}_i(\mathbf{x}_i^f)$ .  
 240 The two are the same if the observation operator  $\mathcal{H}_i$  is linear. However, in the general case,  $\mathbf{Y}_i \mathbf{w}$ , 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  $\mathbf{X}_i^a = (\mathbf{x}_i^{a,1}, \dots, \mathbf{x}_i^{a,(r+1)})$ , whose columns are the ensemble members, is obtained by

$$245 \quad \mathbf{X}_i^a = \mathbf{x}_i^a \mathbb{1}_{1 \times (r+1)} + \mathbf{L}_i^a \mathbf{C}_i^{aT} \boldsymbol{\Omega}_i^{aT} \mathbf{W}^{-\frac{1}{2}}. \quad (26)$$

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  $k \in \{1, \dots, r + 1\}$ ,

$$250 \quad \tilde{\mathbf{x}}_i^{f,k} = \mathcal{M}_i(\mathbf{x}_{i-1}^{a,k}), \quad (27)$$

$$\tilde{\mathbf{X}}_i^f = (\tilde{\mathbf{x}}_i^{f,1}, \dots, \tilde{\mathbf{x}}_i^{f,(r+1)}). \quad (28)$$

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

$$\mathbf{x}_i^f = \tilde{\mathbf{X}}_i^f \mathbf{w}. \quad (29)$$

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

$$\tilde{\mathbf{L}}_i^f = \tilde{\mathbf{X}}_i^f \mathbf{T}. \quad (30)$$

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

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

260 where  $\mathbf{Q}_i$  is the  $N \times N$  covariance matrix of 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 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

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

265 where  $\rho$  is the forgetting factor, which is added to the equation to introduce inflation (see e.g., Pham et al., 1998; Carrassi et al., 2018). The last term in equation (32) is one of the novel elements of the present work. It represents the projection of the parametric uncertainty matrix  $\mathbf{Q}_i$  on the uncertainty subspace defined by the ensemble. While the use of  $\mathbf{Q}_i$  in equation (31) follows Pham (2001), equation (32) projects  $\mathbf{Q}_i$  in a novel non-orthogonal way that is induced by the scalar product defined

by  $\mathbf{Q}_i^{-1}$ . This approach can be interpreted as a form of hybridization that aims at focusing on the effects of the parametric uncertainty in the ensemble uncertainty subspace.

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

Finally, equation (31) is conveniently rewritten as

$$\mathbf{P}_i^f \approx \mathbf{L}_i^f \mathbf{L}_i^{fT}, \quad (33)$$

275 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

$$\mathbf{L}_i^f = \tilde{\mathbf{L}}_i^f \mathbf{S}_i^f, \quad (34)$$

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

$$\mathbf{S}_i^f \mathbf{S}_i^f = \mathbf{A}_i^f, \quad (35)$$

280 which can be computed by an eigenvalue decomposition of  $\mathbf{A}_i^f$ .

### 3.1.5 Forecast high-order sampling

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

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

285 The sampling procedure is similar to the one of Section 3.1.3; therefore,  $\mathbf{C}_i^f$  and  $\mathbf{\Omega}_i^f$  are computed as  $\mathbf{C}_i$  and  $\mathbf{\Omega}_i$  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  $\mathbf{Q}_i$  effects in equations (31) and (32), which would be otherwise neglected. If  $\mathbf{Q}_i$  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

$$290 \quad \mathbf{x}_i^{f,k} = \mathbf{x}_i^f + \frac{1}{\sqrt{\rho}} (\tilde{\mathbf{x}}_i^{f,k} - \mathbf{x}_i^f), \quad (37)$$

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

### 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

295 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  $\mathcal{O}(r^2)$  steps.

Finally, the covariance matrices  $\mathbf{Q}_i^{-1}$  and  $\mathbf{R}_i^{-1}$  and the scaling matrix  $\mathbf{A}_i^{-1}$  are considered sparse or low rank, such that, for  
300 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  $\mathcal{O}(nr^2)$ ) and of its inverse ( $\mathcal{O}(r^3)$ ), and the change of basis of equation (23), which is a  $\mathcal{O}(Nr^2)$  operation. Excluding the cost of the observation operator  $\mathcal{H}_i$ , the analysis computational complexity sums to  $\mathcal{O}(nr^2 + r^3 + Nr^2)$ .

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

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

All together, the asymptotic computational complexity of the GHOSH filter is  $\mathcal{O}(nr^2 + r^3 + Nr^2)$ . 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  
310 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.

### 315 **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 (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  
320  $N$  rows and returns its localized counterpart with  $l$  rows with respect to the domain point  $p$ . The simplest 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 delocalization operator.

### 325 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

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

where  $\mathbf{Q}_i^p$  is the  $l \times l$  covariance matrix at the point  $p$  of the parametric uncertainty non-dependent on the ensemble.

330 The global basis is built by changing the basis and delocalizing, i.e.,

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

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); thus, the delocalization operator does not need to include averaging or other smoothing procedures to avoid discontinuities.

### 335 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.,

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

where  $\mathbf{R}_i^p$  is the  $l' \times l'$  observation uncertainty covariance matrix at the point  $p$ . This matrix can be extracted by  $\mathbf{R}_i^{-1}$ , but it is  
 340 convenient to increase the uncertainty at the points far from  $p$  (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.,

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

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  
 345 and Cohn, 1999).

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

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

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

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

$$350 \quad \mathbf{x}_i^a = \mathbf{x}_i^f + \mathcal{D} \left( \left\{ \mathcal{L}_p \left( \tilde{\mathbf{L}}_i^a \right) \mathbf{A}_i^{p,a} \mathcal{L}_p^{\mathcal{H}} (\mathbf{Z}_i)^T (\mathbf{R}_i^p)^{-1} \mathcal{L}_p^{\mathcal{H}} (\mathbf{y}_i - \mathbf{Y}_i \mathbf{w}) \right\}_p \right). \quad (43)$$

## 4 Experimental setup

The GHOSH filter has been tested in a very large set of twin experiments based on the Lorenz96 model (Lorenz, 1996) 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; Fertig et al., 2007; Gharamti, 2018; Grooms and Robinson, 2021; Nerger, 2022).

In each experiment, the *truth* trajectory is provided by integrating the model for a certain time interval. Then, at regular time intervals, observations of a subset of the system variables have been extracted 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), and one using the GHOSH filter with fifth approximation order (i.e.,  $h = 5$ , see Section 3). The SEIK and GHOSH assimilation experiments are initialised with 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 evaluated by computing, during the whole simulation, the root mean square error (RMSE) between the filter forecast and the *truth* before the assimilation step. The RMSEs on assimilated and non-assimilated variables have been evaluated for each filter. In order to produce reliable statistics, the same procedure has been repeated 90000 times varying the *truth* trajectory, the observations, the filters initial conditions and some filters hyper-parameters such as the ensemble size and the forgetting factor.

### 4.1 The Lorenz96 model

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

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

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 vector is  $N = 62$  and the forcing term is  $F = 8$ , which is a common value used to generate chaotic behaviour.

The equations have been implemented in python and 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  $\mathbf{x} = 0$ . The 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, 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*. In the same way, a longer block of 150 time units has been used as *truth* in a set of twin experiments aimed to test long-run performances.

## 4.2 Observations

Observations are generated for each *truth* trajectory by extracting from the state vector the values of the even indexed variables  
385 (i.e.,  $x_2, x_4, \dots, x_{62}$ ) every  $\Delta t$  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  $\Delta t \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$ .

## 4.3 Filters setup

The same settings are used in both SEIK and GHOSH filters. The ensemble size is chosen among 7, 15, 31 and 63 members.  
The initial conditions of the ensemble mean is chosen randomly around a *truth* initial condition, adding a random Gaussian  
390 noise with 0-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-exact sampling for SEIK, high-order sampling with  
 $h = 5$  for GHOSH), setting the prior covariance matrix accordingly to the PCA approximation of the climatological covariance  
(i.e.,  $r$  principal components are taken into account if the ensemble size is  $r + 1$ ).

Both filters use the same forgetting factor (equation (32)), chosen among 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95 and  
395 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. Hence, the values adopted for  $s$  are 2, 3, 4 and 5, respectively for  
7, 15, 31 and 63 ensemble members.

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

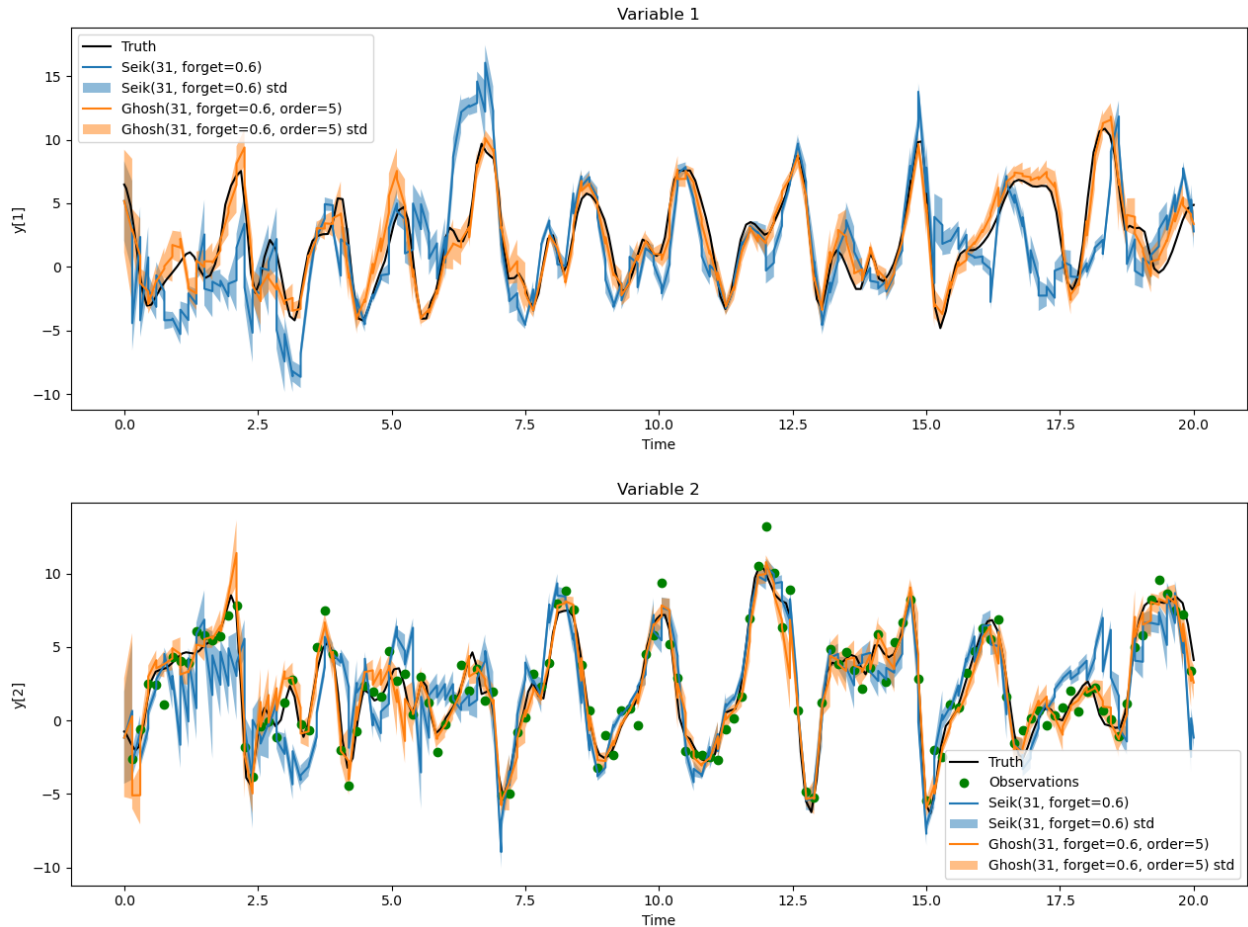
## 4.4 Experimental design

The twin experiments have been 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 20-time-units-long *truth* trajectories, 100 twin experiments have been launched for each  
405 of the 5 observation frequencies, 4 ensemble sizes and 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.

