Progress in Oceanography 216 (2023) 103050

Contents lists available at ScienceDirect

Progress in Oceanography

journal homepage: www.elsevier.com/locate/pocean

# Bayesian learning of coupled biogeochemical–physical models

Abhinav Gupta, Pierre F.J. Lermusiaux *

*Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Mass. Ave., Cambridge, MA 02139, United States of America*
*Center for Computational Science and Engineering, Massachusetts Institute of Technology, 77 Mass. Ave., Cambridge, MA 02139, United States of America*

ARTICLE INFO

*Keywords:*
Dynamical systems
Bayesian data assimilation
Uncertainty quantification
Dynamically orthogonal
Gaussian mixture models
Model learning
Machine learning
Stochastic PDEs
Ocean and weather prediction
Ecosystem modeling

ABSTRACT

Predictive dynamical models for marine ecosystems are used for a variety of needs. Due to the sparse measurements and limited understanding of the myriad of ocean processes, there is however significant uncertainty. There is model uncertainty in the parameter values, functional forms with diverse parameterizations, and level of complexity needed, and thus in the state variable fields. We develop a Bayesian model learning methodology that allows interpolation in the space of candidate dynamical models and discovery of new models from noisy, sparse, and indirect observations, all while estimating state variable fields and parameter values, as well as the joint probability distributions of all learned quantities. We address the challenges of high-dimensional and multidisciplinary dynamics governed by partial differential equations (PDEs) by using state augmentation and the computationally efficient Gaussian Mixture Model — Dynamically Orthogonal filter. Our innovations include stochastic formulation parameters and stochastic complexity parameters to unify candidate models into a single general model as well as stochastic expansion parameters within piecewise function approximations to generate dense candidate model spaces. These innovations allow handling many compatible and embedded candidate models, possibly none of which are accurate, and learning elusive unknown functional forms that augment these models. Our new Bayesian methodology is generalizable and interpretable. It seamlessly and rigorously discriminates among existing models, but also extrapolates out of the space of models to discover new ones. We perform a series of twin experiments based on flows past a ridge coupled with three-to-five component ecosystem models, including flows with chaotic advection. We quantify the learning skill, and evaluate convergence and the sensitivity to hyper-parameters. Our PDE framework successfully discriminates among functional forms and model complexities, and learns in the absence of prior knowledge by searching in dense function spaces. The probabilities of known, uncertain, and unknown model formulations, and of biogeochemical–physical fields and parameters, are updated jointly using Bayes' law. Non-Gaussian statistics, ambiguity, and biases are captured. The parameter values and the model formulations that best explain the noisy, sparse, and indirect data are identified. When observations are sufficiently informative, model complexity and model functions are discovered.

## 1. Introduction

The ability to predict and understand marine ecosystems is essential for addressing many of the grand challenges faced by humanity, such as climate change, food security, and sustainability. In broad terms, marine ecosystems can be seen as food webs, or flows of food/energy from nutrients to phytoplankton, to zooplankton, to fish, and finally recycling back to the nutrients (Lalli and Parsons, 1997; Fennel and Neumann, 2014). However, there does not yet exist a single generic model that accurately represents all the components in marine food webs due to the presence of highly complex biological processes with many unknown interactions. Therefore, many approximations are made in such ecosystem models. In addition, only parts of a food web are commonly modeled. The interactions of what is modeled with the portions of the food web that are not modeled are then either neglected or parameterized in terms of the modeled variables. Biology is also forced by complex nonlinear physics. Most biogeochemical–physical modeling systems thus broadly categorize the nutrients and individual species, representing them as continuous state variable fields, defined as concentrations of nutrients, biomass, or number of organisms per unit volume of water. The dynamics of these fields consists of reaction terms representing biogeochemical processes such as nutrient uptake, grazing, death, etc., and of forcing by physical processes such as advection, diffusion, and sunlight. Each reaction term or physical process is commonly modeled mathematically, using functional forms or terms that contain multiple parameters and have different levels of accuracy.

* Corresponding author at: Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Mass. Ave., Cambridge, MA 02139, United States of America.
*E-mail addresses:* guptaa@mit.edu (A. Gupta), pierrel@mit.edu (P.F.J. Lermusiaux).

https://doi.org/10.1016/j.pocean.2023.103050
Received 28 August 2022; Received in revised form 13 May 2023; Accepted 18 May 2023
Available online 30 May 2023

A plethora of biogeochemical modeling systems have been proposed, each of which with many model formulations (Hofmann and Friedrichs, 2002; Fennel et al., 2022). The models differ in their complexity, or ability to resolve different biological processes. Models of higher complexity have more biological components, functional terms, and parameters. However, process terms and parameters are often poorly known, which hampers the utility of highly complex models (Franks, 2002; Ward et al., 2010; Denman, 2003). The simplest models are 3-component nutrient–phytoplankton–zooplankton (NPZ) models (Franks et al., 1986; Flierl and McGillicuddy, 2002). NPZ models are easily understood and serve an important role in ocean research. Including the intermediate state of detritus leads to four component NPZ-Detritus biological models (Davis and Steele, 1994). Intermediate complexity models involve around 7 to 10 components, adding bacteria, nitrate, ammonium, and dissolved organic nitrogen (Fasham et al., 1990), or related state variables (Beşiktepe et al., 2003). One of the most complex lower-trophic-level marine ecosystem models is the European Regional Seas Ecosystem Model (ERSEM, Baretta et al., 1995; Baretta, 1997; Blackford et al., 2004), originally developed for the North Sea. Many choices of functional forms exist for each of the biological processes (Franks, 2002), leading to application-specific variants of the above models.

Biogeochemical models are commonly developed semi-empirically, leading to uncertainty in their parameters, functional forms, and level of complexity. What is adequate for a particular ocean region may not work elsewhere or may need to be updated due to seasonal or other variabilities (Lermusiaux et al., 2004). Such model uncertainties transfer to the state variables predicted, complicating model learning by direct comparisons of state variables with in situ data. As a result, when observations are employed to develop models, it is often in an offline mode, fitting parameter values or functional forms to data in controlled experiments. With data assimilation, we could however use observations in a direct Bayesian sense, to jointly learn state variables, parameter values, and discriminate/discover functional forms with quantifiable uncertainty (Lermusiaux, 2007). Most biogeochemical data assimilation (Robinson and Lermusiaux, 2002; Dowd et al., 2014) can be categorized broadly into two categories. The first is parameter estimation, where model parameters are calibrated by minimizing misfits between model output fields and independent observations (Friedrichs et al., 2007; Losa et al., 2004; Ward et al., 2010; Mattern et al., 2012; Toyoda et al., 2013). The second is sequential estimation, where observations collected are used to update model states during the forward model integration (Beşiktepe et al., 2003; Mattern et al., 2010; Allen et al., 2003; Natvik and Evensen, 2003; Hu et al., 2012). However, very few studies deal with the simultaneous estimation of parameters, state variables, and model equations. Doron et al. (2011) used a Monte Carlo ensemble of 200 simulations lasting 30-days in the North Atlantic and conducted idealized twin experiments with surface observations of phytoplankton to estimate parameters and states with a Kalman filter-based scheme and state augmentation. Jones et al. (2010) performed state and parameter estimation in a nonlinear phytoplankton–zooplankton model using two Markov Chain Monte Carlo (MCMC) algorithms in an identical-twin setting. Mattern et al. (2013) used a nonlinear particle filter scheme to assimilate satellite sea surface color and jointly estimate the state and parameters of a three-dimensional biological ocean model. Lately, along with state and parameter estimation, the selection of optimal complexity of biogeochemical models has become a new area of research (Lermusiaux, 2007; Ward et al., 2010; Giricheva, 2015; Ward et al., 2013). Because of the multiscale and intermittent variability of marine ecosystems, there is also a need for generalized and adaptive modeling, where models can learn and adapt during run-time (Lermusiaux et al., 2004; Evangelinos et al., 2003; Tian et al., 2004; Lermusiaux et al., 2011).

Several machine learning methods have been developed for the discovery of model equations. The sparse regression-based methods (SINDy; Brunton et al., 2016; Rudy et al., 2019) are promising as they do not require prior knowledge, however, they often require large data sets. Variations of SINDy include weak SINDy to learn PDEs (Messenger and Bortz, 2021), adaptive generation of features to increase the library of models (Kulkarni et al., 2020), and extensions to Bayesian identification (Niven et al., 2020). Deep learning methods have been derived to obtain marine ecosystem closure models (Gupta and Lermusiaux, 2021, 2023). Genetic algorithms (Maslyaev et al., 2019) and reinforcement learning (Bassenne and Lozano-Durán, 2019; Novati et al., 2021; Wang et al., 2019) have been used to search the space of candidate models. However, most of these approaches do not provide uncertainty estimates for the discovered models. Methods have also combined prior knowledge about underlying governing equations for model recovery and refinement. For example, Raissi and Karniadakis (2018) used Gaussian processes to learn the values of the parametric response of partially-known nonlinear differential equations. Unfortunately, data and knowledge of governing laws are luxuries in the case of realistic biogeochemical models.

It is clear that fundamental methods for identifying dynamical models that best explain sparse data in accord with prior governing laws and uncertainties would be most useful. The Bayesian theory and schemes of Lu and Lermusiaux (2014, 2021) address several of the above needs and drawbacks, using noisy, sparse, and indirect observations for joint Bayesian inference of states and parameters along with probabilistic discrimination among candidate models. However, several questions remain: Could we avoid assimilating observations independently in each candidate model when there are so many models to choose from? And if none of these models are that accurate, could the Bayesian machine find the elusive true formulations? Could it interpolate within and extrapolate out of known model spaces, while providing accurate joint probability distributions for model states, parameters, and formulations? Could such Bayesian learning be efficient and accurate with high-dimensional and multidisciplinary physical–biogeochemical stochastic PDEs? The overall goal of the present paper is thus to extend and generalize the discrimination-based model learning developed in Lu and Lermusiaux (2014, 2021) to allow for interpolation in the space of candidate models and discovery of new models, in an efficient fashion. Our novel learning and discovery of differential models are achieved by introducing stochastic formulation parameters, stochastic complexity parameters, and piecewise function approximations assembled with stochastic expansion parameters. We address the challenges of multidisciplinary dynamics and develop a rigorous PDE Bayesian learning framework using state augmentation and the Gaussian Mixture Model — Dynamically Orthogonal (GMM-DO) filter (Sondergaard and Lermusiaux, 2013a,b). The final estimates are notably joint probability distributions for all learned quantities. To our knowledge, it is the first time that sequential Bayesian data assimilation is developed to predict and update the joint probability distributions of state variables, parameters, and known, uncertain, and unknown model formulations, enabling the Bayesian discovery of model functional forms and model complexities, with applications to high-dimensional ocean physical–biogeochemical dynamical systems.

In Section 2, we present the problem statement. In Section 3, we develop the general Bayesian learning methodology with novel parameters for model learning and discovery. In Section 4, we describe the stochastic biogeochemical–physical equations and simulated experiments. In Section 5, we apply our methodology to four sets of experiments of varying complexities and learning objectives. Conclusions are provided in Section 6.

## 2. Problem statement

A single mathematical model that exactly captures all the physical and biological processes occurring in the real world does not yet exist. Hence, there is inherent model uncertainty that manifests in many forms, including: initial and boundary condition uncertainties; unreliable parameter values; multiple competing candidate model functions;

unknown functional forms; missing model terms; and, debatable complexity of the model. In this work, we consider discriminating among candidate models, learning among compatible models, and discovering new model formulations. Compatible models are models that can be related to a single dynamical system theoretically and that can also be combined numerically. Compatible models can nonetheless represent different dynamics, e.g., our goals include learning which dynamics are or are not present based on observations.

In general, we consider a stochastic dynamical modeling system defined on a domain $D$, governing the uncertain spatiotemporal dynamics of $\boldsymbol{\phi}(\boldsymbol{x},t;\omega) : \mathbb{R}^n \times [0,T] \to \mathbb{R}^{N_v}$, the stochastic state vector comprising $N_v$ dynamical state variable fields (e.g., physical fields, biogeochemical concentration fields, etc.). The realization index $\omega$ belongs to a measurable sample space $\Omega$ and the model depends on a vector $\boldsymbol{\theta}(\omega)$ of $N_\theta$ uncertain parameters. The main notation used is defined in Table D.2. To encompass the majority of scenarios, we write the general form of the uncertain dynamical modeling system as follows,

$$\begin{aligned} &\frac{\partial \boldsymbol{\phi}(\boldsymbol{x},t;\omega)}{\partial t} = \mathcal{L}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega),\boldsymbol{\theta}(\omega),\boldsymbol{x},t] + \widehat{\mathcal{L}}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega);\omega] + \widetilde{\mathcal{L}}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega);\omega]\,, \\ &\qquad\qquad \boldsymbol{x}\in D,\ t\in[0,T], \quad \omega\in\Omega\,, \\ &\text{with}\quad \boldsymbol{\phi}(\boldsymbol{x},0;\omega) = \boldsymbol{\phi}_o(\boldsymbol{x};\omega)\,, \\ &\text{and}\quad \mathcal{B}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega)] = \boldsymbol{b}(\boldsymbol{x},t;\omega),\ \boldsymbol{x}\in\partial D,\ t\in[0,T],\ \omega\in\Omega\,, \end{aligned} \tag{1}$$

where $\boldsymbol{\phi}_o(\boldsymbol{x};\omega)$, $\mathcal{B}$, and $\boldsymbol{b}(\boldsymbol{x},t;\omega)$ are the stochastic initial conditions, boundary condition operators, and boundary values respectively. The functional form of the first dynamics term $\mathcal{L}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega),\boldsymbol{\theta}(\omega),\boldsymbol{x},t]$ is assumed to be known, but with uncertain parameters $\boldsymbol{\theta}(\omega)$. The second term $\widehat{\mathcal{L}}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega);\omega] \in \{\widehat{\mathcal{L}}_1[\boldsymbol{\phi}(\boldsymbol{x},t;\omega);\omega],\dots,\widehat{\mathcal{L}}_{N_m}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega);\omega]\}$, represents a set of compatible candidate functional forms, where $N_m$ is the number of candidates. For example, for reaction terms, model functions are often from the polynomial, exponential, and/or sinusoidal families, and can be rational or irrational functions. The third term $\widetilde{\mathcal{L}}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega);\omega]$ has a functional form completely unknown. Each of these three functional terms has uncertainties, hence the $\omega$ dependence. Their summation encompasses common scenarios, e.g., their multiplication simply absorbs the more known types into the most unknown type. The stochastic initial and boundary condition formulations can also have uncertain function forms, similar to the dynamical modeling system itself, i.e., they can be known, belonging to a family, or unknown.