The RMSE score of the first phase experiments has been used 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 the 5 observation frequencies and 4 ensemble sizes (for a total  
410 of 2000 longer twin experiments), using for each filter its best forgetting factor in each configuration.

All the code was written in python and executed in a Linux computer equipped with an Intel(R) Core(TM) i7-11800H  
@2.30GHz.

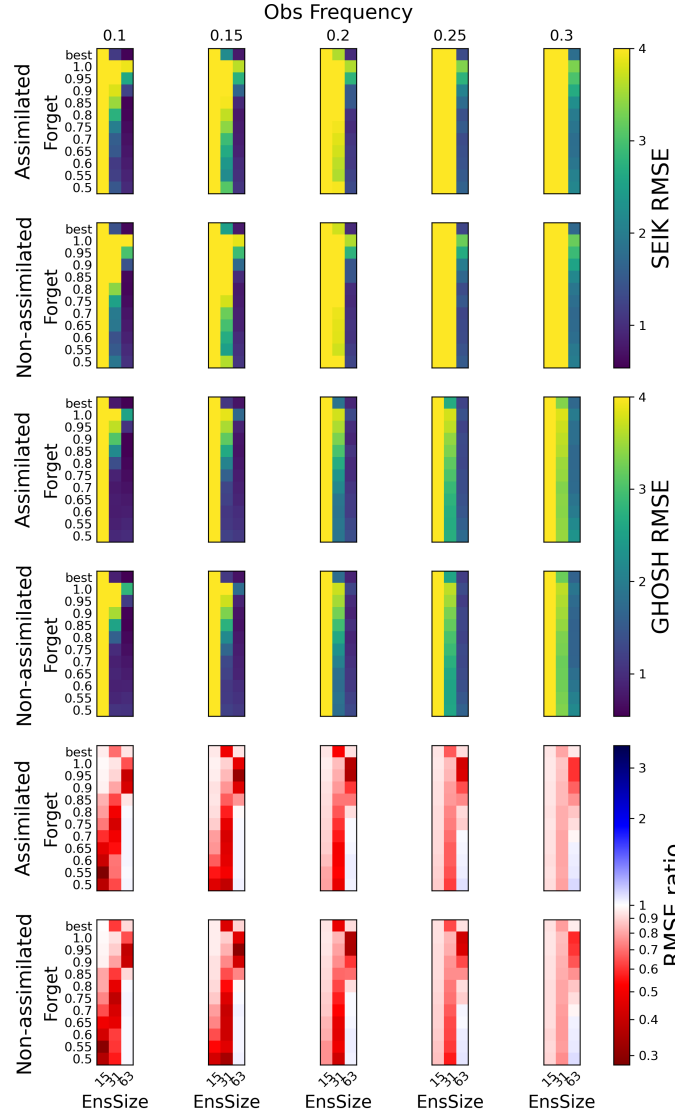


**Figure 3.** Results from one twin experiment example: time between observation (green dots) is 0.15, forgetting factor is 0.6, ensemble size is 31. Shady areas around blue (SEIK) and orange (GHOSH) lines represent the ensemble standard deviation, *truth* is the black line. Top panel: time evolution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

## 5 Results

### 5.1 20-time-units-long experiments

Figure 3 reports an example of evolution of the first two variables (one assimilated and one non-assimilated) of the Lorenz96 model for the two filters (GHOSH and SEIK). The selected simulation out of 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.



**Figure 4.** Result summary of 20-time-units-long twin experiments: each square in the colour-maps represents the aggregated results of 400 twin experiments, changing *truth*, observations and initial conditions. The results are summarized with 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 colour-maps has different observation frequency, with the numbers on the top indicating the time elapsed between each observation/assimilation. Each colour-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 88000 experiments under different conditions have been computed and  
420 reported in aggregated form in Fig. 4. 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.  
Aggregating 400 experiments together reduces the outcome variability by a factor 20 (i.e.,  $\sqrt{400}$ ) with respect to a single twin  
experiment, removing stochasticity concerns in the analysis of the twin experiment results.

The first two rows of colour-maps refer to the SEIK filter assimilated and non-assimilated variables. In each colour-map, the  
425 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 colour-map, labelled "*best*", represents the best result, in terms of lowest RMSE,  
obtained among the set of tested 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 rows of Fig. 4, while the last two rows show the  
430 ratio between GHOSH and SEIK RMSE values, with red color indicating RMSE reduction of GHOSH with respect to SEIK.

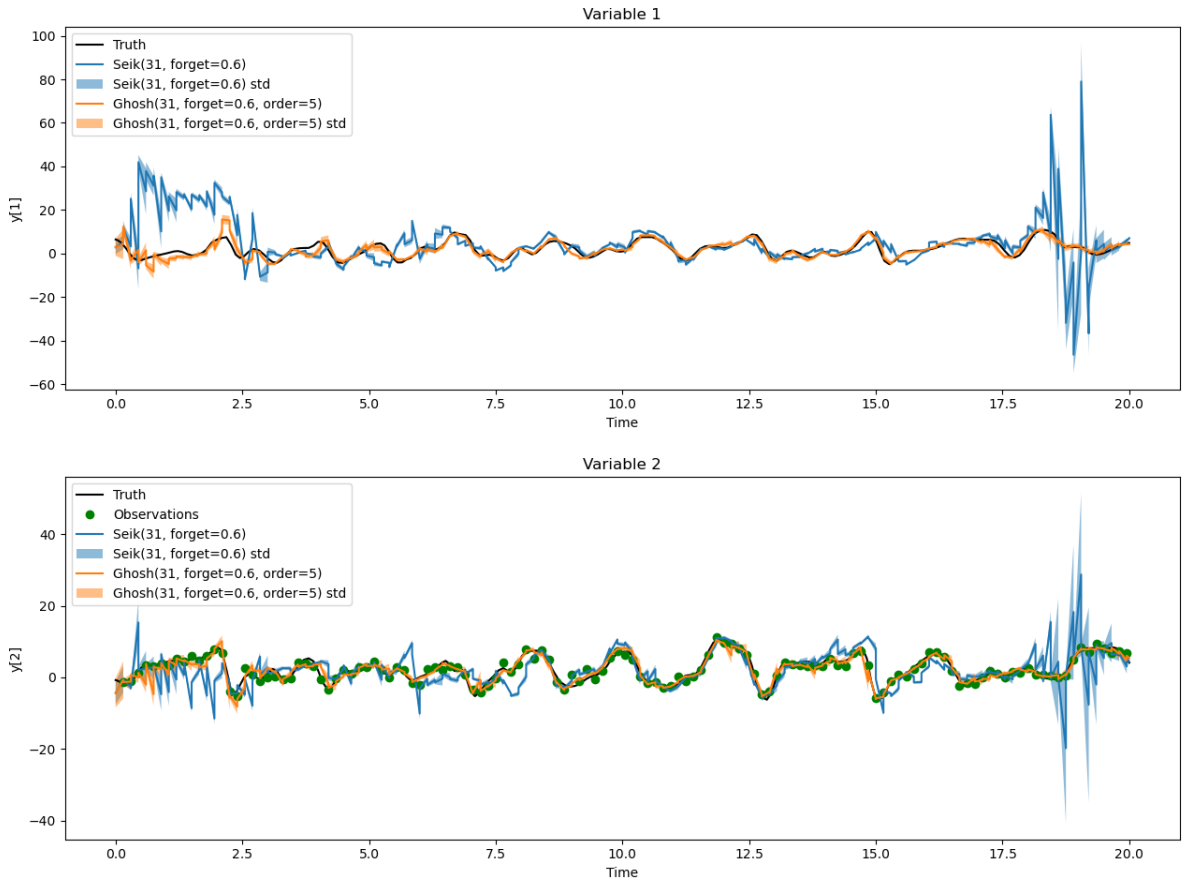
The column corresponding to the smallest ensemble size (i.e., 7 members) is not shown, since in this case SEIK and GHOSH  
behave very similarly, with very poor performances, due to lack of capability of describing the system complexity with an  
overly small ensemble size.

The yellow 15-members columns also show poor performances in all configurations but, looking at the last two rows of Fig.  
435 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  
440 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.

445 Looking at the last two rows of Fig. 4, the GHOSH filter outperforms the SEIK filter in most conditions, up to a 3 times RMSE  
reduction. Furthermore GHOSH proves to be always at least as good as SEIK. In the range of explored forgetting factors, the  
largest improvements (dark red, RMSE ratio 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 uncertainty subspace (i.e., the ensemble size is large).

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

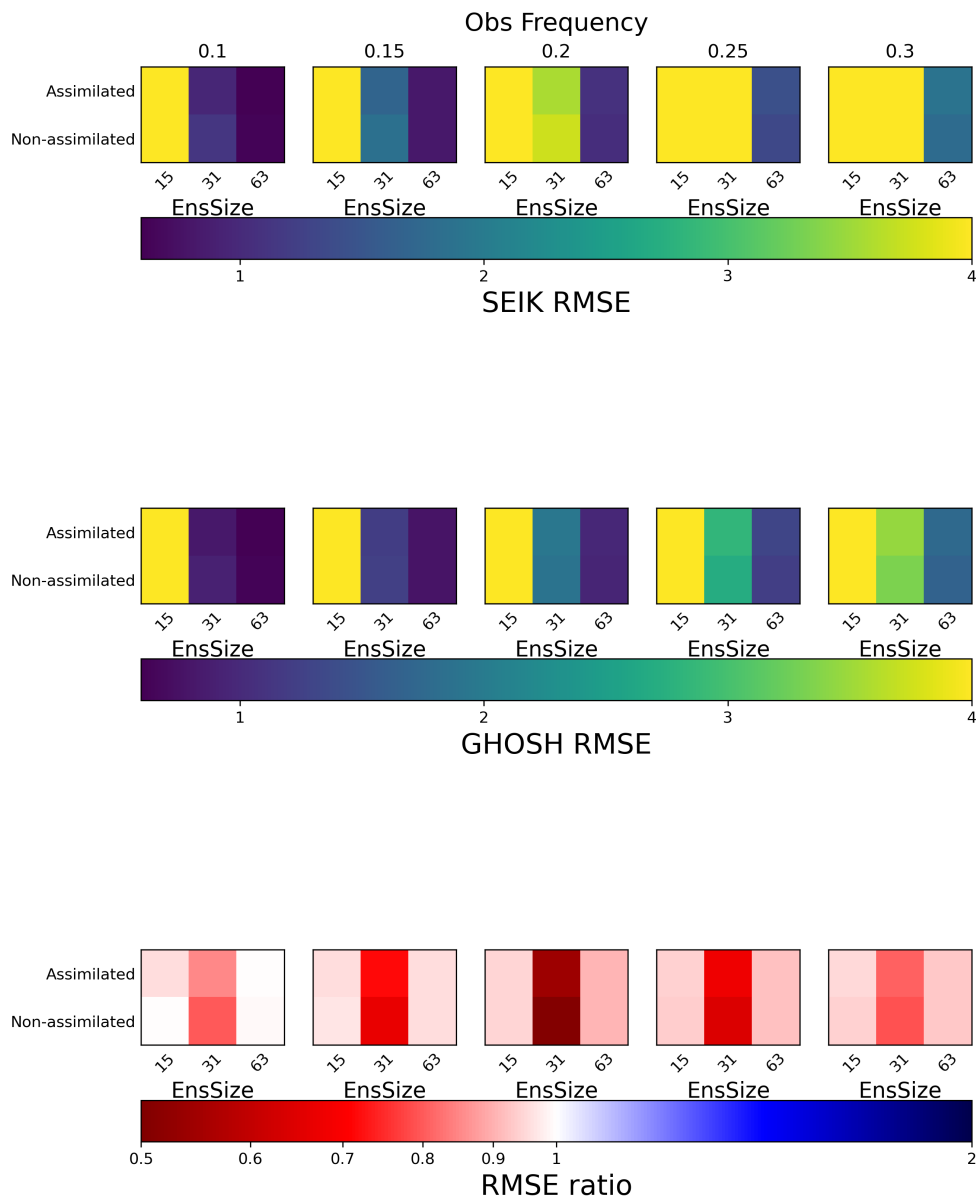


**Figure 5.** Example of a single twin experiment, with SEIK (blue line) diverging behaviour: time between observations (green dots) is 0.15, forgetting factor is 0.6, ensemble size is 31. Values over time of a non-assimilated (top) and an assimilated (bottom) variable.

The first four lines of Fig. 4 makes evident that the GHOSH filter, compared to SEIK, is less dependent on inflation, reaching  
455 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 inter-  
460 estingly, the RMSE ratio shows that the GHOSH advantage over SEIK is slightly larger in all the non-assimilated variables compared to the assimilated ones with the same settings (up to 0.08 RMSE ratio improvement in the *best* forgetting factor configuration, from 0.69 for assimilated to 0.61 for non-assimilated variables).



**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).

## 5.2 150-time-units-long experiments

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 experiments with different random observations and initial conditions. In the top panel, the 5 colour-maps correspond to different time intervals between observations, increasing from left to right. Each colour-map shows, for assimilated and non-assimilated 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, the 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).

The long-run experiments also confirm that the improvement is slightly larger for non-assimilated variables than for assimilated ones, scoring a better RMSE ratio (maximum ratio improvement of 0.05) in almost all configurations.

## 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 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, 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., 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 uncertainty covariance matrix is obtained by averaging the ensemble covariance

and a parametric covariance matrix non-depending by the ensemble, as described in, 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) compares order 1 algorithms (e.g., SEEK Pham et al., 1998) and order 2 algorithms (e.g., SEIK Pham, 2001). Results of Section 5 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 70% lower RMSE) a typical second-order method like SEIK and features higher stability and better performances.

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 uncertainty subspace spanned by the ensemble is big enough to be representative. If the number of ensemble members is too small, the assimilation simply fails independently from the 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). GHOSH filter showed a lower need for inflation with respect to SEIK, removing one of the instability causes observed in the twin experiment (Section 5). The lower need of inflation 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., 2020; Fertig et al., 2007; Gharamti, 2018; Grooms and Robinson, 2021; Grooms, 2022; Kirchgessner et al., 2014; Lei and Bickel, 2011; Nerger, 2022; Scheffler et al., 2022) 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 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.

In addition to a 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 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, resampled between the forecast and analysis phase (Fig. 2). In this way,  $Q_i$  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  $\mu$  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  $\Lambda_i$  used in the sampling procedure plays an important role in the filter, since  $\Lambda_i$  defines how variables are compared to each others in the computation of the most relevant components of the state uncertainty (Equation (7)). In the Lorenz96 case, it is straightforward to set  $\Lambda_i$  equal to the identity matrix, since each variable is equivalent to any other. In more complex applications,  $\Lambda_i$  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  $\Lambda_i$  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 constraints; 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 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 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 (e.g., van Leeuwen et al., 2019) and from other types of weighted ensemble filters, 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 when weights “collapse”, e.g., van Leeuwen

et al., 2019). In the case of GHOSH, the ensemble is used to estimate specific moments and then resampled to keep constant the weights, 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, 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, mainly depends on the ensemble size, 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 second-order data assimilation schemes, 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 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 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 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.

While increasing the order usually involves increasing the number of ensemble members (and consequently the computational cost), the GHOSH 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 parallel implementation of a Mediterranean physical-biogeochemical model.

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 nonlinearities 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 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.14230318>).

## 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  $h$ . Thus, they have the same mean after applying any polynomial operator of order  $h$ . In fact, given a probability density function (pdf)  $p : \mathbb{R}^s \rightarrow \mathbb{R}$ , if  $\mathcal{F} : \mathbb{R}^s \rightarrow \mathbb{R}$  is a polynomial operator such that

$$\mathcal{F}(\mathbf{x}) = \sum_{\xi=0}^h \sum_{j_1, \dots, j_\xi=1}^s f_\xi^{j_1, \dots, j_\xi} x_{j_1} \cdots x_{j_\xi}, \quad (\text{A1})$$

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

$$\bar{\mathcal{F}} = \int_{\mathbb{R}^s} \mathcal{F}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \sum_{\xi=0}^h \sum_{j_1=1}^s \cdots \sum_{j_\xi=1}^s f_\xi^{j_1, \dots, j_\xi} \int_{\mathbb{R}^s} x_{j_1} \cdots x_{j_\xi} p(\mathbf{x}) d\mathbf{x}, \quad (\text{A2})$$

depends only on the first  $h$  moments of the 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 (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 by the nonlinear system

$$\begin{cases} \vdots \\ \sum_{m=1}^{r+1} x_{j_1, m} \cdots x_{j_\xi, m} w_m = \mu_{j_1, \dots, j_\xi}, \\ \vdots \end{cases} \quad (\text{A3})$$

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

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

$$\mu_{j_1, \dots, j_\xi} = \int_{\mathbb{R}^s} x_{j_1} \cdots x_{j_\xi} p(\mathbf{x}) d\mathbf{x}. \quad (\text{A4})$$

The system is solved considering  $x_{j,m}$  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  $\tilde{\mathbf{X}} = (x_{j,m})$  (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  $\mathbf{X}$  with arbitrary mean  $\bar{\mathbf{x}}$  and covariance matrix  $\mathbf{L}\mathbf{L}^T$  through the projection

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

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.,  $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}, 0, \mathbf{I}_s)$ , the moments  $\mu_{j_1, \dots, j_\xi}$  are given by

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

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

$$\int_{\mathbb{R}^s} x_1^{k_1} \dots x_s^{k_s} \mathcal{N}(\mathbf{x}; 0, \mathbf{I}_s) d\mathbf{x} = \frac{k_1!}{\sqrt{2}^{k_1} (\frac{k_1}{2}!)} \dots \frac{k_s!}{\sqrt{2}^{k_s} (\frac{k_s}{2}!)} \quad (\text{A7})$$

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 (A3) the substitution

$$\begin{cases} w_m = v_m^2, \\ x_{j,m} = \frac{u_{m,j}}{v_m}, \end{cases} \quad (\text{A8})$$

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

$$\begin{cases} \vdots \\ \sum_{m=1}^{r+1} u_{m,j_1} \dots u_{m,j_\xi} v_m^{2-\xi} = \mu_{j_1, \dots, j_\xi}, \\ \vdots \end{cases} \quad (\text{A9})$$

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

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

where  $\mathbf{v} \in \mathbb{R}^{r+1}$  is the column vector with coordinates  $v_m$  and  $\mathbf{\Omega}^h$  is an  $(r+1) \times s$  matrix with coordinates  $u_{m,j}$ .

Note that equation (A10) can be solved for any unit vector  $\mathbf{v}$  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, 2001), which leads to, in the formalism of equation (A5),

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

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

640 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  $\mathbf{v}$  and  $\mathbf{\Omega}^h$ , given by a particular solution of system (A9). Then, exploiting symmetry and rotational invariance, randomness is added to avoid biases by multiplying  $\mathbf{\Omega}^h$  by any random  $s \times s$  orthogonal matrix  $\mathbf{\Omega}^{rnd}$ . Finally, the  $(r+1) \times (r-s)$  matrix  $\mathbf{\Omega}^\perp$  is chosen to fulfil

$$645 \left( \begin{array}{c|c|c} \mathbf{v} & \mathbf{\Omega}^h \mathbf{\Omega}^{rnd} & \mathbf{\Omega}^\perp \end{array} \right)^T \left( \begin{array}{c|c|c} \mathbf{v} & \mathbf{\Omega}^h \mathbf{\Omega}^{rnd} & \mathbf{\Omega}^\perp \end{array} \right) = \mathbf{I}_{r+1}. \quad (\text{A12})$$

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

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

$$\mathbf{\Omega} = \left( \begin{array}{c|c} & \mathbf{\Omega}^\perp \end{array} \right). \quad (\text{A13})$$

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

$$\mathbf{L}^T \mathbf{\Lambda}^{-1} \mathbf{L} = \mathbf{C}^T \mathbf{E} \mathbf{C}. \quad (\text{A14})$$

In the last equation, the right-hand side is an eigenvalue decomposition of the left-hand side, with  $\mathbf{E}$  being an  $r \times r$  diagonal  
655 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 2.2.