In some cases, candidate models have different complexities,

$$\mathcal{M}_i : \begin{cases} \frac{\partial \phi^i_1(\boldsymbol{x},t;\omega)}{\partial t} = & \mathcal{L}^i_1[\phi^i_1(\boldsymbol{x},t;\omega),\dots,\phi^i_{N_v(i)}(\boldsymbol{x},t;\omega),\boldsymbol{\theta}^i(\omega),\boldsymbol{x},t;\omega] \\ \qquad\vdots & \qquad\qquad , i=1,\dots,N_m \\ \frac{\partial \phi^i_{N_v(i)}(\boldsymbol{x},t;\omega)}{\partial t} = & \mathcal{L}^i_{N_v(i)}[\phi^i_1(\boldsymbol{x},t;\omega),\dots,\phi^i_{N_v(i)}(\boldsymbol{x},t;\omega),\boldsymbol{\theta}^i(\omega),\boldsymbol{x},t;\omega] \end{cases}, \tag{2}$$

where each model, $\mathcal{M}_i$, has $N_v(i)$ number of state variable fields ($\{\phi^i_1,\dots,\phi^i_{N_v(i)}\}$) from a pool of candidates, and their aggregates. In such situations, the candidate models can often remain compatible with each other, for example, low-complexity models are embedded in higher-complexity ones. We refer to such classes of candidate models as, *compatible-embedded models*. The number $N_v$ of dynamical state variable fields then denotes the number of state variables needed to encompass all models $\mathcal{M}_i$'s. Of course, in general, uncertainty in parameter values, functional forms, and complexities occur simultaneously, thus, each term $\{\mathcal{L}^i_1,\dots,\mathcal{L}^i_{N_v(i)}\}$ in Eq. (2) can encompass the $\widehat{\mathcal{L}}$ and $\widetilde{\mathcal{L}}$ terms introduced in Eq. (1). Such scenarios are exemplified in our series of experiments in Section 5. For example, in Experiments-2, three- and four-component biogeochemical models are considered to be candidate $\mathcal{M}_i$'s; in Experiments-1 & 4, the zooplankton mortality function is considered to be either linear or quadratic, corresponding to $\widehat{\mathcal{L}}$; and in Experiments-3, the zooplankton mortality function is assumed completely unknown, corresponding to $\widetilde{\mathcal{L}}$.

Let $\boldsymbol{\Phi}(t;\omega) \in \mathbb{R}^{N_v N_x}$ denote the spatially discretized state vector of the continuous field $\boldsymbol{\phi}(\boldsymbol{x},t;\omega)$. where $N_x$ denotes the dimension of the discretized state space. In the experiments, we assume that all observations $\mathcal{Y}(t;\omega)$ are noisy, sparse, and indirectly related to $\boldsymbol{\Phi}(t;\omega)$ by a stochastic linear measurement model from the state to the data space,

$$\mathcal{Y}(t;\omega) = \boldsymbol{H}\boldsymbol{\Phi}(t;\omega) + \boldsymbol{V}(t;\omega), \qquad \boldsymbol{V}(t;\omega) \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{R}) \tag{3}$$

where $N_y$ is the number of available observations; $\boldsymbol{H} \in \mathbb{R}^{N_y \times N_v N_x}$ the observation matrix; and $\boldsymbol{V} \in \mathbb{R}^{N_y}$ a zero-mean, uncorrelated Gaussian observation noise with covariance matrix $\boldsymbol{R} \in \mathbb{R}^{N_y \times N_y}$. The latter noise in the observations is larger than the sensor noise as it also contains representation errors (Janjić et al., 2018): in our experiments, it will be a fraction of the variability in the state variables (Lermusiaux et al., 2000; Lermusiaux, 2002). The noisy observations are also sparse ($N_y \ll N_v N_x$) and indirect: they are available only at discrete time-instants, $t_k$ for $k = 1,2,\dots,K$, at a very limited number of spatial locations, and only for one state variable. As in reality, the specific variable being measured will vary from experiment to experiment.

In summary, our specific objectives are thus two-fold, first to solve the stochastic forward-modeling system (Eqs. (1) & (2)), taking into account all the associated uncertainties including compatible, compatible-embedded, and unknown model terms; and second to simultaneously learn, in the Bayesian sense, the state fields, parameters, and model equations based on the stochastic, sparse, and indirect observation model (Eq. (3)). Our Bayesian learning thus needs to evolve the prior and posterior joint probabilities of state fields, parameters, and model formulations, given the noisy observations available and all other uncertainties. The overall goal is to accurately represent these probability density functions (pdfs), including the marginal probabilities of known, uncertain, and unknown model formulations. It is only if the noisy observations are sufficiently informative about either the state fields, parameters, and model formulations, that the Bayesian machine can identify the true state variables, true parameters, and true model. If the noisy observations are not sufficiently informative, the perfect Bayesian machine will not lead to a unique identification, but provide the exact posterior probabilities of the models, parameter values, and state variable fields.

## 3. General Bayesian learning methodology

Prior to developing our general methodology, we briefly review the Bayesian learning for rigorous discrimination among candidate differential dynamical models (Lu and Lermusiaux, 2014, 2021). In this Bayesian discrimination, each candidate model then evolves the joint pdf of its state variables and parameters, independently from other models, and provides probability distributions that are conditional on the candidate model. In other words, each model runs its own probabilistic forecast. When noisy observations are made, the model-conditional state variables and parameters (all contained in $\boldsymbol{\Phi}$), and also the model pdfs themselves, are updated using Bayes' rules (Bayes and Price, 1763; Bertsekas and Tsitsiklis, 2008),

$$\begin{aligned} p_{\boldsymbol{\Phi}|\mathcal{Y},\mathcal{M}}(\boldsymbol{\Phi}|\boldsymbol{y},\mathcal{M}_i) &= \frac{p_{\mathcal{Y}|\boldsymbol{\Phi},\mathcal{M}}(\boldsymbol{y}|\boldsymbol{\Phi},\mathcal{M}_i)}{p_{\mathcal{Y}|\mathcal{M}}(\boldsymbol{y}|\mathcal{M}_i)} \\ &\qquad \times p_{\boldsymbol{\Phi}|\mathcal{M}}(\boldsymbol{\Phi}|\mathcal{M}_i)\,, \quad \forall\,\boldsymbol{\Phi}\in\mathbb{R}^{N_v N_x}, \forall\, i\in\{1,\dots,N_m\}\,, \\ p_{\mathcal{M}|\mathcal{Y}}(\mathcal{M}_n|\boldsymbol{y}) &= \frac{p_{\mathcal{Y}|\mathcal{M}}(\boldsymbol{y}|\mathcal{M}_i)}{p_{\mathcal{Y}}(\boldsymbol{y})} p_{\mathcal{M}}(\mathcal{M}_i)\,, \quad \forall\, i\in\{1,\dots,N_m\}\,, \end{aligned} \tag{4}$$

where $\mathcal{M}_i$ is the $i$th model candidate and the pdfs $p_{\boldsymbol{\Phi}|\mathcal{M}}(\boldsymbol{\Phi}|\mathcal{M}_i)$ and $p_{\boldsymbol{\Phi}|\mathcal{Y},\mathcal{M}}(\boldsymbol{\Phi}|\boldsymbol{y},\mathcal{M}_i)$ are the prior and posterior model-conditional state variable distributions, respectively. The model distribution $p_{\mathcal{M}}(\bullet)$ is the prior probability for each of the candidates being the true model and $p_{\mathcal{M}|\mathcal{Y}}(\bullet|\boldsymbol{y})$ is the corresponding posterior model distribution. This pdf $p_{\mathcal{M}|\mathcal{Y}}(\bullet|\boldsymbol{y})$ allows learning by exact Bayesian discrimination among candidate models and is one of the main novelty of Lu and Lermusiaux (2014, 2021). In particular, when the noisy observations are not sufficient to achieve unequivocally the ultimate learning objective, these

posterior pdfs will adequately represent the ambiguity including possible multimodal distributions and the effects of biases in the candidate models (Lu and Lermusiaux, 2021).

The above Bayesian learning evolves each stochastic candidate model separately. To increase efficiency and allow the discrimination among many more models, all the way to a continuous space of models, this should be circumvented. For example, when models are compatible or compatible-embedded, the learning could interpolate in these model spaces, or even extrapolate out of them. New capabilities are also needed to enable Bayesian learning of unknown models. Next, we thus develop new stochastic parameterizations that unify all such candidate models into a single general modeling system. We recast the model learning into new parameter estimation problems, using stochastic formulation and complexity parameters (Section 3.1) and piece-wise function approximation theory with stochastic expansion parameters (Section 3.2). We then evolve the joint probabilities of the state fields, the regular parameters, and these new stochastic formulation, complexity, and expansion parameters, using stochastic DO equations (Section 3.3). At each observation time, we perform Bayesian learning using the GMM-DO filter with state augmentation (Section 3.3). Our methodology does not need to compute the discrete marginal likelihoods, $p_{\mathcal{Y}|\mathcal{M}}(\boldsymbol{y}|\mathcal{M}_i)$; instead, it learns in a parameterized continuous model space. We thus extend learning among discrete model formulations to learning within a continuous infinite range of formulations as well as across models of different complexities and into unknown models. In other words, we remain able to discriminate among existing models, but we can now also interpolate in or extrapolate out of the space of models to discover new ones.

### 3.1. Stochastic formulation and complexity parameters: Compatible and compatible-embedded models

Let us first consider the case where, when according to prior scientific knowledge, the uncertain model belongs to a set of compatible candidate functional forms ($\hat{\mathcal{L}}[\bullet]$; Eq. (1)). In order to recast this learning problem with multiple models into a learning problem with a single model and parameter estimation, the compatible candidate model functions are added to each other but only after being multiplied with novel stochastic parameters. Each of the candidates is thus assigned a new stochastic formulation parameter that can take discrete or continuous values depending on the learning objectives and prior knowledge. For example, binary values would be utilized to discriminate between the presence or absence of certain functions, while other values would be utilized to allow some linear interpolation within the space defined by the compatible candidate models. To complete Bayesian learning, when noisy observations are collected, the probability distributions of these *stochastic formulation parameters*, $\alpha_k(\omega)$'s, $k = 1, \dots, N_m$, are updated and their mean values estimated alongside these of other regular parameters $\boldsymbol{\theta}(t;\omega)$, using state augmentation. Summarizing, the general model can thus be written as a stochastic linear combination of the candidates,

$$\hat{\mathcal{L}}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega),t;\omega] = \sum_{k=1}^{N_m} \alpha_k(\omega)\mathcal{L}_k[\boldsymbol{\phi}(\boldsymbol{x},t;\omega),\boldsymbol{x},t;\omega] \,. \tag{5}$$

where the distributions of the $\alpha_k(\omega)$'s are updated at each observation time. This new formulation can thus both help select active candidate functions and identify their linear combinations. It allows interpolating in the space of known candidate functions.

Next, we extend this approach to learn model complexity (Eq. (2)). This is achieved by defining new states $\phi'_k$ that are the original states $\phi_k$ multiplied with new *stochastic complexity parameters* $\beta_k(\omega)$. Hence, we define $\phi'_k = \beta_k(\omega)\phi_k$ and a new general model, $\mathcal{L}'_k$, which encompasses all the candidates in the class of compatible-embedded models,

$$\begin{aligned}&\frac{\partial \phi'_k(\boldsymbol{x},t;\omega)}{\partial t}\\ &= \mathcal{L}'_k[\phi'_1(\boldsymbol{x},t;\omega),\dots,\phi'_{N_v}(\boldsymbol{x},t;\omega),\boldsymbol{\theta}(t;\omega),\boldsymbol{\beta}(\omega),\boldsymbol{x},t;\omega], \quad k = 1,\dots,N_v\end{aligned} \tag{6}$$

where $N_v = \max\{N_v(i)\}_{i=1}^{N_m}$ or in general the number of state variables needed to encompass all models $\mathcal{M}_i$'s. By learning this vector of new complexity parameters $\boldsymbol{\beta}(\omega)$, we can eliminate certain state variables or aggregate them to form new states, and determine the model of appropriate complexity that best explains the noisy observed data.

To illustrate such combinations of compatible-embedded models into a general model, let us consider a case with only two candidate models ($N_m = 2$ in Eq. (2)). Let us further assume that the set of states of the first model ($\{\phi_1,\dots,\phi_{N_v(1)}\}$) are fully contained within the set of states of the second model ($\{\phi_1,\dots,\phi_{N_v(1)},\dots,\phi_{N_v(2)}\}$), and the goal is to discriminate between the presence or absence of either of the models. Using a new complexity parameter $\beta(\omega)$ that is allowed to take only binary values and substituting new states variables defined as $\phi'_1 = \phi_1$, ..., $\phi'_{N_v(1)} = \phi_{N_v(1)}$, $\phi'_{N_v(1)+1} = \beta(\omega)\phi_{N_v(1)+1}$, ..., $\phi'_{N_v(2)} = \beta(\omega)\phi_{N_v(2)}$, the general model can be written as (based on Eq. (2) and omitting explicit dependence on $\boldsymbol{x}$, $t$, & $\omega$ for brevity),

$$\begin{aligned}
\frac{\partial \phi'_1}{\partial t} &= (1-\beta)\mathcal{L}^1_1[\phi'_1,\dots,\phi'_{N_v(1)},\theta^1]\\
&\quad + \beta\mathcal{L}^2_1[\phi'_1,\dots,\phi'_{N_v(1)},\phi'_{N_v(1)+1},\dots,\phi'_{N_v(2)},\theta^2]\,,\\
&\vdots\\
\frac{\partial \phi'_{N_v(1)}}{\partial t} &= (1-\beta)\mathcal{L}^1_{N_v(1)}[\phi'_1,\dots,\phi'_{N_v(1)},\theta^1]\\
&\quad + \beta\mathcal{L}^2_{N_v(1)}[\phi'_1,\dots,\phi'_{N_v(1)},\phi'_{N_v(1)+1},\dots,\phi'_{N_v(2)},\theta^2]\,,\\
\frac{\partial \phi'_{N_v(1)+1}}{\partial t} &= \beta\mathcal{L}^2_{N_v(1)+1}[\phi'_1,\dots,\phi'_{N_v(1)},\phi'_{N_v(1)+1},\dots,\phi'_{N_v(2)},\theta^2]\,,\\
&\vdots\\
\frac{\partial \phi'_{N_v(2)}}{\partial t} &= \beta\mathcal{L}^2_{N_v(2)}[\phi'_1,\dots,\phi'_{N_v(1)},\phi'_{N_v(1)+1},\dots,\phi'_{N_v(2)},\theta^2]\,,
\end{aligned} \tag{7}$$

where $\beta(\omega) = 0$ leads to the first candidate model, and $\beta(\omega) = 1$ to the second candidate model. In similar fashion, we can derive the general model for cases with more than two candidate models, with states in one model being aggregate of states in other models, etc.