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 such that the principal component are on the left side of the matrix product  $\mathbf{L}\mathbf{C}^T$ . This is compliant with the ensemble stored in  $\mathbf{\Omega}^T \mathbf{W}^{-\frac{1}{2}}$ , and the sampling is finally obtained by

$$660 \mathbf{X} = \bar{\mathbf{x}} \mathbf{1}_{1 \times (r+1)} + \mathbf{L}\mathbf{C}^T \mathbf{\Omega}^h \mathbf{\Omega}^{rnd} \mathbf{W}^{-\frac{1}{2}}. \quad (\text{A15})$$

## Appendix B: Orthogonal matrices

The GHOSH sampling algorithm requires to randomly generate orthogonal matrices, as in the case of  $\mathbf{\Omega}^{rnd}$ , or to complete a set of orthonormal vectors to form an orthogonal matrix, as in the case of  $\mathbf{\Omega}^\perp$  (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  
665 have implemented in the provided python code.

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

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

is an orthogonal matrix.

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

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

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

675 (b) normalize  $\mathbf{v}$  to get  $\mathbf{u} = \frac{\mathbf{v}}{\sqrt{\mathbf{v}^T \mathbf{v}}}$

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

$$\mathbf{u}^\perp = \Omega^\perp (\Omega^\perp)^T \mathbf{u},$$

$$\mathbf{u}^\parallel = \mathbf{u} - \mathbf{u}^\perp;$$

3. normalize  $\mathbf{u}^\parallel$  to obtain  $\mathbf{w} = \frac{\mathbf{u}^\parallel}{\sqrt{(\mathbf{u}^\parallel)^T \mathbf{u}^\parallel}}.$

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

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

## Appendix C: Examples

685 The aim of this appendix is to help the reader understanding the concept of  $h^{\text{th}}$ -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  $\mathbf{P}_1, \mathbf{P}_2, \dots$ , 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 ( $r = 1$ )

690 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):  $\mu_{x^3} = 0$ ,
- 695 – the fourth moment (kurtosis):  $\mu_{x^4} = 3$ ,
- the fifth moment:  $\mu_{x^5} = 0, \dots$

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

$$P_1: x_1 = 1; \tag{C1}$$

$$P_2: x_2 = -1.$$

The ensemble first moment (mean) is:

$$700 \quad \frac{1}{2}(x_1 + x_2) = \frac{1}{2}(1 + (-1)) = 0. \tag{C2}$$

The ensemble second moment (variance) is:

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

Thus, the ensemble  $P_1, P_2$  matches the moments of the target pdf up to order 2.

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

705 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:

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

Thus, the ensemble  $P_1, P_2$  matches the moments of the target pdf up to order 3.

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

$$\begin{aligned} P_1: \quad x_1 &= \sqrt{3}, & w_1 &= \frac{1}{6}; \\ P_2: \quad x_2 &= -\sqrt{3}, & w_2 &= \frac{1}{6}; \\ P_3: \quad x_3 &= 0, & w_3 &= \frac{2}{3}. \end{aligned} \tag{C5}$$

The ensemble first moment (mean) is:

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

The ensemble second moment (variance) is:

$$715 \quad w_1x_1^2 + w_2x_2^2 + w_3x_3^2 = \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 3 + \frac{2}{3} \cdot 0 = 1. \quad (\text{C7})$$

The ensemble third moment (skewness) is:

$$w_1x_1^3 + w_2x_2^3 + w_3x_3^3 = \frac{1}{6} \cdot \sqrt{3}^3 + \frac{1}{6} \cdot (-\sqrt{3})^3 + \frac{2}{3} \cdot 0 = 0. \quad (\text{C8})$$

The ensemble fourth moment (kurtosis) is:

$$w_1x_1^4 + w_2x_2^4 + w_3x_3^4 = \frac{1}{6} \cdot 9 + \frac{1}{6} \cdot 9 + \frac{2}{3} \cdot 0 = 3. \quad (\text{C9})$$

720 The ensemble fifth moment is:

$$w_1x_1^5 + w_2x_2^5 + w_3x_3^5 = \frac{1}{6} \cdot \sqrt{3}^5 + \frac{1}{6} \cdot (-\sqrt{3})^5 + \frac{2}{3} \cdot 0 = 0. \quad (\text{C10})$$

Thus, the ensemble  $P_1, P_2, P_3$  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  $\mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I}_2)$ . According to equations (A6) and (A7), the

725 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:

730 – 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:  $\mu_{x^3} = 0$ ,

735 – 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:  $\mu_{x^4} = 3$ ,
  - 740 – 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:
  - 745 – 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:  $\mu_{xy^4} = 0$ ,
  - 750 – fifth moment associated to  $y^5$  is:  $\mu_{y^5} = 0$ , ...

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

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

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).

755 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:

$$760 \quad \frac{1}{3} (x_1^2 + x_2^2 + x_3^2) = \frac{1}{3} \left( \frac{3}{2} + \frac{3}{2} + 0 \right) = 1. \quad (\text{C14})$$

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

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

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

$$\frac{1}{3} (x_1 y_1 + x_2 y_2 + x_3 y_3) = \frac{1}{3} \left( \left( \sqrt{\frac{3}{2}} \cdot \left( -\frac{\sqrt{2}}{2} \right) + \left( -\sqrt{\frac{3}{2}} \right) \cdot \left( -\frac{\sqrt{2}}{2} \right) + 0 \cdot \sqrt{2} \right) \right) = 0. \quad (\text{C16})$$

765 Thus, the ensemble  $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$  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.

770 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:

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

775 Thus, the ensemble  $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$  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

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

780 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:

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

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

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

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

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

Thus, the ensemble  $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$  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

$$\begin{aligned} \mathbf{P}_1 : \quad (x_1, y_1) &= (\sqrt{2}, 0); \\ \mathbf{P}_2 : \quad (x_2, y_2) &= (-\sqrt{2}, 0); \\ 790 \quad \mathbf{P}_3 : \quad (x_3, y_3) &= (0, \sqrt{2}); \\ \mathbf{P}_4 : \quad (x_4, y_4) &= (0, -\sqrt{2}). \end{aligned} \quad (\text{C22})$$

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:

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

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

$$795 \quad \frac{1}{4} (y_1 + y_2 + y_3 + y_4) = \frac{1}{4} (0 + 0 + \sqrt{2} + (-\sqrt{2})) = 0. \quad (\text{C24})$$

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

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

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

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

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

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

The ensemble third moment associated to  $x^3$  is:

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

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

$$805 \quad \frac{1}{4} (x_1^2 y_1 + x_2^2 y_2 + x_3^2 y_3 + x_4^2 y_4) = \frac{1}{4} (2 \cdot 0 + 2 \cdot 0 + 0 \cdot \sqrt{2} + 0 \cdot (-\sqrt{2})) = 0. \quad (\text{C29})$$

The ensemble third moment associated to  $xy^2$  is:

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

The ensemble third moment associated to  $y^3$  is:

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

810 Thus, the ensemble  $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4$  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

$$\begin{aligned} \mathbf{P}_1 : \quad (x_1, y_1) &= (\sqrt{3}, 0), & w_1 &= \frac{1}{6}; \\ \mathbf{P}_2 : \quad (x_2, y_2) &= (-\sqrt{3}, 0), & w_2 &= \frac{1}{6}; \\ \mathbf{P}_3 : \quad (x_3, y_3) &= \left(0, \sqrt{\frac{3}{2}}\right), & w_3 &= \frac{1}{3}; \\ \mathbf{P}_4 : \quad (x_4, y_4) &= \left(0, -\sqrt{\frac{3}{2}}\right), & w_4 &= \frac{1}{3}. \end{aligned} \quad (\text{C32})$$

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  $\mathbf{P}_3$  and  $\mathbf{P}_4$  collapsing in the same point on the  $x$ -axis. Differently from ensemble (C18),

815 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:

$$820 \quad w_1 y_1 + w_2 y_2 + w_3 y_3 + w_4 y_4 = \frac{1}{6} \cdot 0 + \frac{1}{6} \cdot 0 + \frac{1}{3} \cdot \sqrt{\frac{3}{2}} + \frac{1}{3} \cdot \left(-\sqrt{\frac{3}{2}}\right) = 0. \quad (\text{C33})$$

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

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

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

$$w_1 x_1 y_1 + w_2 x_2 y_2 + w_3 x_3 y_3 + w_4 x_4 y_4 = \frac{1}{6} \cdot \sqrt{3} \cdot 0 + \frac{1}{6} \cdot (-\sqrt{3}) \cdot 0 + \frac{1}{3} \cdot 0 \cdot \sqrt{\frac{3}{2}} + \frac{1}{3} \cdot 0 \cdot \left(-\sqrt{\frac{3}{2}}\right) = 0. \quad (\text{C35})$$

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

$$w_1x_1^2y_1 + w_2x_2^2y_2 + w_3x_3^2y_3 + w_4x_4^2y_4 = \frac{1}{6} \cdot 3 \cdot 0 + \frac{1}{6} \cdot 3 \cdot 0 + \frac{1}{3} \cdot 0 \cdot \sqrt{\frac{3}{2}} + \frac{1}{3} \cdot 0 \cdot \left(-\sqrt{\frac{3}{2}}\right) = 0. \quad (\text{C36})$$

The ensemble third moment associated to  $xy^2$  is:

$$w_1x_1y_1^2 + w_2x_2y_2^2 + w_3x_3y_3^2 + w_4x_4y_4^2 = \frac{1}{6} \cdot \sqrt{3} \cdot 0 + \frac{1}{6} \cdot (-\sqrt{3}) \cdot 0 + \frac{1}{3} \cdot 0 \cdot \frac{3}{2} + \frac{1}{3} \cdot 0 \cdot \frac{3}{2} = 0. \quad (\text{C37})$$

The ensemble third moment associated to  $y^3$  is:

$$830 \quad w_1y_1^3 + w_2y_2^3 + w_3y_3^3 + w_4y_4^3 = \frac{1}{6} \cdot 0 + \frac{1}{6} \cdot 0 + \frac{1}{3} \cdot \left(\sqrt{\frac{3}{2}}\right)^3 + \frac{1}{3} \cdot \left(-\sqrt{\frac{3}{2}}\right)^3 = 0. \quad (\text{C38})$$

Thus, the ensemble  $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4$  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

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

835 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:

$$w_1x_1 + w_2x_2 + \dots + w_7x_7 = \frac{1}{12} \cdot \sqrt{3} + \frac{1}{12} \cdot \sqrt{3} + \frac{1}{12} \cdot (-\sqrt{3}) + \frac{1}{12} \cdot (-\sqrt{3}) + \frac{1}{12} \cdot 0 + \frac{1}{12} \cdot 0 + \frac{1}{2} \cdot 0 = 0. \quad (\text{C40})$$