### 3.2. Stochastic piecewise linear function approximations: Unknown models

Model learning with the above new stochastic formulation and complexity parameters requires a set of candidate functional forms to choose from. However, in some cases, there might be no such prior information/candidates available, hence there is an unknown part $\widetilde{\mathcal{L}}$ in the dynamical model (Eq. (1)). Such dynamical model functions then need to be discovered. A way to achieve this is to parameterize the function space of the unknown dynamics, for example, using orthogonal polynomials. For the scalar right-hand-side of a single dynamical equation, if the polynomials are defined over the whole range of the scalar biogeochemical state variable, one would need high-order polynomials for the function approximation to achieve sufficient accuracy. The result is then susceptible to spurious oscillations. To remedy this, as in finite elements, one can divide the range of the biogeochemical variable into pieces and use low-order polynomials within each piece that are stitched together to approximate the unknown dynamical model function over its whole range. Generalizing, we thus propose to parameterize the unknown function space using stochastic piece-wise continuous functions, building on approximation theory (Trefethen, 2019). In the present work, we consider dense piecewise linear functions as this representation is both rich and simple, and provides practical approximations of any unknown function. As we will showcase, it expands the functional space in which the Bayesian search is performed and enables searching outside of known models and the discovery of new learned functions.

For brevity, let us only illustrate the scalar case, where $\widetilde{\mathcal{L}}[\phi(\boldsymbol{x},t;\omega);\omega]$ is the unknown function (Eq. (1)) of a single scalar state variable. As biological concentration is finite, we assume that prior

information about the range of values taken by the state variable is available, $\phi(\boldsymbol{x},t;\omega) \in [\phi_L,\phi_R]$, $\forall \boldsymbol{x} \in \mathcal{D}$ and $t \in [0,T]$. Now, to define a parameterization using continuous piece-wise linear segments, this range $\mathcal{H} = [\phi_L,\phi_R]$ is divided into an indexed collection of $N_I$ number of intervals with non-zero measure, $\{I_i = [\phi_L^i,\phi_R^i]\}_{1\le i\le N_I}$, forming a partition of the range $\mathcal{H}$, i.e.,

$$\mathcal{H} = \bigcup_{i=1}^{N_I} I_i \quad \text{and} \quad I_i \cap I_j = \emptyset \quad \text{for } i \neq j \,, \tag{8}$$

and we use $N_I + 1$ points to discretize the range, such that,

$$\phi_L = \phi_L^1 < \phi_R^1 = \phi_L^2 < \cdots < \phi_R^{N_I-1} = \phi_L^{N_I} < \phi_R^{N_I} = \phi_R \,. \tag{9}$$

Let $\{\Psi_1,\ldots,\Psi_{N_I+1}\}$ be the linear functions defined on these elements,

$$\Psi_1(\phi) = \begin{cases} \frac{1}{(\phi_R^1-\phi_L)}(\phi_R^1-\phi) & \text{if } \phi \in I_1 \,, \\ 0 & \text{otherwise} \end{cases}$$

$$\Psi_k(\phi) = \begin{cases} \frac{1}{(\phi_R^{k-1}-\phi_L^{k-1})}(\phi-\phi_L^{k-1}) & \text{if } \phi \in I_{k-1} \,, \\ \frac{1}{(\phi_R^{k}-\phi_L^{k})}(\phi_R^{k}-\phi) & \text{if } \phi \in I_{k} \,, \\ 0 & \text{otherwise} \end{cases} \quad \text{for } k \in \{2,\ldots,N_I\} \,, \tag{10}$$

$$\Psi_{N_I+1}(\phi) = \begin{cases} \frac{1}{(\phi_R-\phi_L^{N_I})}(\phi-\phi_L^{N_I}) & \text{if } \phi \in I_{N_I} \,, \\ 0 & \text{otherwise} \end{cases}$$

and $\gamma_k(\omega)'s$, $k \in 1,\ldots,N_I+1$ be $N_I+1$ *stochastic expansion parameters* that parameterize the unknown function space by taking a linear combination of the functions defined on each element. Hence, all together we obtain:

$$\widetilde{\mathcal{L}}[\phi(\boldsymbol{x},t;\omega);\omega] = \sum_{k=1}^{N_I+1} \gamma_k(\omega)\Psi_k(\phi(\boldsymbol{x},t;\omega)) \,. \tag{11}$$

Thus, estimating the stochastic expansion parameters $\gamma_k$'s, in turn, leads to the learning of the unknown model function. The above formulation ensures $C^0$ continuity in the functional space and is equivalent to linear interpolating splines (Lambers, 2023). The prior pdf of these parameters defines the functional space in which the search is performed. By construction, this parameterized space can be made as dense as desired, by increasing the number of realization $\omega$ of $\gamma_k(\omega)$'s. Throughout this paper, we only consider linear segments as the basis, however, this formulation can be extended to any other basis, such as higher-degree splines.

### 3.3. Bayesian learning: stochastic DO PDEs, GMM-DO filter, and learning skill

To provide an accurate and informative prior for our new Bayesian learning paradigm with uncertain and unknown nonlinear dynamics and PDEs, we employ Dynamically Orthogonal (DO) equations (Sapsis and Lermusiaux, 2009, 2012; Feppon and Lermusiaux, 2018b). The DO equations are instantaneously optimal reduced-order differential equations that evolve, based on the governing nonlinear dynamics, the dominant probabilistic subspace. Their derivation with uncertain parameters is outlined in Appendix B and in Section 4.3 for biogeochemical specifics.

For the Bayesian learning at each observation time, the GMM-DO filter (Sondergaard and Lermusiaux, 2013a,b) is used to perform nonlinear, non-Gaussian updates of the probability distribution of all quantities estimated, as detailed in Appendix C. For the joint Bayesian learning of state variables and parameters, we combine the GMM-DO filter with state augmentation (Gelb, 1974; Lu and Lermusiaux, 2021) ($\boldsymbol{\Phi}_{aug}$; Appendix D). In essence, the DO equations evolve the initial joint probability distribution for the augmented state variable, $p_{\boldsymbol{\Phi}_{aug}}(\boldsymbol{\Phi}_{aug}(0;\omega))$, to obtain the forecast/prior joint distribution at the observation time $t$, $p_{\boldsymbol{\Phi}^f_{aug}}(\boldsymbol{\Phi}^f_{aug}(t;\omega))$. This prior pdf is approximated with multivariate Gaussian Mixture Models (GMMs)

$$p_{\boldsymbol{\Phi}^f_{aug}}(\boldsymbol{\Phi}^f_{aug}(t;\omega)) \approx \sum_{j=1}^{N_{\text{GMM}}} \pi^f_{\boldsymbol{\Phi}_{aug},j} \times \mathcal{N}(\boldsymbol{\Phi}^f_{aug}(t;\omega); \boldsymbol{\mu}^f_{\boldsymbol{\Phi}_{aug},j}, \boldsymbol{\Sigma}^f_{\boldsymbol{\Phi}_{aug},j}) \tag{12}$$

using an efficient fit in the DO subspace (Appendix C). Given the property that GMMs are conjugate priors to Gaussian observation models (Eq. (3)), their Bayesian update remains a GMM (Sondergaard and Lermusiaux, 2013a; Casella and Berger, 2021), providing the posterior pdf, $p_{\boldsymbol{\Phi}^a_{aug}}(\boldsymbol{\Phi}^a_{aug}(t;\omega))$,

$$p_{\boldsymbol{\Phi}^a_{aug}}(\boldsymbol{\Phi}^a_{aug}(t;\omega)) \approx \sum_{j=1}^{N_{\text{GMM}}} \pi^a_{\boldsymbol{\Phi}_{aug},j} \times \mathcal{N}(\boldsymbol{\Phi}^a_{aug}(t;\omega); \boldsymbol{\mu}^a_{\boldsymbol{\Phi}_{aug},j}, \boldsymbol{\Sigma}^a_{\boldsymbol{\Phi}_{aug},j}) \tag{13}$$

where, $\forall j \in \{1,\ldots,N_{\text{GMM}}\}$,

$$\begin{aligned} \pi^a_{\boldsymbol{\Phi}_{aug},j} &= \frac{\pi^f_{\boldsymbol{\Phi}_{aug},j} \times \mathcal{N}(\boldsymbol{y}; \boldsymbol{H}_{aug}\boldsymbol{\mu}^f_{\boldsymbol{\Phi}_{aug},j}, \boldsymbol{H}_{aug}\boldsymbol{\Sigma}^f_{\boldsymbol{\Phi}_{aug},j}\boldsymbol{H}^T_{aug} + \boldsymbol{R})}{\sum_{m=1}^{N_{\text{GMM}}} \pi^f_{\boldsymbol{\Phi}_{aug},m} \times \mathcal{N}(\boldsymbol{y}; \boldsymbol{H}_{aug}\boldsymbol{\mu}^f_{\boldsymbol{\Phi}_{aug},m}, \boldsymbol{H}_{aug}\boldsymbol{\Sigma}^f_{\boldsymbol{\Phi}_{aug},m}\boldsymbol{H}^T_{aug} + \boldsymbol{R})} \,, \\ \boldsymbol{\mu}^a_{\boldsymbol{\Phi}_{aug},j} &= \boldsymbol{\mu}^f_{\boldsymbol{\Phi}_{aug},j} + \boldsymbol{K}_j(\boldsymbol{y} - \boldsymbol{H}_{aug}\boldsymbol{\mu}^f_{\boldsymbol{\Phi}_{aug},j}) \,, \\ \boldsymbol{\Sigma}^a_{\boldsymbol{\Phi}_{aug},j} &= (\boldsymbol{I} - \boldsymbol{K}_j\boldsymbol{H}_{aug})\boldsymbol{\Sigma}^f_{\boldsymbol{\Phi}_{aug},j} \,, \\ \boldsymbol{K}_j &= \boldsymbol{\Sigma}^f_{\boldsymbol{\Phi}_{aug},j}\boldsymbol{H}^T_{aug}(\boldsymbol{H}_{aug}\boldsymbol{\Sigma}^f_{\boldsymbol{\Phi}_{aug},j}\boldsymbol{H}^T_{aug} + \boldsymbol{R})^{-1} \,. \end{aligned} \tag{14}$$

Further, using the properties of affine transformation, the above Bayesian update is performed in the DO subspace (Appendix C), thus rendering it computationally tractable.

Our novel schemes allow for efficient simultaneous Bayesian estimation of state variable fields, parameters, and model equations themselves, all while using a single modeling system. They recast the learning of compatible and compatible-embedded models into formulation and complexity parameter estimations (Section 3.1) and, to allow the discovery of formulations, parameterize the space of unknown model functions using piece-wise linear continuous functions (Section 3.2). For the former, the learning occurs within the space of candidate models while for the latter, it occurs outside of that space and into the space of unknown model functions, hence providing the capability for model discovery. Importantly, this discovery is interpretable as it is in the form of piece-wise continuous functions. In addition, all of our Bayesian estimations provide much more than maximum likelihood estimates: they predict and update the complete joint probability distribution of states, parameters, and models. If the noisy, sparse, and indirect observations are not sufficiently informative to learn and eliminate all but one model, parameter value, or state variable field, our Bayesian learning estimates the correct multi-modal pdfs. Our learning can indeed represent ambiguity, e.g. multiple options are possible, or even equifinality (Hart et al., 2000), e.g. a set of model estimates have the same likelihood. It can also signal the presence of bias in competing model formulations. Such capabilities will be showcased in Section 5.

To evaluate the learning skill, we first compare the mean fields and parameters with the noisy observations, using several error metrics. We also analyze the evolution of the pdfs of fields and parameters, as well as the convergence of these pdfs with stochastic resolution.

The definitions and notation for the hyper-parameters used in the DO methodology and the GMM-DO filter are provided in Table D.2. To summarize, in Fig. 1, we provide an overview of our general Bayesian model learning methodology with references to relevant equations, sections, and appendices.

## 4. Biogeochemical–physical equations and simulated experiments setup

In this section, we describe the specifics of our simulated Bayesian learning experiments. We start with the biogeochemical differential equations, their coupling with the physics PDEs, and the stochastic DO decomposition with uncertain and unknown terms. This is followed by

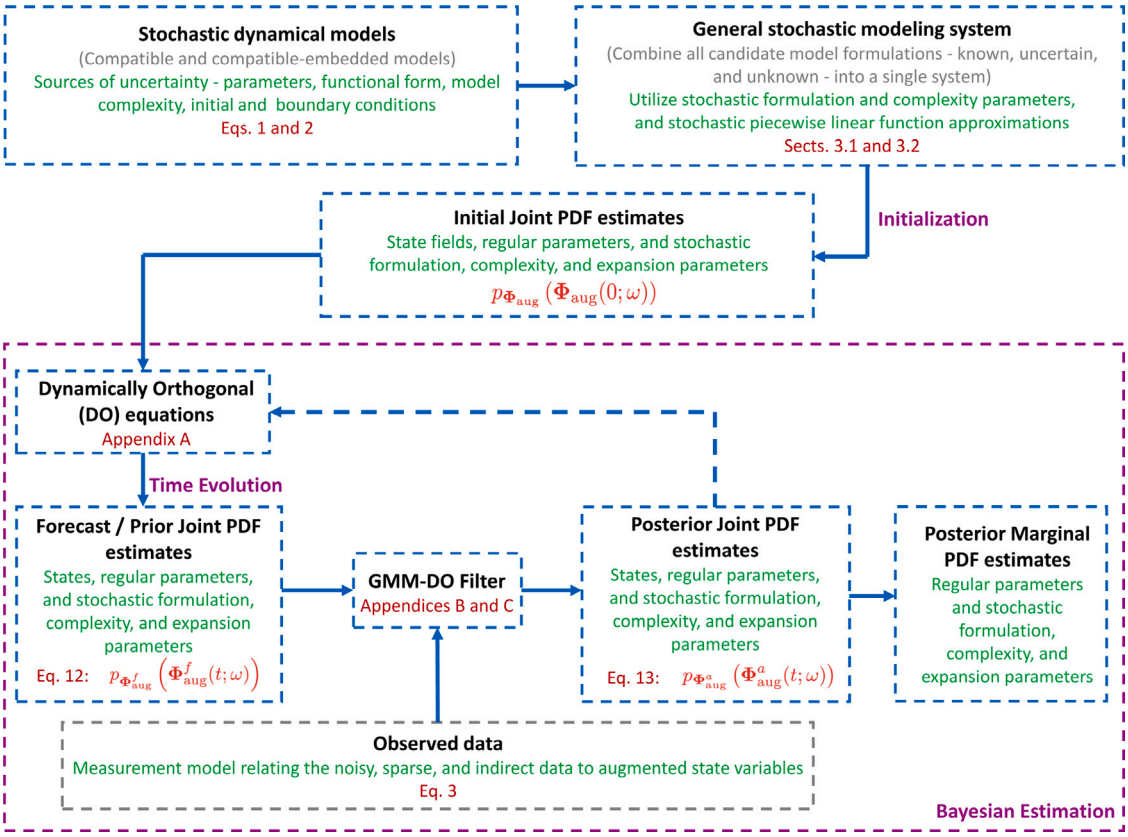



**Fig. 1.** Overview of the general Bayesian model learning methodology.

details of the modeling domain, numerical methods, initialization of the stochastic simulations, true solution generation, simulated noisy, sparse, and indirect observations, and learning metrics.

### 4.1. Biogeochemical models

The biogeochemical differential equations that we employ are adapted from Tian et al. (2004) and Lermusiaux (2007) and references therein, and from Newberger et al. (2003). They meet the criterion of being compatible with each other, with low complexity models being embedded in higher complexity models (compatible-embedded-models). We will utilize three reaction models: the three-component NPZ model, i.e., nutrients ($N$), phytoplankton ($P$), and zooplankton ($Z$); the four-component NPZD model that adds Detritus ($D$); and, the five-component NNPZD model that adds a second nutrient, simulating ammonia ($NH_4$), nitrate ($NO_3$), $P$, $Z$, and $D$.

The NPZ biogeochemical reaction model is given by,

$$\begin{aligned}
\frac{dN}{dt} &= -\underbrace{G\frac{PN}{N+K_u}}_{Nutrient\ Uptake} + \underbrace{\Xi P}_{Phyt.\ Mortality} + \underbrace{\Gamma Z}_{Zoo.\ Mortality} + \underbrace{R_m\gamma Z(1-\exp^{-\Lambda P})}_{Zoo.\ Egestion},\\
\frac{dP}{dt} &= \underbrace{G\frac{PN}{N+K_u}}_{Nutrient\ Uptake} - \underbrace{\Xi P}_{Phyt.\ Mortality} - \underbrace{R_m Z(1-\exp^{-\Lambda P})}_{Zoo.\ Grazing},\\
\frac{dZ}{dt} &= \underbrace{R_m(1-\gamma)Z(1-\exp^{-\Lambda P})}_{Zoo.\ Ingestion} - \underbrace{\Gamma Z}_{Zoo.\ Mortality},
\end{aligned} \tag{15}$$

where $G$ representing the optical model,

$$G = V_m\frac{\alpha I}{(V_m^2+\alpha^2 I^2)^{1/2}}, \quad \text{and} \quad I(z) = I_0\exp^{k_w z}, \tag{16}$$

$z$ is depth, and $I(z)$ models the availability of sunlight for photochemical reactions. The parameters in Eqs. (15) & (16) are: $k_w$, light attenuation by sea water; $\alpha$, initial slope of the $P-I$ curve; $I_0$, surface photosynthetically available radiation; $V_m$, phytoplankton maximum uptake rate; $K_u$, half-saturation constant for phytoplankton uptake of nutrients; $\Xi$, phytoplankton specific mortality rate; $R_m$, zooplankton maximum grazing rate; $\Lambda$, Ivlev grazing constant; $\gamma$, fraction of zooplankton grazing egested; and $\Gamma$, zooplankton specific excretion/mortality rate. In this NPZ model (Eq. (15)), the nutrient uptake by phytoplankton is governed by a Michaelis–Menten formulation, which amounts to a linear uptake relationship at low nutrient concentrations that saturates to a constant at high concentrations. The grazing of phytoplankton by zooplankton follows a similar behavior: their growth rate becomes independent of $P$ in case of abundance, but proportional to available $P$ when resources are scarce; hence, zooplankton grazing is modeled by an Ivlev function. The death rates of both $P$ and $Z$ are linear, and a portion of zooplankton grazing in the form of excretion goes directly to nutrients.

For the NPZD model, the only change is in the addition of detritus, which is the intermediate state through which dead plankton is converted to nutrients,

$$\begin{aligned}
\frac{dN}{dt} &= -G\frac{PN}{N+K_u} + \underbrace{\Phi D}_{Remineralization} + \Gamma Z,\\
\frac{dD}{dt} &= R_m\gamma Z(1-\exp^{-\Lambda P}) + \Xi P - \underbrace{\Phi D}_{Remineralization}.
\end{aligned} \tag{17}$$

However, for the NNPZD model, the nutrients are divided into ammonia and nitrates, which are the two most important forms of nitrogen in the ocean (Lalli and Parsons, 1997; Fennel and Neumann, 2014). This helps to capture new processes such as phytoplankton cells preferentially taking up ammonia over nitrates because the presence of

ammonia inhibits the activity of the enzyme nitrate reductase essential for the uptake kinetics, the pool of ammonia coming from remineralization of detritus, and part of this ammonia pool getting oxidized to become a source of nitrates referred to as nitrification, etc. Lalli and Parsons (1997), Fennel and Neumann (2014) and Beşiktepe et al. (2003). The NNPZD model is given by,

$$
\begin{aligned}
\frac{d\mathrm{NO}_3}{dt} &= \Omega \mathrm{NH}_4 - \underbrace{G\left[\frac{\mathrm{NO}_3}{\mathrm{NO}_3+K_u}\exp^{-\Psi_I \mathrm{NH}_4}\right]P}_{Nitrate\ Uptake}\,,\\
\frac{d\mathrm{NH}_4}{dt} &= -\Omega \mathrm{NH}_4 + \Phi D + \Gamma Z - \underbrace{G\left[\frac{\mathrm{NH}_4}{\mathrm{NH}_4+K_u}\right]P}_{Ammonia\ Uptake}\,,\\
\frac{dP}{dt} &= \underbrace{G\left[\frac{\mathrm{NO}_3}{\mathrm{NO}_3+K_u}\exp^{-\Psi_I \mathrm{NH}_4}+\frac{\mathrm{NH}_4}{\mathrm{NH}_4+K_u}\right]P}_{Nitrate\ +\ Ammonia\ Uptake}\\
&\quad - \Xi P - R_m Z(1-\exp^{-\Lambda P})\,,\\
\frac{dZ}{dt} &= R_m(1-\gamma)Z(1-\exp^{-\Lambda P}) - \Gamma Z\,,\\
\frac{dD}{dt} &= R_m \gamma Z(1-\exp^{-\Lambda P}) + \Xi P - \Phi D\,.
\end{aligned}
\tag{18}
$$

The above three reaction models aim to capture the lower-trophic-level (LTL) interactions in the ocean ecosystem. They are the Lagrangian or ordinary differential equation (ODE) versions of these models. For realistic ocean field simulations, the above rates of change are material derivatives of dynamic tracers that are coupled with the physics using advection–diffusion–reaction PDEs. Of course, these models are not directly applicable in every ocean region without parameter tuning or modifying the functional form of the reaction terms. Regional diversity is one of the reasons for parameter and functional form (model) uncertainties.

### 4.2. Coupling with the physics

In biogeochemical–physical models, the physics is provided by solving PDEs for the conservation of mass and momentum (Navier–Stokes), internal energy, and salt, e.g., the ocean primitive equations (Haley and Lermusiaux, 2010; Haley et al., 2015). These models often contain parameterizations to represent subgrid-scale processes (McWilliams, 2008; Hecht and Hasumi, 2013). In the present work, we employ the incompressible nonhydrostatic Reynolds-averaged Navier–Stokes (RANS) PDEs (Ferziger et al., 2002),

$$
\begin{aligned}
&\nabla\cdot\boldsymbol{u}(\boldsymbol{x},t) = 0, \qquad \boldsymbol{x}\in\mathcal{D}\,,\\
&\frac{\partial \boldsymbol{u}(\boldsymbol{x},t)}{\partial t} + \boldsymbol{u}(\boldsymbol{x},t)\cdot\nabla\boldsymbol{u}(\boldsymbol{x},t) = -\nabla p(\boldsymbol{x},t) + \nu_E\nabla^2\boldsymbol{u}(\boldsymbol{x},t)\,, \qquad \boldsymbol{x}\in\mathcal{D}\,,
\end{aligned}
\tag{19}
$$

where $\boldsymbol{u}(\boldsymbol{x},t)$ is the velocity field, $p(\boldsymbol{x},t)$ the pressure field, and $\nu_E$ the turbulent eddy viscosity.

The Lagrangian biogeochemical models (Section 4.1) are coupled with the physics using stochastic advection–diffusion–reaction (ADR) PDEs. For $N_\phi$ stochastic biogeochemical tracers, $\phi^i(\boldsymbol{x},t;\omega)$'s (concentration per unit volume, Kulkarni and Lermusiaux (2019)), we obtain,

$$
\begin{aligned}
&\frac{\partial \phi^i(\boldsymbol{x},t;\omega)}{\partial t} + \underbrace{\boldsymbol{u}(\boldsymbol{x},t)\cdot\nabla\phi^i(\boldsymbol{x},t;\omega)}_{\text{Advection}} - \underbrace{\mathcal{K}_E\nabla^2\phi^i(\boldsymbol{x},t;\omega)}_{\text{Diffusion}}\\
&= \underbrace{S^{\phi^i}(\phi^1,\dots,\phi^{N_\phi},\theta^1(\omega),\dots,\theta^{N_\theta}(\omega),\boldsymbol{x},t;\omega)}_{\text{Reaction}}\,,\\
&\qquad \text{for } i\in\{1,\dots,N_\phi\},
\end{aligned}
\tag{20}
$$

where $\boldsymbol{u}(\boldsymbol{x},t)$ is the deterministic velocity field governed by Eq. (19), $\mathcal{K}_E$ is the eddy diffusivity, $S^{\phi^i}(\phi^1,\dots,\phi^{N_\phi},\theta^1(\omega),\dots,\theta^{N_\theta}(\omega),\boldsymbol{x},t;\omega)$ are the reaction or source terms defined by the right-hand-side of the ODEs of Section 4.1, and the $\theta^l(\omega)$'s, $l=\{1,\dots,N_\theta\}$, are the uncertain biogeochemical parameters. Biogeochemical reactions are nonlinear in nature, hence, the PDEs (20) form a set of strongly nonlinear, stiff, and coupled PDEs.

### 4.3. Biogeochemical–physical stochastic dynamically-orthogonal PDEs

To solve the system of Eqs. ((19) & (20)) efficiently, we now develop the DO equations for the stochastic ADR PDEs (20) with model and parameter uncertainty. We first separate the reactions into known, uncertain, and unknown terms, and write Eq. (20) in vector form,

$$
\begin{aligned}
&\frac{\partial \boldsymbol{\phi}(\boldsymbol{x},t;\omega)}{\partial t} + \nabla\cdot(\boldsymbol{u}(\boldsymbol{x},t)\boldsymbol{\phi}(\boldsymbol{x},t;\omega)) - \mathcal{K}_E\nabla^2\boldsymbol{\phi}(\boldsymbol{x},t;\omega)\\
&\qquad = \boldsymbol{S}^{\boldsymbol{\phi}}(\boldsymbol{\phi}(\boldsymbol{x},t;\omega),\boldsymbol{\theta}(\omega),\boldsymbol{\beta}(\omega),\boldsymbol{x},t;\omega)\\
&\qquad + \widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}(\boldsymbol{\phi}(\boldsymbol{x},t;\omega),\boldsymbol{\theta}(\omega),\boldsymbol{\alpha}(\omega),\boldsymbol{\beta}(\omega),\boldsymbol{x},t;\omega)\\
&\qquad + \widetilde{\boldsymbol{S}}^{\boldsymbol{\phi}}(\boldsymbol{\phi}(\boldsymbol{x},t;\omega),\boldsymbol{\gamma}(\omega),\boldsymbol{x},t;\omega)\,,
\end{aligned}
\tag{21}
$$

where $\boldsymbol{\phi} = \left[\phi^i\right]_{i=1}^{N_\phi}$. The functional form of the first reaction term $\boldsymbol{S}^{\boldsymbol{\phi}}(\bullet) = \left[S^{\phi^i}(\bullet)\right]_{i=1}^{N_\phi}$ is assumed to be known, however it contains $N_\theta$ uncertain regular parameters $\boldsymbol{\theta}(\omega) = \left[\theta^k\right]_{k=1}^{N_\theta}$. The second term $\widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}(\bullet) = \left[\widehat{S}^{\phi^i}(\bullet)\right]_{i=1}^{N_\phi}$ is uncertain: it belongs to a family of candidate functions, parameterized using $N_\alpha$ stochastic formulation parameters $\boldsymbol{\alpha}(\omega) = \left[\alpha^k\right]_{k=1}^{N_\alpha}$, and may contain uncertain regular parameters $\boldsymbol{\theta}(\omega)$. The candidate models of different complexities are combined using $N_\beta$ stochastic complexity parameters $\boldsymbol{\beta}(\omega) = \left[\beta^k\right]_{k=1}^{N_\beta}$. The $\beta_k(\omega)$'s multiplied with the original biological tracer fields (as described in Section 3.1) are absorbed into $\phi_i$'s and not explicitly shown; however, $\beta_k(\omega)$'s usually appear on the right-hand-side (RHS) in $\boldsymbol{S}^{\boldsymbol{\phi}}(\bullet)$ and $\widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}(\bullet)$. The third term $\widetilde{\boldsymbol{S}}^{\boldsymbol{\phi}}(\bullet) = \left[\widetilde{S}^{\phi^i}(\bullet)\right]_{i=1}^{N_\phi}$ has a functional form completely unknown, and is parameterized using $N_\gamma$ stochastic expansion parameters $\boldsymbol{\gamma}(\omega) = \left[\gamma^k\right]_{k=1}^{N_\gamma}$.