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

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

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

$$w_1x_1^2 + w_2x_2^2 + \dots + w_7x_7^2 = \frac{1}{12} \cdot 3 + \frac{1}{12} \cdot 3 + \frac{1}{12} \cdot 3 + \frac{1}{12} \cdot 3 + \frac{1}{12} \cdot 0 + \frac{1}{12} \cdot 0 + \frac{1}{2} \cdot 0 = 1. \quad (\text{C42})$$

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

$$845 \quad w_1 y_1^2 + w_2 y_2^2 + \dots + w_7 y_7^2 = \frac{1}{12} \cdot 1 + \frac{1}{12} \cdot 1 + \frac{1}{12} \cdot 1 + \frac{1}{12} \cdot 1 + \frac{1}{12} \cdot 4 + \frac{1}{12} \cdot 4 + \frac{1}{2} \cdot 0 = 1. \quad (\text{C43})$$

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

$$\begin{aligned} w_1 x_1 y_1 + w_2 x_2 y_2 + \dots + w_7 x_7 y_7 = \\ = \frac{1}{12} \cdot \sqrt{3} \cdot 1 + \frac{1}{12} \cdot \sqrt{3} \cdot (-1) + \frac{1}{12} \cdot (-\sqrt{3}) \cdot 1 + \frac{1}{12} \cdot (-\sqrt{3}) \cdot (-1) + \frac{1}{12} \cdot 0 \cdot 2 + \frac{1}{12} \cdot 0 \cdot (-2) + \frac{1}{2} \cdot 0 \cdot 0 = 0. \end{aligned} \quad (\text{C44})$$

Given the comparably high number of moments to be matched, it could be useful here to introduce a result that relieves from  
850 checking some moments. In fact, if a zero-mean ensemble is symmetric (i.e., if  $\mathbf{P}$  is a non-zero ensemble member then also  $-\mathbf{P}$  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.

855 The ensemble fourth moment associated to  $x^4$  is:

$$w_1 x_1^4 + w_2 x_2^4 + \dots + w_7 x_7^4 = \frac{1}{12} \cdot 9 + \frac{1}{12} \cdot 9 + \frac{1}{12} \cdot 9 + \frac{1}{12} \cdot 9 + \frac{1}{12} \cdot 0 + \frac{1}{12} \cdot 0 + \frac{1}{2} \cdot 0 = 3. \quad (\text{C45})$$

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

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

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

$$w_1 x_1^2 y_1^2 + w_2 x_2^2 y_2^2 + \dots + w_7 x_7^2 y_7^2 = \frac{1}{12} \cdot 3 \cdot 1 + \frac{1}{12} \cdot 3 \cdot 1 + \frac{1}{12} \cdot 3 \cdot 1 + \frac{1}{12} \cdot 3 \cdot 1 + \frac{1}{12} \cdot 0 \cdot 4 + \frac{1}{12} \cdot 0 \cdot 4 + \frac{1}{2} \cdot 0 \cdot 0 = 1. \quad (\text{C47})$$

The ensemble fourth moment associated to  $xy^3$  is:

$$\begin{aligned} w_1 x_1 y_1^3 + w_2 x_2 y_2^3 + \dots + w_7 x_7 y_7^3 = \\ = \frac{1}{12} \cdot \sqrt{3} \cdot 1 + \frac{1}{12} \cdot \sqrt{3} \cdot (-1) + \frac{1}{12} \cdot (-\sqrt{3}) \cdot 1 + \frac{1}{12} \cdot (-\sqrt{3}) \cdot (-1) + \frac{1}{12} \cdot 0 \cdot 8 + \frac{1}{12} \cdot 0 \cdot (-8) + \frac{1}{2} \cdot 0 \cdot 0 = 0. \end{aligned} \quad (\text{C48})$$

865 The ensemble fourth moment associated to  $y^4$  is:

$$w_1 y_1^4 + w_2 y_2^4 + \dots + w_7 y_7^4 = \frac{1}{12} \cdot 1 + \frac{1}{12} \cdot 1 + \frac{1}{12} \cdot 1 + \frac{1}{12} \cdot 1 + \frac{1}{12} \cdot 16 + \frac{1}{12} \cdot 16 + \frac{1}{2} \cdot 0 = 3. \quad (\text{C49})$$

Thus, the ensemble  $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4$  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  $\mathcal{N}(x; \mathbf{0}, \mathbf{I}_3)$ . According to equations (A6) and (A7), the  
870 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:  $\mu_z = 0$ ,

875 – 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:  $\mu_{xy} = 0$ ,
- 880 – 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

$$\begin{aligned}
 P_1 : \quad (x_1, y_1, z_1) &= \left( \sqrt{2}, 0, -\frac{1}{2} \right); \\
 P_2 : \quad (x_2, y_2, z_2) &= \left( -\sqrt{2}, 0, -\frac{1}{2} \right); \\
 P_3 : \quad (x_3, y_3, z_3) &= \left( 0, \sqrt{2}, \frac{1}{2} \right); \\
 P_4 : \quad (x_4, y_4, z_4) &= \left( 0, -\sqrt{2}, \frac{1}{2} \right).
 \end{aligned} \tag{C50}$$

This tetrahedron-shaped ensemble is a second-order ensemble. It is one of the many possible outcomes of the SEIK's sampling  
885 (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:

$$890 \quad \frac{1}{4} (z_1 + z_2 + z_3 + z_4) = \frac{1}{4} \left( \left( -\frac{1}{2} \right) + \left( -\frac{1}{2} \right) + \frac{1}{2} + \frac{1}{2} \right) = 0. \tag{C51}$$

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

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

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

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

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

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

Thus, the ensemble  $P_1, P_2, P_3, P_4$  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.

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

$$\begin{aligned} P_1 : \quad (x_1, y_1, z_1) &= (\sqrt{3}, 0, -\sqrt{2}), \quad w_1 = \frac{1}{6}; \\ P_2 : \quad (x_2, y_2, z_2) &= (-\sqrt{3}, 0, -\sqrt{2}), \quad w_2 = \frac{1}{6}; \\ P_3 : \quad (x_3, y_3, z_3) &= \left(0, \sqrt{\frac{3}{2}}, \frac{\sqrt{2}}{2}\right), \quad w_3 = \frac{1}{3}; \\ P_4 : \quad (x_4, y_4, z_4) &= \left(0, -\sqrt{\frac{3}{2}}, \frac{\sqrt{2}}{2}\right), \quad w_4 = \frac{1}{3}. \end{aligned} \quad (\text{C55})$$

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  $z$ .

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

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

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

$$w_1z_1^2 + w_2z_2^2 + w_3z_3^2 + w_4z_4^2 = \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 2 + \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2} = 1. \quad (\text{C57})$$

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

$$910 \quad w_1x_1z_1 + w_2x_2z_2 + w_3x_3z_3 + w_4x_4z_4 = \frac{1}{6} \cdot \sqrt{3} \cdot (-\sqrt{2}) + \frac{1}{6} \cdot (-\sqrt{3}) \cdot (-\sqrt{2}) + \frac{1}{3} \cdot 0 \cdot \frac{\sqrt{2}}{2} + \frac{1}{3} \cdot 0 \cdot \frac{\sqrt{2}}{2} = 0. \quad (\text{C58})$$

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

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

Thus, the ensemble  $P_1, P_2, P_3, P_4$  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.

915 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

$$\begin{aligned}
P_1: & (x_{1,1}, x_{2,1}, \dots, x_{r,1}) = (\sqrt{r}, 0, 0, \dots, 0); \\
P_{-1}: & (x_{1,-1}, x_{2,-1}, \dots, x_{r,-1}) = (-\sqrt{r}, 0, 0, \dots, 0); \\
P_2: & (x_{1,2}, x_{2,2}, \dots, x_{r,2}) = (0, \sqrt{r}, 0, \dots, 0); \\
P_{-2}: & (x_{1,-2}, x_{2,-2}, \dots, x_{r,-2}) = (0, -\sqrt{r}, 0, \dots, 0); \\
& \vdots \\
P_r: & (x_{1,r}, x_{2,r}, \dots, x_{r,r}) = (0, 0, \dots, 0, \sqrt{r}); \\
P_{-r}: & (x_{1,-r}, x_{2,-r}, \dots, x_{r,-r}) = (0, 0, \dots, 0, -\sqrt{r}).
\end{aligned} \tag{C60}$$

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,

920 it only remains to prove that variances are equal to 1.

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

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

Thus, the ensemble  $P_1, P_{-1}, P_2, P_{-2}, \dots, P_r, P_{-r}$  matches the moments of the target pdf (i.e., the  $r$ -dimensional standard normal distribution  $\mathcal{N}(x; 0, \mathbf{I}_r)$ ) up to order 3.

925 *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.

930 *Acknowledgements.* We acknowledge the CINECA award under the ISCRA initiative and the HPC-TRES program, for the availability of 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

- 935 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.
- 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>, 2017.
- Bannister, R. N.: A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error  
940 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.
- 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](https://doi.org/https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2), 2001.
- 945 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.
- 950 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.
- 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.
- 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  
955 Lorenz-96, *Tellus A: Dynamic Meteorology and Oceanography*, <https://doi.org/10.1111/j.1600-0870.2006.00205.x>, 2007.
- 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.
- 960 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 –  
965 2919, [https://doi.org/10.1175/1520-0493\(2000\)128<2905:AHEKFV>2.0.CO;2](https://doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2), 2000.
- 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.

- 970 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.
- 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.
- 975 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.
- 980 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.
- 985 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.
- 990 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.
- 995 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.
- 1000 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.: 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](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.
- 1005 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.
- 1010 Roth, M., Hendeb, 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.
- 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.
- 1015 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.
- 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.
- 1020 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.
- 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](https://doi.org/10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2), 2003.
- 1025 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.
- 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.
- 1030 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.
- 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.
- 1035

# A novel Gauss-Hermite High-Order Sampling Hybrid ensemble filter for computationally efficient data assimilation in geosciences - Part 2: OGSTM-BFM-GHOSH

Simone Spada<sup>1</sup>, Anna Teruzzi<sup>1</sup>, Stefano Maset<sup>2</sup>, Stefano Salon<sup>1</sup>, Cosimo Solidoro<sup>1</sup>, and Gianpiero Cossarini<sup>1</sup>

<sup>1</sup>National Institute of Oceanography and Applied Geophysics - OGS, 34010 Trieste, Italy

<sup>2</sup>University of Trieste, 34127 Trieste, Italy

**Correspondence:** Simone Spada (sspada@ogs.it)

**Abstract.** Data assimilation (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, aims to improve DA performances by featuring an approximation order higher than in other ensemble DA filters like SEIK (Singular Evolutive Interpolated Kalman filter) or ETKF (Ensemble Transform Kalman Filter).