The DO decomposition for the biogeochemical fields consists of the sum of the statistical mean $\bar{\boldsymbol{\phi}}$ with the sum of the $N_s$ modes $\tilde{\boldsymbol{\phi}}_i$ multiplied by their stochastic coefficients $Y_i$, all of which will be evolved dynamically using DO differential equations,

$$
\boldsymbol{\phi}(\boldsymbol{x},t;\omega) = \bar{\boldsymbol{\phi}}(\boldsymbol{x},t) + \sum_{i=1}^{N_s}\tilde{\boldsymbol{\phi}}_i(\boldsymbol{x},t)Y_i(t;\omega)\,.
\tag{22}
$$

The uncertain regular and formulation and complexity parameters are split into means and deviations, $\boldsymbol{\theta}(\omega) = \bar{\boldsymbol{\theta}} + \mathfrak{D}^{\boldsymbol{\theta}}(\omega)$, $\boldsymbol{\alpha}(\omega) = \bar{\boldsymbol{\alpha}} + \mathfrak{D}^{\boldsymbol{\alpha}}(\omega)$, and $\boldsymbol{\beta}(\omega) = \bar{\boldsymbol{\beta}} + \mathfrak{D}^{\boldsymbol{\beta}}(\omega)$. For the nonlinear reaction terms in $\boldsymbol{S}^{\boldsymbol{\phi}}(\bullet)$ and $\widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}(\bullet)$, as for the nonlinear path planning optimal propulsion term (Subramani and Lermusiaux, 2016; Subramani et al., 2018), we utilize a local Taylor series expansion around the statistical means, $\bar{\boldsymbol{\phi}}(\boldsymbol{x},t)$, $\bar{\boldsymbol{\theta}}$, $\bar{\boldsymbol{\alpha}}$, and $\bar{\boldsymbol{\beta}}$, to locally represent the nonlinear stochastic effects in the reaction equations as nonlinear mean terms plus stochastic deviations. As we will exemplify, for most uncertainties, such stochastic approximation is efficient for Bayesian learning as it maintains the significant computational advantages of DO with respect to the other methods (Branicki and Majda, 2013). Finally, for maximum accuracy, we evaluate the $\widetilde{\boldsymbol{S}}[\bullet]$ terms for every state realization in a Monte-Carlo fashion. Details on DO schemes are provided in Appendix B. Next, we directly provide the DO differential equations for the mean and for $N_s$ modes and stochastic coefficients with $i\in\{1,\dots,N_s\}$ (omitting

function arguments and using $j$, $n$, and $m$ as summation indices),

$$
\begin{aligned}
\frac{\partial \bar{\boldsymbol{\phi}}}{\partial t} &= -\nabla\cdot(\boldsymbol{u}\bar{\boldsymbol{\phi}}) + \mathcal{K}_E\nabla^2\bar{\boldsymbol{\phi}} + \boldsymbol{S}^{\boldsymbol{\phi}}\big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \beta=\bar{\beta}}} + \widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}\big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \alpha=\bar{\alpha},\\ \beta=\bar{\beta}}} + \mathbb{E}[\widetilde{\boldsymbol{S}}^{\boldsymbol{\phi}}]\,,\\
\frac{\partial \tilde{\boldsymbol{\phi}}_i}{\partial t} &= \boldsymbol{Q}_i - \sum_{j=1}^{N_s}\langle \boldsymbol{Q}_i, \tilde{\boldsymbol{\phi}}_j\rangle\tilde{\boldsymbol{\phi}}_j\,,\\
\frac{dY_i}{dt} &= \sum_{m=1}^{N_s}\langle \boldsymbol{F}_m, \tilde{\boldsymbol{\phi}}_i\rangle Y_m + \sum_{m=1}^{N_\theta}\left\langle \frac{\partial \boldsymbol{S}^{\boldsymbol{\phi}}}{\partial\theta_i}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \beta=\bar{\beta}}}, \tilde{\boldsymbol{\phi}}_i\right\rangle \mathfrak{D}^{\theta}_m\\
&\quad + \sum_{m=1}^{N_\beta}\left\langle \frac{\partial \boldsymbol{S}^{\boldsymbol{\phi}}}{\partial\beta}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \beta=\bar{\beta}}}, \tilde{\boldsymbol{\phi}}_i\right\rangle \mathfrak{D}^{\beta}_m \qquad (23)\\
&\quad + \sum_{m=1}^{N_\theta}\left\langle \frac{\partial \widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}}{\partial\theta_i}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \alpha=\bar{\alpha},\\ \beta=\bar{\beta}}}, \tilde{\boldsymbol{\phi}}_i\right\rangle \mathfrak{D}^{\theta}_m + \sum_{m=1}^{N_\alpha}\left\langle \frac{\partial \widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}}{\partial\alpha_i}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \alpha=\bar{\alpha},\\ \beta=\bar{\beta}}}, \tilde{\boldsymbol{\phi}}_i\right\rangle \mathfrak{D}^{\alpha}_m\\
&\quad + \sum_{m=1}^{N_\beta}\left\langle \frac{\partial \widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}}{\partial\beta_i}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \alpha=\bar{\alpha},\\ \beta=\bar{\beta}}}, \tilde{\boldsymbol{\phi}}_i\right\rangle \mathfrak{D}^{\beta}_m\\
&\quad + \left\langle \widetilde{\boldsymbol{S}}^{\boldsymbol{\phi}} - \mathbb{E}[\widetilde{\boldsymbol{S}}^{\boldsymbol{\phi}}], \tilde{\boldsymbol{\phi}}_i\right\rangle\,,
\end{aligned}
$$

where,

$$
\begin{aligned}
\boldsymbol{Q}_i &= -\nabla\cdot(\boldsymbol{u}\tilde{\boldsymbol{\phi}}_i) + \mathcal{K}_E\nabla^2\tilde{\boldsymbol{\phi}}_i + \frac{\partial \boldsymbol{S}^{\boldsymbol{\phi}}}{\partial\boldsymbol{\phi}}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \beta=\bar{\beta}}}\tilde{\boldsymbol{\phi}}_i + \sum_{j=1}^{N_s}\sum_{n=1}^{N_\theta} C^{-1}_{Y_iY_j}C_{\mathfrak{D}^{\theta}_nY_j}\frac{\partial \boldsymbol{S}^{\boldsymbol{\phi}}}{\partial\theta_n}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \beta=\bar{\beta}}}\\
&\quad + \sum_{j=1}^{N_s}\sum_{n=1}^{N_\beta} C^{-1}_{Y_iY_j}C_{\mathfrak{D}^{\beta}_nY_j}\frac{\partial \boldsymbol{S}^{\boldsymbol{\phi}}}{\partial\beta_n}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \beta=\bar{\beta}}} + \frac{\partial \widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}}{\partial\boldsymbol{\phi}}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \alpha=\bar{\alpha},\\ \beta=\bar{\beta}}}\tilde{\boldsymbol{\phi}}_i\\
&\quad + \sum_{j=1}^{N_s}\sum_{n=1}^{N_\theta} C^{-1}_{Y_iY_j}C_{\mathfrak{D}^{\theta}_nY_j}\frac{\partial \widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}}{\partial\theta_n}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \alpha=\bar{\alpha},\\ \beta=\bar{\beta}}} \qquad (24)\\
&\quad + \sum_{j=1}^{N_s}\sum_{n=1}^{N_\alpha} C^{-1}_{Y_iY_j}C_{\mathfrak{D}^{\alpha}_nY_j}\frac{\partial \widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}}{\partial\alpha_n}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \alpha=\bar{\alpha},\\ \beta=\bar{\beta}}} + \sum_{j=1}^{N_s}\sum_{n=1}^{N_\beta} C^{-1}_{Y_iY_j}C_{\mathfrak{D}^{\beta}_nY_j}\frac{\partial \widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}}{\partial\beta_n}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \alpha=\bar{\alpha},\\ \beta=\bar{\beta}}}\\
&\quad + \sum_{j=1}^{N_s} C^{-1}_{Y_iY_j}\mathbb{E}[Y_j\widetilde{\boldsymbol{S}}^{\boldsymbol{\phi}}]\,,\\
\boldsymbol{F}_m &= -\nabla\cdot(\boldsymbol{u}\tilde{\boldsymbol{\phi}}_m) + \mathcal{K}_E\nabla^2\tilde{\boldsymbol{\phi}}_m + \frac{\partial \boldsymbol{S}^{\boldsymbol{\phi}}}{\partial\boldsymbol{\phi}}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \beta=\bar{\beta}}}\tilde{\boldsymbol{\phi}}_m + \frac{\partial \widehat{\boldsymbol{S}}^{\boldsymbol{\phi}}}{\partial\boldsymbol{\phi}}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \theta=\bar{\theta},\\ \alpha=\bar{\alpha},\\ \beta=\bar{\beta}}}\tilde{\boldsymbol{\phi}}_m\,,
\end{aligned}
$$

with $C_{\bullet,\bullet}$ representing cross-covariances, $\mathbb{E}[\bullet]$ expectations, and $\langle\bullet,\bullet\rangle$ spatial inner-products operators.

### 4.4. Modeling domain and boundary conditions

Our modeling domain is inspired by Stellwagen Bank at the edge of Massachusetts Bay, which is a whale feeding ground (McGillicuddy et al., 1998; Lermusiaux, 2001; Beşiktepe et al., 2003; Pineda et al., 2015; Tian et al., 2015; Pershing et al., 2019; Silva, 2021), as well as by many other coastal banks and ridges. The experimental setup consists of a two-dimensional domain with a bathymetry obstacle (Fig. 2), which could be considered a slice of a wide seamount or bank, an idealized sill, or a cross-section of a ridge. In the rest of the text, we will consider our experimental setup as flow over an idealized ridge. The mean flow occurs from left to right in the positive $x$-direction over the ridge. Such flows can create an upwelling of nutrients, leading to phytoplankton blooms, zooplankton responses, and nutrient uptake and recycling.

A horizontal length scale of $D \approx 1$ km is chosen for the ridge, while the vertical height scale is $H \approx 50$ m. The overall transverse height of the domain is $H_{in} = 100$ m. The cross-ridge length of the domain is $L = 20$ km, with center of the ridge at $X_c = 7.5$ km.

Further, we only consider deterministic boundary conditions (BCs) models. The inlet at the left boundary has Dirichlet BCs for velocity, and zero Neumann for biological tracers,

$$
u = U,\ v = 0 \qquad \text{and} \qquad \frac{\partial \boldsymbol{\phi}^i}{\partial x} = 0, \qquad \text{at } x = 0\,,\ \text{for } i \in \{1,\dots,N_\phi\}\,. \qquad (25)
$$

On the top and bottom boundary, free slip for velocity and again zero Neumann for tracers are applied,

$$
\frac{\partial u}{\partial z} = 0,\ v = 0 \text{ and } \frac{\partial \boldsymbol{\phi}^i}{\partial z} = 0, \qquad \text{at } z = 0 \text{ and } z = h\,,\ \text{for } i \in \{1,\dots,N_\phi\}\,. \qquad (26)
$$

At the outlet on the right boundary, we have open BCs with zero Neumann for all the state variables,

$$
\frac{\partial u}{\partial x} = 0,\ \frac{\partial v}{\partial x} = 0 \qquad \text{and} \qquad \frac{\partial \boldsymbol{\phi}^i}{\partial x} = 0, \qquad \text{at } x = L\,,\ \text{for } i \in \{1,\dots,N_\phi\}\,. \qquad (27)
$$

Finally, on the obstacle surface, no-slip for velocity and zero Neumann for tracers are used,

$$
u = 0,\ v = 0 \qquad \text{and} \qquad \frac{\partial \boldsymbol{\phi}^i}{\partial x} = \frac{\partial \boldsymbol{\phi}^i}{\partial z} = 0, \qquad \text{at } z = He^{-(x-X_c)^2/D^2}\,,\ \text{for } i \in \{1,\dots,N_\phi\}\,. \qquad (28)
$$

### 4.5. Numerical schemes

The velocity and pressure fields are governed by the incompressible nonhydrostatic RANS PDEs (19). The stochastic biogeochemical fields are coupled with this dynamic RANS flow and governed by a dynamic reduced-order representation of the original stochastic ADR PDEs (20), the DO ADR PDEs we derived (Eqs. (23) & (24)). In all experiments Section 4.5, we solve the deterministic RANS-biogeochemical PDEs for the true solution as well as the RANS-biogeochemical DO equations for the predicted pdfs using our modular finite-volume framework (Ueckermann and Lermusiaux, 2012). The physical domain (Section 4.4) is discretized using a uniform finite-volume staggered C-grid, for both the flow and stochastic biogeochemical fields. The size of finite volumes in each $x-$ and $z-$ direction is equal to $\Delta x = \frac{1}{15}$ and $\Delta z = \frac{1}{15}$ (non-dimensional) respectively, thus, a grid-size of $300 \times 30$. Advection is computed explicitly, using a total variation diminishing (TVD) scheme with a monotonized flux limiter (Van Leer, 1977). Diffusion is treated implicitly, with a second-order central difference scheme. All the reaction terms are computed explicitly. To handle the complex boundaries with the structured Cartesian grid, a ghost cell immersed boundary method is adopted for accurate enforcement of the boundary conditions (Gupta, 2022). For time-marching of the PDEs (RANS, DO mean, and DO modes), we use a first-order forward Euler method, while for the stochastic DO coefficient ODEs, we use a four-stage Runge–Kutta scheme. A non-dimensional time-step of $\Delta t = \frac{1}{240}$ is used in all the experiments. It is also ensured that we satisfy the Courant–Friedrichs–Lewy (CFL) condition at all times. We refer to Ueckermann et al. (2013) and Feppon and Lermusiaux (2018a) for more details on the numerical schemes we employ.