5 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 (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  
10 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 that: i) the GHOSH implementation in a realistic three-dimensional application is just as computationally feasible as other ensemble assimilation filters, since the integration of a  
15 realistic model is by far more computationally expensive than the assimilation scheme; ii) the GHOSH assimilation algorithm improved the agreement between forecasts and observations without producing unrealistic effects on the non-assimilated variables. Moreover, a sensitivity analysis revealed correlation between higher order and a key nutrient (nitrate) estimation improvement, highlighting the importance of the GHOSH enhanced order of approximation in boosting DA performances on non-assimilated variables.

## 20 1 Introduction

The Gauss-Hermite 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.

25 In Part 1, the reliability of the new filter has been proven in a large set of twin experiments based on an idealized model commonly used to test DA methodologies (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 the Copernicus Marine Service (Salon et al., 2019), featuring assimilation of satellite observations.

Biogeochemical data assimilation (DA) is relatively recent both in terms of methods and assimilated variables (Fennel et al., 30 2019). Surface chlorophyll concentration, which is estimated by satellite ocean colour observations, is the most commonly assimilated 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., 2016). The 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 35 (see e.g., Santana-Falcón et al., 2020; Ford, 2020; Fowler et al., 2023; Mamnun et al., 2022; Wang et al., 2020).

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.

In this work, the GHOSH skill is assessed qualitatively, looking at the dynamics of chlorophyll, nutrients and their correlations, and quantitatively, in terms of improved agreement with observations without degradation of non-assimilated variables. 40 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, the feasibility of a real implementation of the 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.

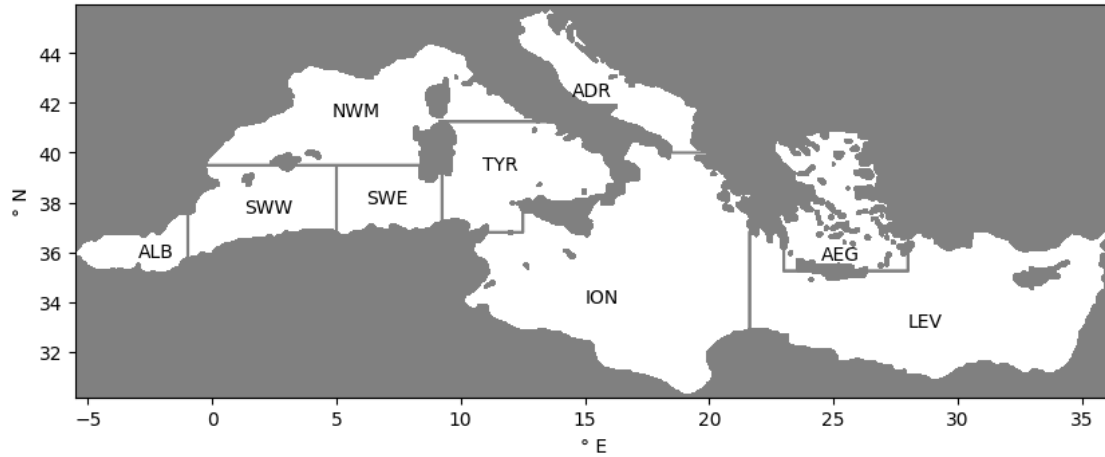
Section 2 presents the experimental design and introduces the OGSTM-BFM model, the data used for assimilation and 45 performance evaluation, and a summary of the relevant GHOSH key elements. Section 3 shows the results of the experiments, that are discussed in Section 4. Finally, Appendix A contains some specific details of the GHOSH implementation.

## 2 Material and methods

### 2.1 The OGSTM-BFM model

The biogeochemical model used in this study is the coupled transport-biogeochemical model OGSTM-BFM, which is the 50 biogeochemical component of the Mediterranean Copernicus Marine Service (Salon et al., 2019). The OGSTM transport model 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 (for a total of 51 tracers as state variables). The OGSTM-BFM 55 model has been applied to study the chlorophyll and primary production (Cossarini et al., 2021; Teruzzi et al., 2021, 2018b;



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

Lazzari et al., 2012), 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).

## 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). 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 datasets for nitrate and phosphate, to be used for validation, are computed from a dataset that integrates the in situ EMODnet data collections (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 represented in Fig. 1 (Salon et al., 2019) and for 4 different seasons (January-March; April-June; July-September; October-December).

### 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., the subspace of the state space spanned by the ensemble members. The GHOSH filter features an enhanced order of approximation (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  $\mu$ ) the prior uncertainty pdf (i.e., 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  $\Lambda$  is used to rescale the state variables, i.e. to define which variables are of more interest. In this implementation,  $\mu$  describes (up to order  $h$ ) a Gaussian distribution, and the scaling matrix  $\Lambda$  is diagonal. Being  $\mathbf{P} = \mathbf{L}\mathbf{L}^T$  the uncertainty covariance matrix,  $\Lambda$  contains the uncertainty variance (i.e., the diagonal of  $\mathbf{P}$ ), which is computed as the squared Euclidean norms of the rows of  $\mathbf{L}$ . In the present application, the  $\Lambda$  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 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  $\frac{1}{\rho}$  ( $\rho$  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).

### 2.4 Experimental design

Implemented in the OGSTM-BFM biogeochemical model system of the Mediterranean Sea at  $\frac{1}{4}^\circ$  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. 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 is 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 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 control run initial condition and the starting ensemble 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 applying a  $10^{-4}$  cutoff on concentrations, for a total of 310 simulated multivariate states. The mean of the multivariate states was used as control initial condition and mean of the starting ensemble, while a PCA of the ensemble of  $\mathbf{\Lambda}$ -scaled states provided the 15 (i.e., the dimension of the uncertainty subspace  $r$ ) 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 covariance matrix of the parametric uncertainty  $\mathbf{Q}$  has been assumed diagonal and invariant in time, 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  $\mathbf{Q}$  uncertainty variance in the coastal areas was reduced by applying a 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.

Also the covariance matrix  $\mathbf{R}$  of the observation uncertainty has been assumed diagonal and constant in time. Using the same fourth-degree polynomial smoothing applied in the  $\mathbf{Q}$  case, the inverse of the observation uncertainty variances is smoothed to zero in the coastal areas (i.e., the  $\mathbf{R}$  variances increase in the coastal areas oppositely to  $\mathbf{Q}$ ). Furthermore, the fourth-degree polynomial smoothing was used to weight the local version of  $\mathbf{R}^{-1}$  to gently remove assimilation of observations beyond the localization radius of 90 km.

In the assimilation experiment, the GHOSH 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 parametric uncertainty (represented in  $\mathbf{Q}$ ) and observation uncertainty ( $\mathbf{R}$ ) may be necessary to produce an effective assimilation when system uncertainties are not fully known (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 degradation on biogeochemical variables.

As a sensitivity analysis, 17 additional tests (summarized in tab. 1) are carried out varying the polynomial order  $h$ , the forgetting factor  $\rho$  and the forecast frequency. In particular, the polynomial order  $h$  (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 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 GHOSH filter settings.

| Test name                             | Order ( $h$ ) | Forgetting factor ( $\rho$ ) | Forecast frequency |
|---------------------------------------|---------------|------------------------------|--------------------|
| $ctrl$                                |               | Free model                   |                    |
| $T_{h2\_rho1\_d}$                     | 2             | 1                            | daily              |
| $T_{h3\_rho1\_d}$                     | 3             |                              |                    |
| $T_{h5\_rho1\_d}$                     | 5             |                              |                    |
| $T_{h2\_rho0.9\_d}$                   | 2             | 0.9                          |                    |
| $T_{h3\_rho0.9\_d}$                   | 3             |                              |                    |
| <b><math>T_{h5\_rho0.9\_d}</math></b> | <b>5</b>      |                              |                    |
| $T_{h2\_rho0.8\_d}$                   | 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.

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).

### 140 3 Results

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 non-assimilated variables (nutrients, Section 3.2), and the

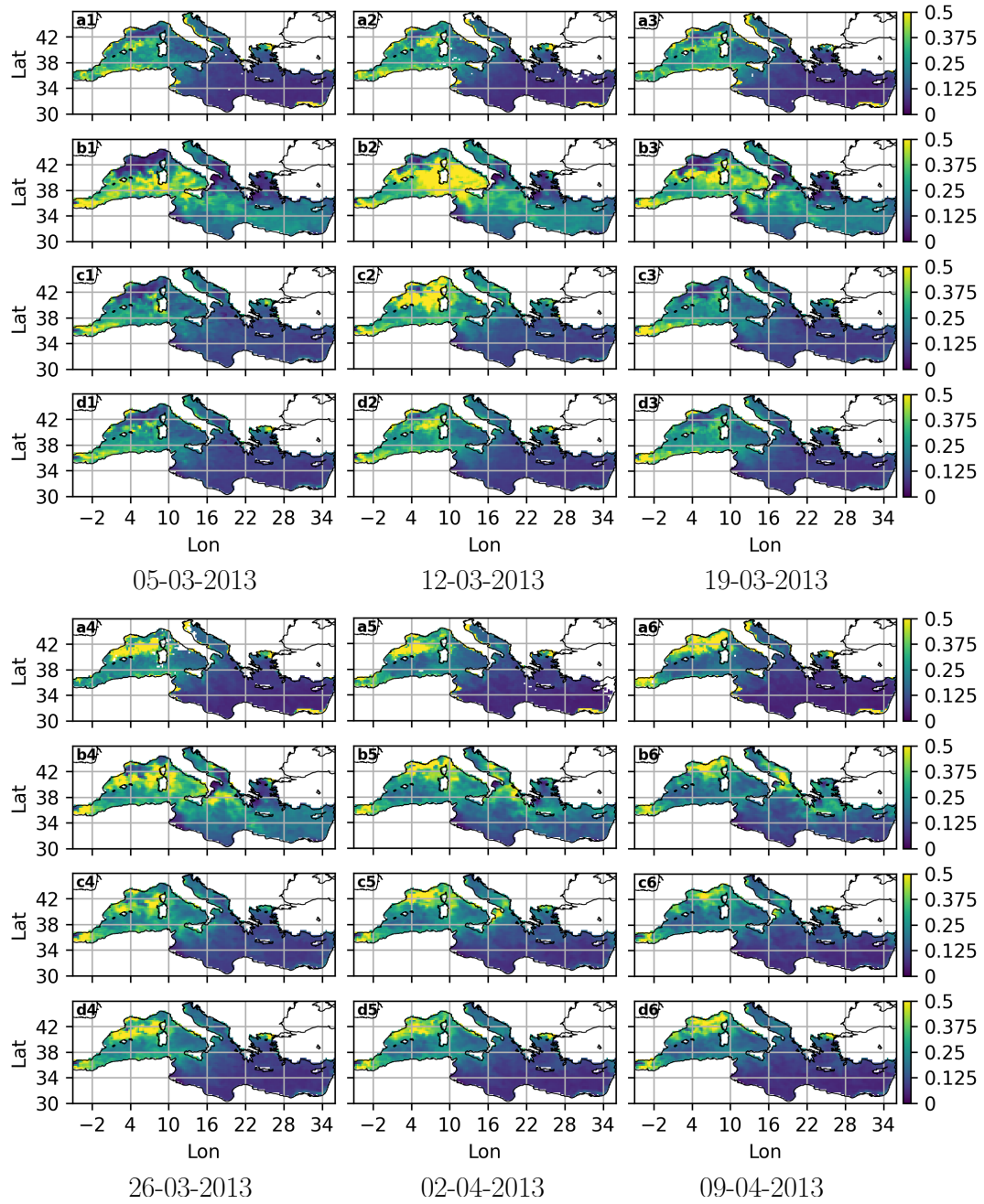
| Test name                   | Order<br>(h) | Chlorophyll<br>RMSD ( $mg/m^3$ ) | Phosphate<br>RMSD ( $mmol/m^3$ ) | Nitrate<br>RMSD ( $mmol/m^3$ ) |
|-----------------------------|--------------|----------------------------------|----------------------------------|--------------------------------|
|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

**Table 2.** Summary of the root mean square deviation (RMSD) indicators. Results for control run, assimilation experiment (in bold) and sensitivity analysis tests.

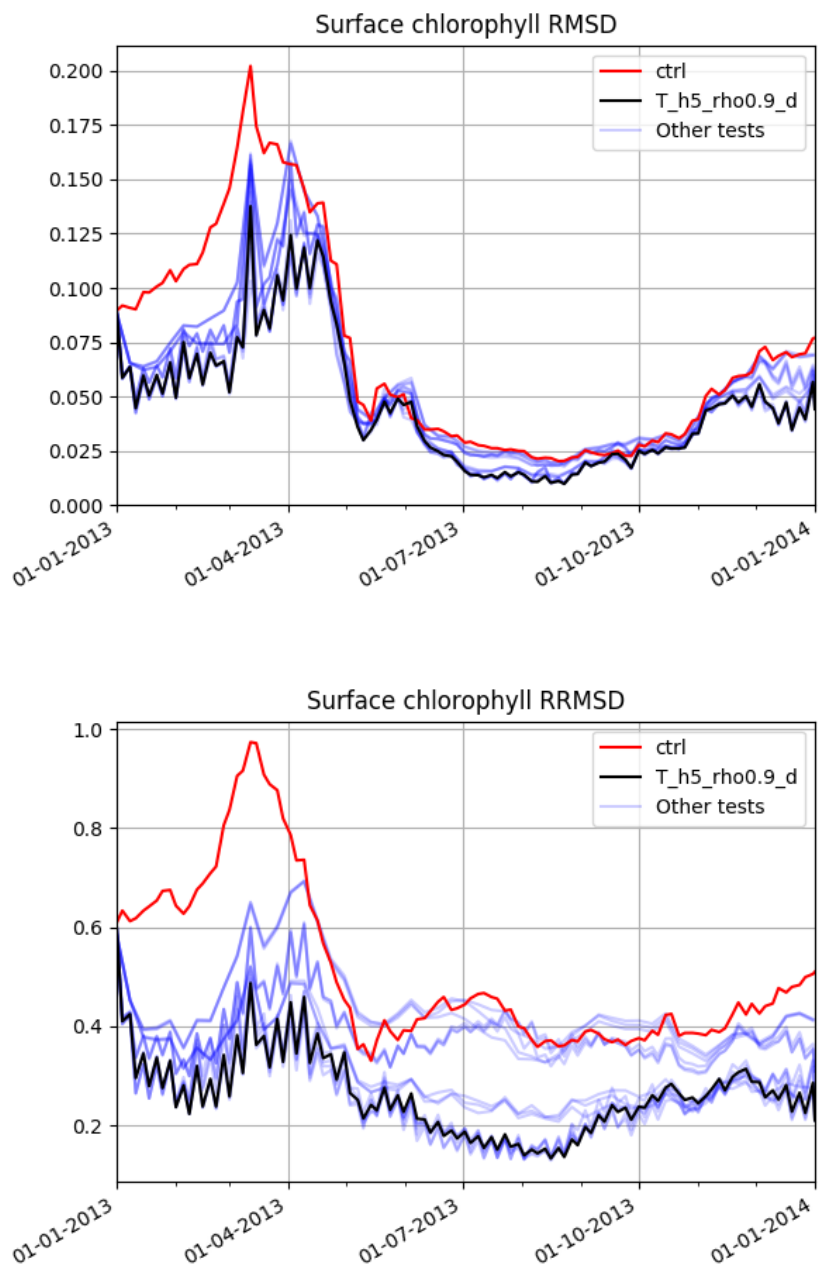
quality of the system state uncertainty estimation (Section 3.3). The results are discussed in detail through a qualitative and quantitative comparison with the control run (i.e., a hindcast run with no assimilation). The RMSD between simulation (before  
145 assimilation) and observations (Section 2.2) is used as a quantitative performance indicator.

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, the 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.



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



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

The surface chlorophyll estimated by the 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, end of March - beginning of April 2013) is shown in Fig. 2. The RMSD time series is shown in Fig. 3.

The 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. 2a), the start of the surface bloom in the whole area of the northern 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. 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. 2a).

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. 2b). The maps of the pre-assimilation for the March 5 and 12 dates of the GHOSH run (Fig. 2, 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, 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 values of surface chlorophyll can vary considerably during the year, the relative root mean square deviation (RRMSD, i.e., 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  
185 summer, due to a general smaller 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.

## 190 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).

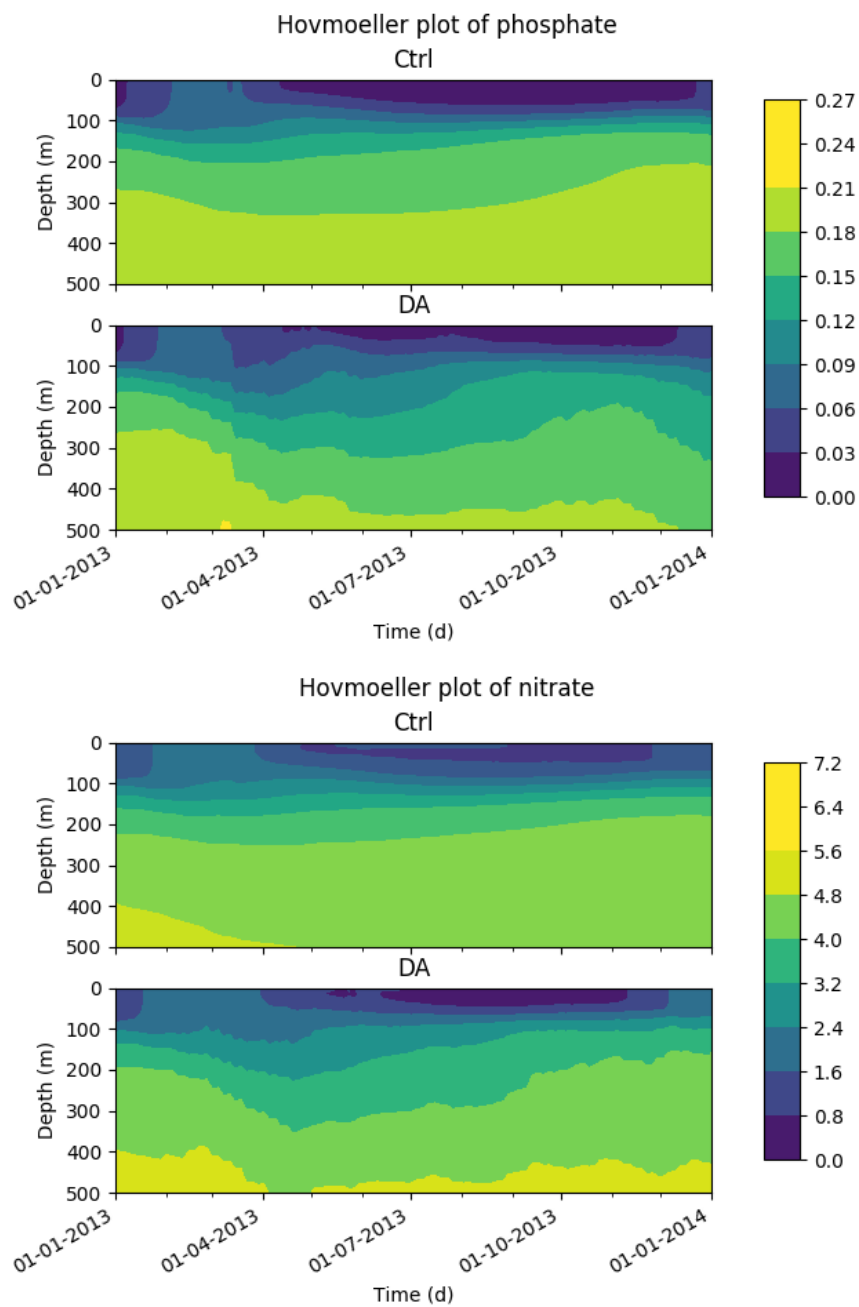
The Hovmöller diagrams of phosphate and nitrate in the DA and the control runs show the effects of the assimilation on the  
195 two nutrient compartments (Fig. 4). Both nitrate and phosphate show lower concentrations in the 150–400 m layer from April in the simulation with satellite assimilation, with 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, and the seasonal sequence of late-spring/winter mixed conditions and of summer stratification is preserved.  
200 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.