### 4.6. Stochastic balanced initialization: Parameters, state variable fields, and probabilities

The values of the parameters for the physics are chosen such that the flow emulates some coastal ocean dynamics. The dimensional barotropic velocity at the inlet is chosen to be $U \approx 10^{-2}$ to $10^{-1}$ m/s. The subgridscale eddy-viscosity is $\nu_E \approx 0.01$ to $0.5$ m$^2$/s. Considering

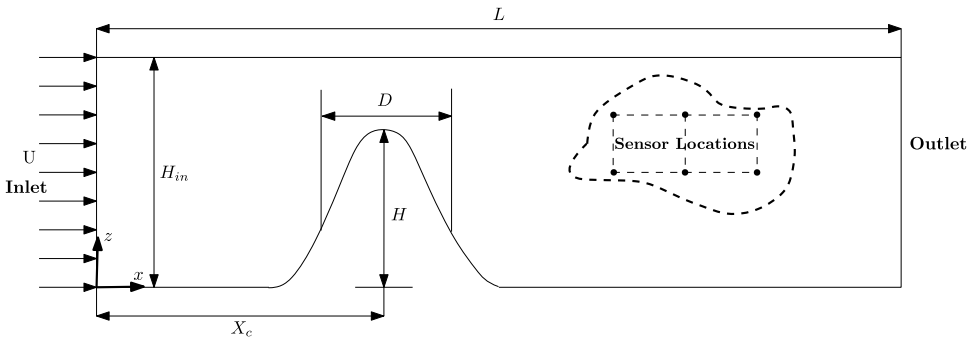



**Fig. 2.** Two-dimensional spatial domain of the flow past a ridge. The ridge is defined by $He^{-(x-X_c)^2/D^2}$, where $D$ is the characteristic width, $H$ the height, and $X_c$ the distance between the inlet and the center of the ridge. Noisy and sparse observations of state variables are collected downstream of the ridge (see example sensor locations inset). The exact observation locations and the state variables being measured vary with the particular experiment.

the vertical length scale of $H \approx 50$ m for the ridge, we obtain an eddy-viscosity Reynolds number of $Re = \frac{UH}{\nu_E} \approx 1$ to 500. Further, we do not consider any wind-forcing explicitly. For the initial velocity, we use a divergence free velocity field that satisfies the inlet and outlet boundary conditions, and so mass conservation in the given domain. The pressure field is initialized to be zero throughout the domain.

The biological parameters are either deterministic or stochastic. The values of the deterministic parameters are kept fixed for every realization. The stochastic parameters are sampled from their respective initial pdf or joint pdfs, if available; when measurements are made, these pdfs are updated along with all other state variables, using the GMM-DO filter. The stochastic parameters are divided into two categories, regular ones that were originally present in the biogeochemical models and have biological origins, and new parameters introduced for the unification of candidate models and parameterizations of unknown functions. The deterministic values and initial pdfs of the biological parameters used in the main experiments are given in Table 1. The initial pdfs of all the stochastic parameters are set to be uniform and independent of each other unless otherwise specified. Several but not all regular parameters are assumed stochastic, aiming to showcase parameter uncertainties that are commonly significant or of different mathematical types (e.g., nonlinear functions) or learning types (e.g., parameters in equations not directly related to the sparse measurements). In the experiments presented in this paper, advection-reaction dominates and the eddy-diffusivity for the biological tracers can be taken as negligible, $\mathcal{K}_E \approx 0$, such that the eddy-diffusivity Peclet number $Pe = \frac{UH}{\mathcal{K}_E} \to \infty$. Other experiments (not shown) were also successful however with non-negligible diffusivity, e.g. Lu and Lermusiaux (2021). In all our simulations, a biological time-scale of the order of 1 day is used for all non-dimensionalization purposes.

Following Lermusiaux et al. (2000), Lermusiaux (2002), Beşiktepe et al. (2003), Lermusiaux (2007) and Lermusiaux et al. (2011), in all the subsequent experiments, biogeochemical fields are initialized in approximate dynamical balance, in accord with their stochastic model PDEs (20) and their sampled parameter values and local total biomass field. Specifically, the initial concentration fields for every sampled realization are obtained by finding a biogeochemical equilibrium solution corresponding to their sampled parameter values and vertical biomass profile. These equilibrium fields are found by solving the ODE nonlinear biogeochemical models of Section 4.1 at all depths. Dynamical equilibrium for initialization is reached when temporal variations become negligible, or when the system reaches a limit cycle or time-variations that are without unrealistic transients (Lermusiaux, 2002; Haley et al., 2015). However, in the current setup, the approximate equilibrium solutions at each spatial location are found by using MATLAB's *fsolve* (MathWorks, 2023) to find the roots of the nonlinear system obtained with time derivatives set to zero (Newberger et al., 2003). Further, we also impose the initial total biomass profile, $\sum_{i=1}^{N_\phi} \phi^i(z;\omega) = T_{bio}(z)$, to be given, with $T_{bio}$ to be linearly increasing from 10 mmol N m$^{-3}$ at the surface to 30 mmol N m$^{-3}$ at the depth of 100 m, for all the biogeochemical models. This depth-dependent equilibrium solution for each realization of the biogeochemical state variables is used to initialize the corresponding fields in space, with the ridge masked at every $x$ location. We also ensure that none of the realizations of the stochastic parameters lead to nonphysical equilibrium solutions, such as negative tracer values. The value of 30 mmol N m$^{-3}$ is used to non-dimensionalize all the biogeochemical fields and parameter values. For the non-dimensionalization of parameters, when needed, we additionally use a length-scale of 50 m (the height $H$ of the ridge) and a time-scale of 1 day.

To initialize the DO decomposition of the biogeochemical fields, after generating the initial fields for each realization, we compute their statistical average and use it to initialize the mean biogeochemical fields. To initialize the DO modes and stochastic coefficients, we take the singular value decomposition (SVD) of the ensemble of mean-removed concatenated fields, keeping the dominant singular values and vectors. We account for the differences in the magnitude of the variability of individual biogeochemical tracers before taking the SVD, by appropriate normalization based on their standard deviations (Appendix B).

### 4.7. True solution generation

In the present work, twin experiments (Bengtsson et al., 1981; Ide and Ghil, 1998a,b; Lermusiaux, 1999a) are conducted and the noisy observations are extracted from a simulated truth. To obtain the simulated truth fields for each experiment, a set of parameters and initial state fields are sampled. Starting from these initial conditions, the Navier–Stokes PDEs (19) and the deterministic version of the ADR PDEs (20) with the true biogeochemical model are numerically integrated. The result is the simulated truth solution. In each experiment, all the remaining deterministic parameters, modeling domain, and numerical schemes are as these of the stochastic simulation using the DO equations.

### 4.8. Observations and inference

Noisy, sparse, and indirect observations are taken from the simulated true solution (Section 4.7). The observation locations are kept in or near the euphotic zone because deeper depths have limited biological variability (Fig. 2). In each experiment, only one of the biological tracer fields is observed at 6 to 9 locations. What is measured thus varies from experiment to experiment, as is common in real sea experiments. The observation schedule is also experiment dependent, however, it is not more frequent than once every non-dimensional

time: data are collected at far-apart discrete time-instants, $t_k$ for $k = 1, 2, \ldots, K$. This use of a few, infrequent, and indirect observations allows us to showcase the capabilities of multivariate GMM-DO Bayesian learning: it can use noisy and sparse measurements of only one state variable to jointly update all pdfs, and thus indirectly estimate multiple state variable fields, parameters, and model functions.

The linear observation matrix $\boldsymbol{H}$ (Eq. (3)) is specified such that it predicts the concentration of the observed tracer field at the observation locations by interpolating the concatenated state fields at the observation locations. The observation error standard deviation matrix ($\sqrt{\boldsymbol{R}}$ in Eq. (3)) is assumed diagonal. It models both sensor noise and representation errors (Janjić et al., 2018). The latter representation errors include subgrid-scale processes and variability in the data not simulated by the dynamical model; the representation errors are thus often much larger than sensor errors. As a result, in our experiments, the observation error standard deviation values contained in $\sqrt{\boldsymbol{R}}$ are set to be a fraction of the local variability in the state variables (Lermusiaux et al., 2000; Lermusiaux, 2002). In most of our at-sea experiments, observation errors are significantly smaller than errors in biogeochemical model equations and variability of their fields, and the relative observation error is here set at five percent. We note that as long as these standard deviations and the actual simulated observation noise are consistent and not larger than the natural variability, their values mainly affect the learning rate achieved by data assimilation as well as the amount of uncertainty remaining in the learned states and parameters (i.e., smaller observation errors accelerate the learning and reduce the uncertainty remaining at the end).

Further, the hyper-parameters related to the DO equations and the GMM-DO filter were chosen based on numerical tests and experience (Sondergaard and Lermusiaux, 2013b; Lolla et al., 2014; Lu and Lermusiaux, 2014; Gupta et al., 2019), for each of the experiments. For the DO equations, the number of modes, number of Monte-Carlo coefficient samples, time-step, etc., were selected so as to be sufficient to capture the dominant uncertainty and evolving probability distribution for each of the state vector fields, parameters, and model equations themselves. For Bayesian learning with the GMM-DO filter, the expectation–maximization (EM) algorithm (Bilmes et al., 1998) and Bayesian Information Criterion (BIC) (Stoica and Selen, 2004; Wornell, 2016) were employed to select the optimal number of GMM components at each data time. Typical BIC-optimized values for $N_{\text{GMM}}$ were found to be 10 for the present experiments.

### 4.9. Learning metrics

We evaluate the performance of our Bayesian learning framework by comparing the learned solution with the true solution from which noisy observations were collected and by examining the posterior joint state-parameter-model probability distributions. For the former solution evaluations, we compare the true fields to the DO mean fields, and the true parameter values to the most probable DO pdf values of the parameters. To quantify performance, we examine the evolution of the Root Mean Square Error (RMSE) of the biogeochemical tracer fields, uncertain regular parameters $\boldsymbol{\theta}(\omega)$, and novel formulation $\boldsymbol{\alpha}(\omega)$, complexity $\boldsymbol{\beta}(\omega)$, and/or expansion $\boldsymbol{\gamma}(\omega)$ parameters. The RMSE between an evolved stochastic state field/parameter estimate $\phi(\boldsymbol{x}, t; \omega)$ and its corresponding true field/parameter $\phi^{true}(\boldsymbol{x}, t)$, is given by, $\sqrt{\frac{1}{|D|}\int_D \mathbb{E}[(\phi(\boldsymbol{x}, t; \omega) - \phi^{true}(\boldsymbol{x}, t))^2] d\boldsymbol{x}}$. The square of RMSE hence consists of two contributions (Lin, 2020), one is the square of the $L_2$ distance between the mean of the variable in the stochastic run and the simulated truth, while the other is the variance of the variable. In every experiment, the RMSE values of each variable are normalized by the corresponding RMSE value just before the first assimilation step. For the latter pdf evaluations, we analyze the evolution of the posterior pdfs of the stochastic DO coefficients, and of the regular and new stochastic parameters. For example, for the DO coefficient realizations, we employ 2-D scatter plots. For the stochastic parameters, we use marginals and kernel-density fits. We also evaluate the convergence of pdf estimates with stochastic resolution, i.e. increasing/decreasing stochastic numerical parameters ($N_s$, $N_r$, etc.), see Section 3.3.

## 5. Application results and discussion

In order to demonstrate the capabilities of our Bayesian learning we utilize four sets of twin experiments with different coupled biogeochemical–physical dynamics and learning objectives. We perform simultaneous Bayesian estimation of state variables, parameters, and model equations, using noisy observations that are sparse in both space and time. To quantify performance, we evaluate several learning metrics, emphasizing the sharpness of the inference and the accuracy of probability distributions. For each of the four sets of experiments, we conduct multiple studies so as to evaluate the sensitivity to hyper-parameters. However, for each set, we present detailed results for only one experiment and summarize the other results and sensitivity studies. The main parameters of the physical–biogeochemical models, the hyper-parameters for the DO equations, and the observation and assimilation parameters are provided in Table 1.

### 5.1. Experiments 1: Discriminating among candidate functional forms

Biologically, mortality is a linear rate process. The mortality terms of phytoplankton and zooplankton however commonly act as "closure" parameterizations in models because as they allow for recycling of nutrients directly from plankton. As a result, due to the missing intermediate states in the recycling model, the zooplankton mortality and recycling processes are often modeled nonlinearly, with a concentration-dependent loss rate (Franks, 2002), which allows representing unresolved processes such as predation by larger predators, diseases, etc. In this first set of experiments, we use the NPZ model with uncertainty introduced by the ambiguity in the presence or absence of a quadratic zooplankton mortality function, along with the uncertainty in the value of the Ivlev grazing parameter ($\Lambda$). This ambiguity in the zooplankton mortality function corresponds to the $\hat{\mathcal{L}}$ term introduced in Eq. (1). Uncertainties in the initial biogeochemical conditions are set in balance with the uncertain parameters and model equations, as explained in Section 4.6. The learning objective is to simultaneously learn all the biological states, regular parameter $\Lambda$, and functional form of zooplankton mortality using a stochastic formulation parameter, by assimilating sparse noisy observations.