## 3.3 Uncertainty estimation

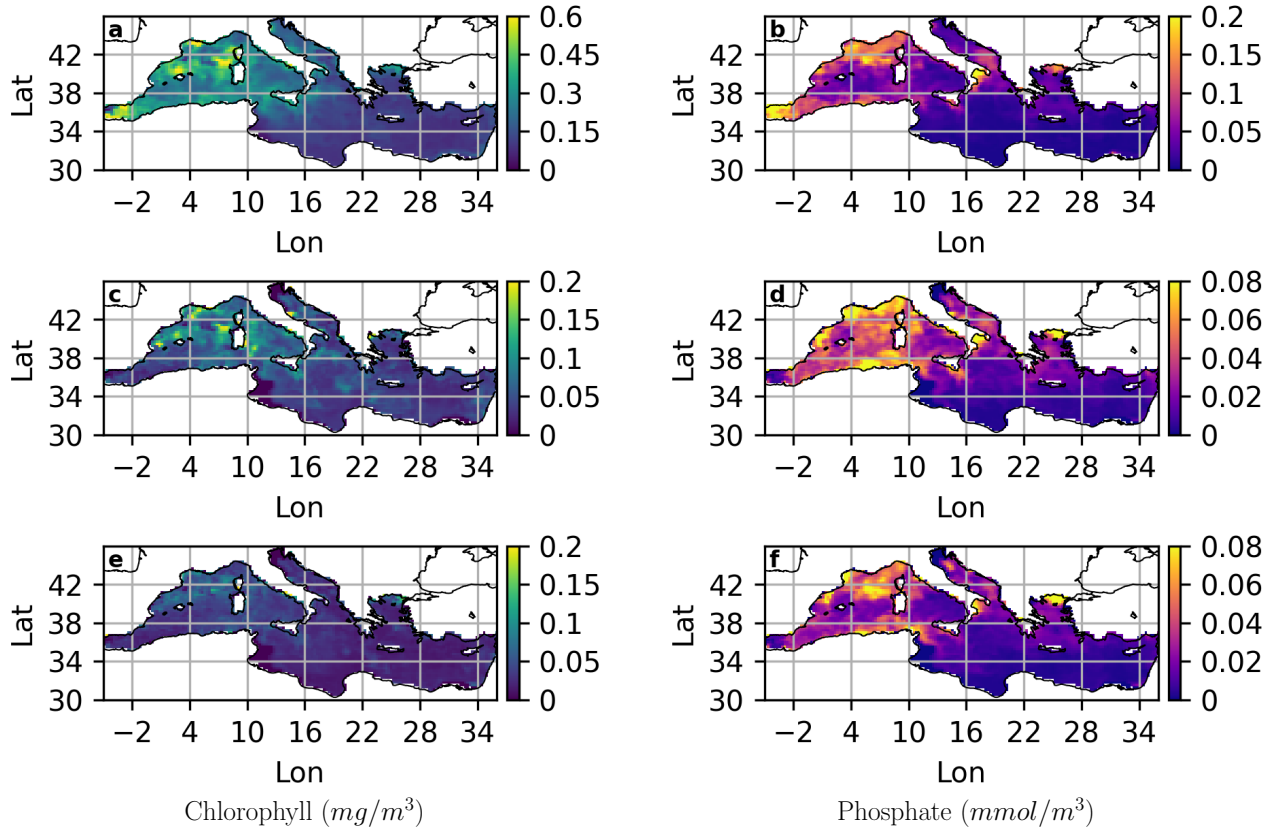
In the GHOSH filter, a weighted ensemble is used to estimate the uncertainty of the forecast and analysis states. The standard  
205 deviation of the weighted ensemble is used to quantify the uncertainty, whereas a comparison of the standard deviations before and after assimilation is calculated to evaluate 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 bloom (March 26), the spatial patterns of chlorophyll and phosphate standard deviations at the surface  
210 (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.

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.  
215 Moreover, a relevant reduction in the standard deviation, in relative terms, is also observed in the less productive eastern Mediterranean Sea.



**Figure 4.** Hovmöller diagrams of nutrients (phosphate and nitrate,  $\text{mmol}/\text{m}^3$ ): comparison between the control run *ctrl* (up) and the assimilation experiment (bottom)



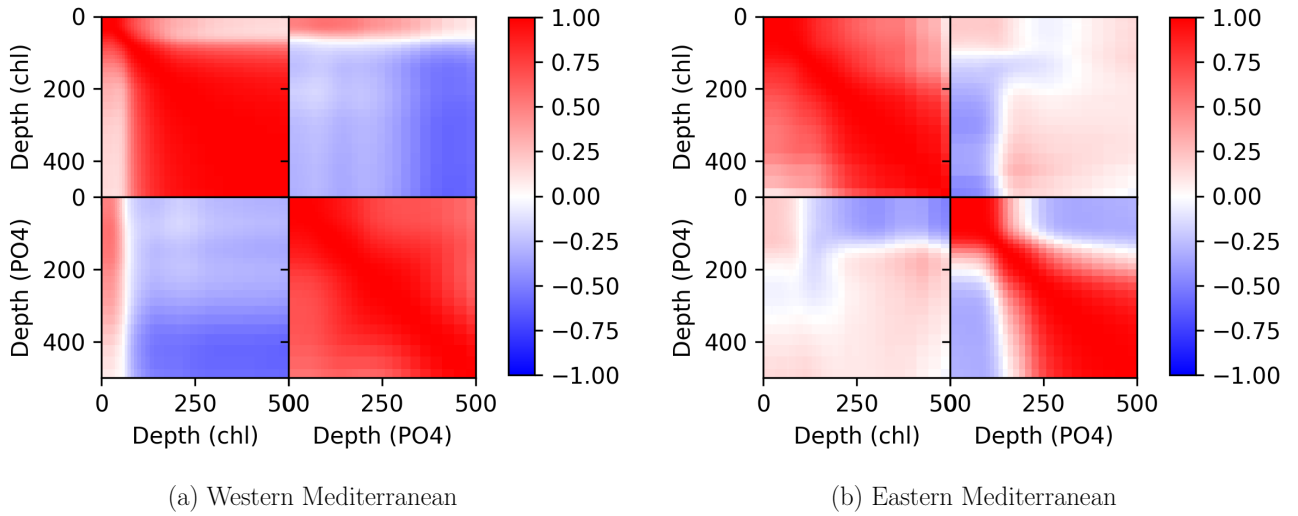
**Figure 5.** Maps of chlorophyll and phosphate (a and b) and maps of their uncertainty (ensemble standard deviation) before assimilation (c and d) and after assimilation (e and f) for the day 26.03.2013.

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.

220 The system has been tuned to focus on open sea and the assimilation in shallow waters is less effective due to imposed higher observation uncertainties in the coastal areas (see Section 2.4 for details).

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  
 225 of the Eastern Mediterranean Sea, chlorophyll and phosphate are less correlated than in the western Mediterranean, whereas



**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.

the impact of surface assimilation on chlorophyll is deeper (i.e., correlations with surface chlorophyll is high down to nearly 400 m).

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

### 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 marginal compared with the model integration time and the I/O operations, being less than 1% of the total computational time.

## 4 Discussion

In the realistic 3D high-dimensional parallel implementation with assimilation of real satellite data, the GHOSH filter proved (in Section 3) to be a reliable data assimilation scheme with 30% RMSD reduction for the assimilated variable compared to the control run. Moreover, the GHOSH assimilation did not degrade non-assimilated variables, actually improving their RMSDs with respect to independent observations (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 and Fig. 6). Further, a sensitivity analysis underlines the role of the high-order sampling in improving performances in a non-assimilated variable (Section 3.4).

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 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, with the normalization rules chosen in this experiment (i.e.,  $\Lambda$  in Section 2.3), the marginal effects of the higher approximation order on chlorophyll and phosphate could be motivated by the fact that 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 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  $\Lambda$  used in the sampling procedure plays an important role in the filter, since it defines which variables are of more interest in the computation of the most relevant components of the state uncertainty. When implementing GHOSH in the realistic OGSTM-BFM application, we used a normalization matrix to represent  $\Lambda$  (Section 2.3) in order to properly scale the different biogeochemical variables. However, this strategy implies that the 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., processes/dynamics occurring in deep layers). Alternative options can be investigated 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),  $\Lambda$  could be designed to focus on the most relevant processes. In particular,  $\Lambda$  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  $\Lambda$ , a preliminary approach is applied for the definition of the parametric uncertainty covariance matrix  $\mathbf{Q}$  (Section 2.3). The diagonal matrix  $\mathbf{Q}$  can not be very representative in the relatively complex OGSTM-BFM application. A more advanced parameterization of  $\mathbf{Q}$  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).

However, even in its basic implementation, the adopted hybrid approach (together with the GHOSH's lower need of inflation proved in Spada et al., 2024) prevented the “collapse into the mean” effect (not shown), 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.,  $\rho = 1$ ).

The implemented statistical moments  $\mu$  (Section 2.3) describe a Gaussian uncertainty pdf (see also Discussion in Spada et al., 2024). However, the log-transformation of the tracers before applying the assimilation filter implies a log-Gaussian un-  
280 certainty pdf on concentrations, which is well-suited for non-negative biogeochemical 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 ca-  
285 pability of managing non-linearity (showed in Part 1) not only with the model operator but also with the non-linear observation operator.

From a computational point of view, the 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. This behaviour is also in line with the GHOSH asymptotic computational complexity  
290 that is the same as in EnKF second-order schemes (see Part 1 Spada et al., 2024).

## 5 Conclusions

GHOSH is a novel DA ensemble filter that was introduced in Part 1, and features an enhanced approximation order. In this work, it has been tested in a 3D 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  
295 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  
300 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. 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).

305 *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 is archived on Zenodo (<https://doi.org/10.5281/zenodo.12819521>).

## 310 Appendix A: High-order sampling solutions

Given a certain maximum approximation order  $h$  and the dimension  $s$  of the subspace in which the  $h^{\text{th}}$ -order approximation is achieved, the GHOSH filter (during its initialization) needs a solution of a specific non-linear system to encode the moments information ( $\mu$ ) in the GHOSH sampling matrix  $\Omega^h$  and in the vector  $\mathbf{v}$  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  $\Omega^h$  matrices (and corresponding  $\mathbf{v}$  vectors) are possible. Since  $\Omega^h$  and  $\mathbf{v}$  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 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  $\mathbf{v}$  and adding orthonormal columns in order to form an orthogonal matrix.

320 In the case with  $h = 3$  and  $s = 8$ , a solution with constant weights ( $v_m = \frac{1}{4}$  for each  $m \in \{1, \dots, 16\}$ ) is chosen, namely,

$$\Omega^{h=3} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}^T. \quad (\text{A1})$$

In the case with  $h = 5$  and  $s = 3$ , the chosen solution is

$$\mathbf{v} = \left( \frac{1}{\sqrt{5}} \quad \frac{1}{\sqrt{5}} \quad \frac{1}{2\sqrt{5}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \cdots \right. \\ \left. \cdots \quad \frac{1}{2\sqrt{5}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \quad \frac{1}{2\sqrt{6}} \right)^T \quad (\text{A2})$$

325 and

$$\Omega^{h=5} = \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & \cdots \\ 0 & 0 & 0 & \frac{1}{\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & \cdots \\ 0 & 0 & 0 & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \cdots \\ \cdots & -\frac{1}{2} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & \cdots \\ \cdots & 0 & \frac{1}{\sqrt{6}} & \frac{1}{2\sqrt{6}} & \frac{1}{2\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & -\frac{1}{2\sqrt{6}} & \cdots \\ \cdots & 0 & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \cdots \end{pmatrix}^T. \quad (\text{A3})$$

*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.
- 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.
- 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.
- 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.
- 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.
- 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.

- 375 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.
- 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.
- 380 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.
- 385 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.
- 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.
- 390 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.
- 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.
- 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](https://doi.org/10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2), 2001.
- 395 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.
- 400 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.
- 405 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.
- 410

- 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.
- 415 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.
- 420 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.
- 425 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.
- 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.
- 430 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.
- 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.