The right-hand-side of the NPZ model (Eq. (15)) with the quadratic zooplankton mortality are the reaction or source terms in the ADR PDEs (20). These NPZ source terms $S^{\phi^i}$ are given by,

$$
\begin{aligned}
S^N &= -G\frac{PN}{N+K_u} + \Xi P + \Gamma Z + \underbrace{\alpha(\omega)(\tilde{\Gamma} Z^2)}_{\text{Quad. Z Mort.}} + R_m \gamma Z(1 - \exp^{-\Lambda(\omega)P}) \\
S^P &= G\frac{PN}{N+K_u} - \Xi P - R_m Z(1 - \exp^{-\Lambda(\omega)P}) \\
S^Z &= R_m(1-\gamma)Z(1 - \exp^{-\Lambda(\omega)P}) - \Gamma Z - \underbrace{\alpha(\omega)(\tilde{\Gamma} Z^2)}_{\text{Quad. Z Mort.}} .
\end{aligned}
\tag{29}
$$

The stochastic parameters are explicitly shown using the realization index ($\omega$), and the ambiguous quadratic mortality term is pointed out. The stochastic formulation parameter, $\alpha(\omega)$, is restricted to binary values, i.e., 0 or 1, corresponding to the absence or presence of the quadratic mortality term, respectively. $\Lambda(\omega)$ is sampled from a uniform probability distribution between the non-dimensional values of 3 and 6, and $\alpha(\omega)$ is assumed to have an initial 50%–50% probability of being 0 or 1. The stochastic ADR PDEs with the above stochastic NPZ reactions (Eq. (29)) are coupled with the RANS flow PDEs, and solved with the DO methodology (Sections 4.3–4.5). The other known physical–biogeochemical model parameters, the hyper-parameters for the DO

**Table 1**
Deterministic or pdf values of the various domain-related, biological, physical, and hyper-parameters used in the four sets of experiments. The values $H = 50$ m, $max\{T_{bio}(z)\} = 30$ mmol N m$^{-3}$, and time-scale of 1 day are the characteristic scales used for non-dimensionalization.

| Parameters | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 |
|---|---|---|---|---|
| Biogeochemical model | NPZ | NPZ & NPZD | NPZ | NNPZD |
| Biological parameters | | | | |
| Light attenuation due to sea water, $k_w$ (m$^{-1}$) | 0.067 | 0.067 | 0.067 | 0.067 |
| Initial slope of the P-I curve, $\alpha$ ((W m$^{-2}$ day)$^{-1}$) | 0.025 | 0.025 | 0.025 | 0.025 |
| Surface photosynthetically available radiation, $I_o$ (W m$^{-2}$) | 158.075 | 158.075 | 158.075 | 158.075 |
| Phytoplankton maximum uptake rate, $V_m$ (day$^{-1}$) | 1.5 | 1.5 | 1.5 | 1.5 |
| Half-saturation for phytoplankton uptake of nutrients, $K_u^*$ (mmol N m$^{-3}$) | 1 | 1 | 1 | 1 |
| $NH_4$ inhibition parameter, $\Psi_I$ ((mmol N m$^{-3}$)$^{-1}$) | – | – | – | 1.46 |
| $NH_4$ oxidation coefficient, $\Omega$ (day$^{-1}$) | – | – | – | 0.25 |
| Phytoplankton specific mortality rate, $\Xi$ (day$^{-1}$) | 0.1 | 0.1 | 0.1 | unif(0.01, 0.08) |
| Zooplankton specific excretion and mortality rate, $\Gamma$ (day$^{-1}$) | 0.145 | 0.145 | 0.145 | unif(0.125, 0.150) |
| Presence/absence of quadratic zooplankton term, $\alpha$ | unif{0, 1} | unif{0, 1} | – | unif{0, 1} |
| Quadratic zooplankton specific excretion and mortality rate, $\tilde{\Gamma}$ (day$^{-1}$) | 0.2 | 0.2 | 0.2 | 0.2 |
| Zooplankton maximum grazing rate, $R_m$ (day$^{-1}$) | 0.52 | 0.52 | 0.52 | unif(0.52, 0.72) |
| Ivlev constant, $\Lambda$ ((mmol N m$^{-3}$)$^{-1}$) | unif(0.1, 0.2) | unif(0.1, 0.2) | 0.13 | unif(0.052, 0.072) |
| Fraction of zooplankton grazing egested, $\gamma$ | 0.3 | 0.3 | 0.2 | 0.3 |
| Detritus decomposition rate, $\Phi$ (day$^{-1}$) | 1.03 | 1.03 | 1.03 | 1.03 |
| Diffusion constants – horizontal & vertical, ($\mathcal{K}_E$) | 0 | 0 | 0 | 0 |
| Modeling domain | | | | |
| Height of the ridge, $H$ (m) | 50 | 50 | 50 | 50 |
| Characteristic width of the ridge, $D$ (km) | 1 | 1 | 1 | 1 |
| Distance between inlet and center of ridge, $X_c$ (km) | 7.5 | 7.5 | 7.5 | 7.5 |
| Domain height, $H_{in}$ (m) | 100 | 100 | 100 | 100 |
| Domain length, $L$ (km) | 20 | 20 | 20 | 20 |
| Physical parameters | | | | |
| Inverse of Eddy-viscosity Reynolds nb., ($\Lambda_{Re}$) | 1 | 1 | 1 | 1/500 |
| DO parameters | | | | |
| Number of Modes, $N_s$ | 20 | 40 | 20 | 15 |
| Number of Monte-Carlo samples, $N_r$ | 10,000 | 10,000 | 1000 | 10,000 |
| GMM-DO Filtering – Observation and assimilation parameters | | | | |
| State variables being observed | $Z$ | $Z$ | $N$ | $P$ |
| Observation error standard deviation, ($\sqrt{\boldsymbol{R}}$) (non-dim.) | 0.05 | 0.05 | 0.035 | 0.04 |
| Size of Observation vector, $N_y$ | 6 | 6 | 8 | 9 |
| Observation start time (non-dim.) | 5 | 5 | 1 | 2 |
| Time interval between assimilations (non-dim.) | 2 | 2 | 2 | 1 |
| Observation end time (non-dim.) | 25 | 25 | 25 | 25 |

equations, and the observation and assimilation parameters are given in Table 1.

*True solution generation:* The true solution corresponds to the non-dimensional values, 3.6 for $\Lambda$, and 1 for $\alpha$, i.e., the quadratic mortality term present. The true state fields are initialized and evolved as described in Section 4.7, and noisy observations are extracted from this simulation.

*Observations and learning parameters:* The simulated observations are sparse in both space and time. They consist of noisy zooplankton measurements at six locations downstream of the ridge, only at every two non-dimensional times, starting at $t = 5$. The non-dimensional $Z$-data error standard deviation is 0.05. The data shown in Fig. 3 is all that the Bayesian learning framework gets to assimilate over the course of the experiment. Other hyper-parameters related to the GMM-DO filtering are provided in Table 1.

*Learning metrics:* As time advances, the sparse noisy data are assimilated using the Bayesian GMM-DO filter in the augmented state space. We compare the true fields and paramaters to their DO estimates (mean and most probable values). To quantify performance, we examine the evolution of the normalized RMSEs (Section 4.9) for the N, P, and Z fields, and for the $\Lambda(\omega)$ and $\alpha(\omega)$ parameters, as well as the pdfs of the stochastic parameters, DO coefficients, and biological states.

#### 5.1.1. Learning results

Fig. 4 shows the initial state and parameters of the system (at $t = 0$), while Fig. 5 shows the evolved prior state and parameters of the system at $t = 5$ (i.e. just before the 1st observational episode). There are significant differences between the true and prior DO mean fields of the biogeochemical tracers. During these first five non-dimensional time units, a phytoplankton bloom develops just downstream (top-right) of the ridge: upwelling of nutrients above the ridge within the euphotic zone feeds the growth in phytoplankton biomass in the wake.

In Fig. 6, we illustrate the evolving statistics of the stochastic dynamical system from $t = 0$ to $t = 5$ just before data assimilation. We show fields of the phytoplankton standard deviation and dominant three DO modes (Panels 6(a) & 6(b)). The standard deviation fields clearly highlight the significant uncertainty around the phytoplankton subsurface maxima and bloom, reaching 30 percent of the mean field maxima. The uncertain subsurface maxima and bloom also clearly affect the DO modes. In Panels 6(c) & 6(d), we show the joint distribution of the top four stochastic coefficients, $Y_i(t;\omega)$ with $i = 1, 2, 3, 4$ in (Eq. (22)) at $t = 0$ and $t = 5$, respectively. In Panel 6(d), we also show the corresponding prior GMM fits at $t = 5$, using $N_{\text{GMM}} = 10$ components (Eq. (C.1) in Appendix C). We use the BIC to find this optimal number of components required (Sondergaard and Lermusiaux, 2013a). The marginalized distributions shown are those of the joint-double and single coefficient spaces for $i = 1, 2, 3, 4$. They demonstrate the highly non-Gaussian nature of the stochastic DO coefficients, which the DO equations evolve, and the GMM-DO filter account for. The strong parametric uncertainties are reflected by the thin joint-double coefficient distributions. In addition, the realizations of the stochastic coefficients are clearly divided into two groups, corresponding to the presence or absence of the quadratic mortality term. This is already the

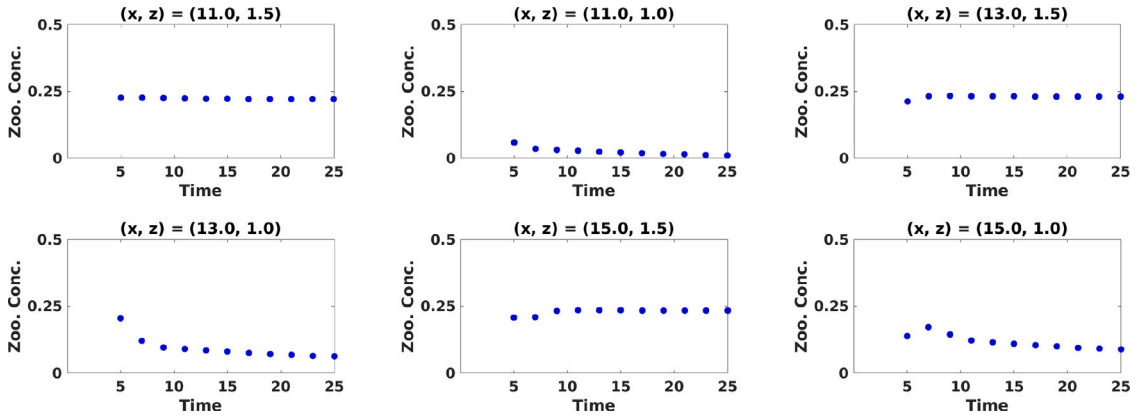



**Fig. 3.** Experiments-1. Time series of non-dimensional zooplankton concentration collected at six observation locations (their coordinates are given in the respective titles). For all experiments-1, the non-dimensional data error standard deviation is 0.05 (see Table 1).

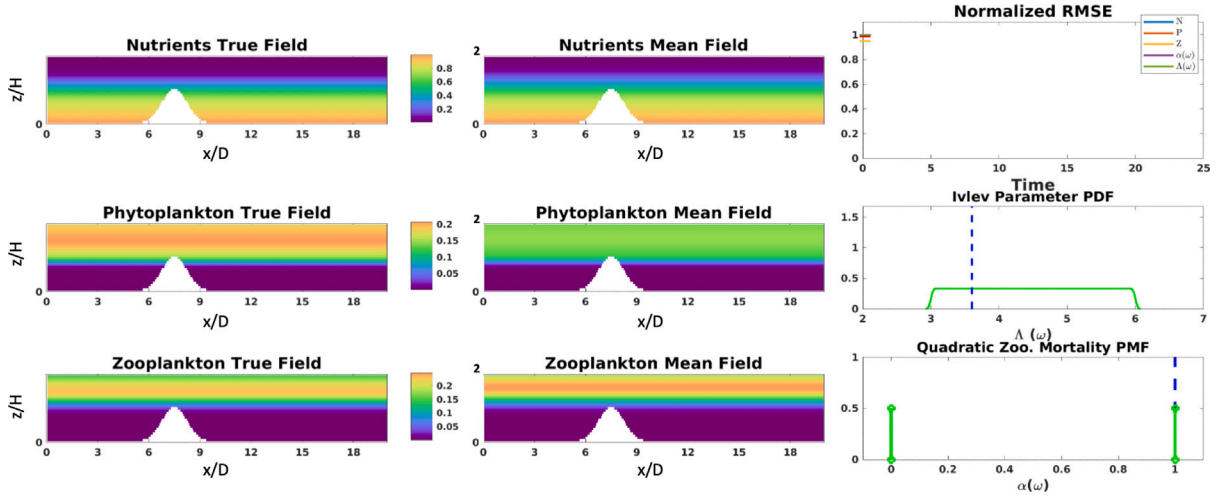



**Fig. 4.** Experiments-1: State of the true and estimate NPZ fields and parameters at $t = 0$ (i.e. initial conditions). The first two columns consist of the non-dimensionalized true (left) and mean estimate (right) tracer fields of N, P, and Z. In the third column, the top panel shows the variation of normalized root-mean-square-error (RMSE) with time for the stochastic state variables and parameters. We explicitly mark the RMSE values at $t = 0$ using short horizontal lines for better visibility. All the RMSE values are concentrated around 1, and for some, the lines overlap. The next two panels contain the pdfs of the non-dimensional $\Lambda(\omega)$ and $a(\omega)$ (to learn the presence or absence of quadratic zooplankton mortality), each marked with solid green lines, with the true unknown parameter values marked with blue dotted lines. The velocity field is deterministic with $Re = 1$.

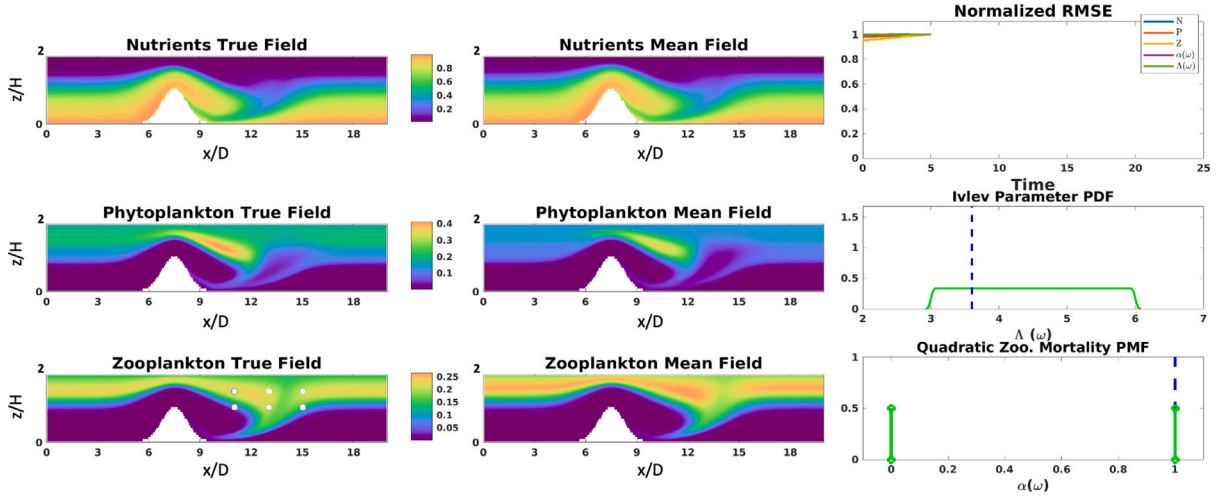
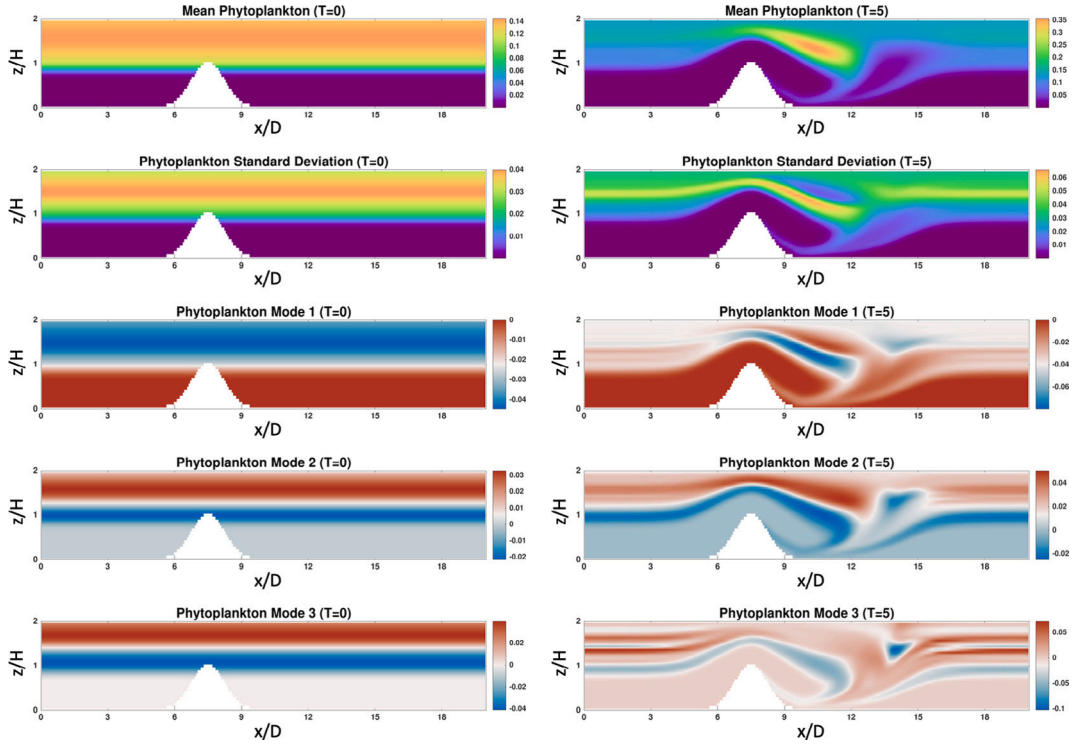
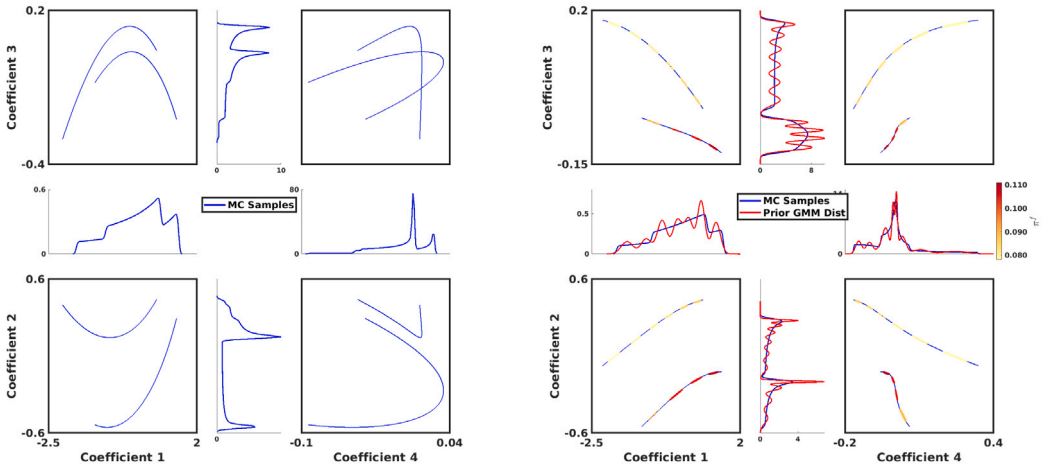



**Fig. 5.** Experiments-1: As Fig. 4, but for the prior fields and parameters at $t = 5$ (i.e. just before the 1st assimilation). The white circles on the zooplankton true field mark the six observation locations. In the first two columns, the axis limits for the state variables have changed so as to follow the bloom, but in the third column, they remain as in Fig. 4 so as to directly highlight the uncertainty evolution.




**(a)** Phytoplankton mean, standard deviation and top three DO modes, at $t = 0$.

**(b)** Phytoplankton mean, standard deviation and top three DO modes, at $t = 5$ (prior).




**(c)** Joint distributions and respective marginals of the top four stochastic DO coefficients, at $t = 0$.

**(d)** Joint distributions and respective marginals of the top four stochastic DO coefficients, along with the GMM fit, at $t = 5$ (prior). In the joint distribution plots, one standard deviation contours of each member of the GMM are marked with solid-line ovals (at the 1-sigma level) colored according to their respective weights (colorbar to the right).

**Fig. 6.** Experiments-1: Statistics for the initial ($t = 0$) and prior ($t = 5$, just before the 1st assimilation) states of the stochastic NPZ ADR dynamical system.

case at $t = 0$ (Panel 6(c)) because our uncertainty initialization scheme (Section 4.6) respects the stochastic dynamical balances.

At $t = 5$, the first sparse noisy data are assimilated. Fig. 7 shows the posterior mean fields, prior and posterior parametric distributions, and the normalized RMSE values for the mean fields and two stochastic parameters. By only observing noisy zooplankton at six locations, the GMM-DO filter simultaneously updates all the biological fields and parameters. This is evident from the mean fields getting aligned with the true fields and quantified by the RMSE reductions of about 20 to 30 percent. Also visible is the slight change in the pdf for $\Lambda(\omega)$ and a higher probability value for $\alpha(\omega)$ being one. The six data are so far much more informative about the mortality term than the Ivlev parameter.

Next, in Fig. 8, we illustrate the same posterior mean fields, prior and posterior parameters, and normalized RMSE values, but at $t = 15$, i.e., at the sixth data assimilation. The flow is fully developed with the biogeochemical fields well learned, as quantified by the normalized

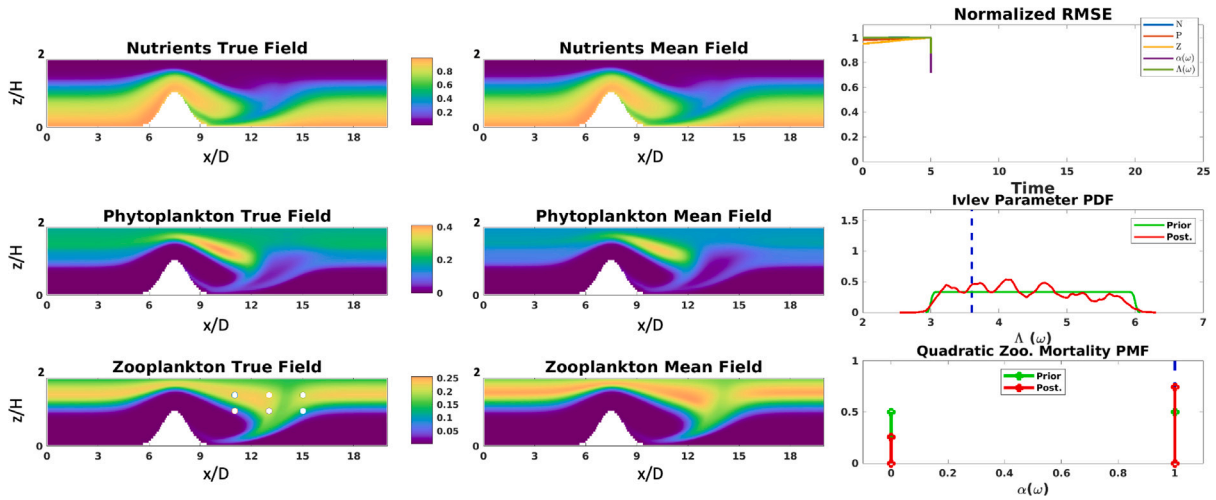



**Fig. 7.** Experiments-1: As Figs. 4 & 5, but for posterior fields and parameters at $t = 5$ (i.e. just after the 1st assimilation). In the last two panels of the third column, the prior pdfs associated with the non-dimensional $\Lambda(\omega)$ and $a(\omega)$ at $t = 5$ are marked with solid green lines, while the posterior pdfs are marked with solid red lines.

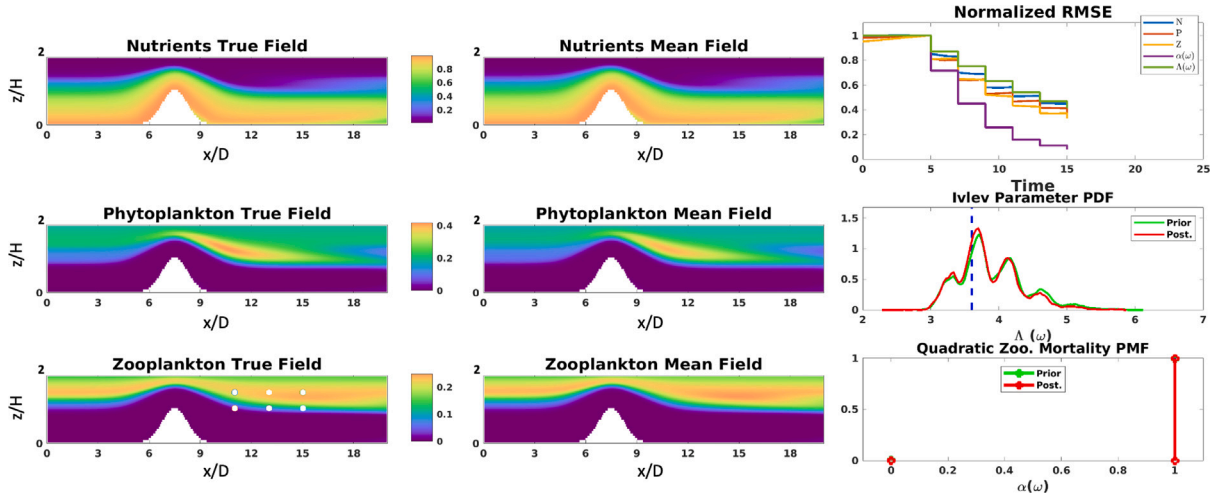



**Fig. 8.** Experiments-1: As Figs. 4, 5 & 7 but for posterior fields and parameters at $t = 15$ (i.e. just after the 6th assimilation). In the first two columns, the axis limits for the state variables have changed so as to follow the bloom, but in the third column, they remain as in Fig. 4 so as to directly highlight the uncertainty evolution.

RMSEs. The GMM-DO filter unambiguously detects the presence of quadratic mortality of $Z$, as confirmed by the RMSE of $\alpha(\omega)$ at 10 percent. The other RMSEs are higher at about 40–50 percent, with that of $Z$ a bit lower (reflecting that $Z$ is the only variable measured at six locations). The pdf of $\Lambda(\omega)$ is also accumulated around its true value, but is now clearly multi-modal, indicating nonlinearities and remaining ambiguity. In other words, with the sparse noisy data assimilated so far, several values of the Ivlev constant $\Lambda(\omega)$ remain very likely, indicating possible biases and equifinality (Duda et al., 2006; Lu and Lermusiaux, 2021).

Finally, at $t = 25$, after 11 assimilation events, the same quantities are shown in Fig. 9. All the biogeochemical mean and true fields match with each other with RMSEs around 20 percent or less. The probability of the presence of the quadratic mortality term is now almost one, while the $\Lambda(\omega)$ pdf has a clear peak near 3.6 with a couple other much lower biased peaks around it. In general, the presence of lower peaks in pdfs of parameters indicate alternative combinations that could explain the data, and also the ability of the GMM-DO filter to capture non-Gaussian pdfs. The learning is also evident from the sustained decrease in the normalized RMSEs at every assimilation step for all the biogeochemical fields and parameters.

*Sensitivity Studies.* Many similar experiments were completed, changing various hyperparameters related to the GMM-DO filter, such as the biological variable being observed, observation locations, frequency, start-time, etc. Noisy observations from simulated truths with different combinations of $\Lambda(\omega)$ and $\alpha(\omega)$ were also used. We found that the biological variable being observed has an impact on the sharpness of the inference or learnability of the given learning objectives. For example, observing $N$ led to the learning of two distinct combinations of $\Lambda(\omega)\&\alpha(\omega)$, 3.1 & 0, and 3.6 & 1, respectively with nearly equal amount of confidence (Gupta, 2016). We also varied the location of observations (e.g., surface versus subsurface) and confirmed that more informative locations (Lermusiaux et al., 2017b) improve the learning rate. Decreasing the amount of observation data, or increasing the value of the observation error standard deviation, led to a slower learning rate and larger uncertainty in the learned states and parameters. We also confirmed the convergence of our GMM-DO Bayesian posteriors by repeating learning experiments with an increasing number of DO modes and coefficients (not shown), until the results converged to those shown. This convergence of the pdfs of the parameters and DO coefficients, and of the DO modes and mean, suggests that our Bayesian GMM-DO filter provides accurate pdf estimates: it thus shows what has been learned without ambiguity or with some ambiguity remaining.

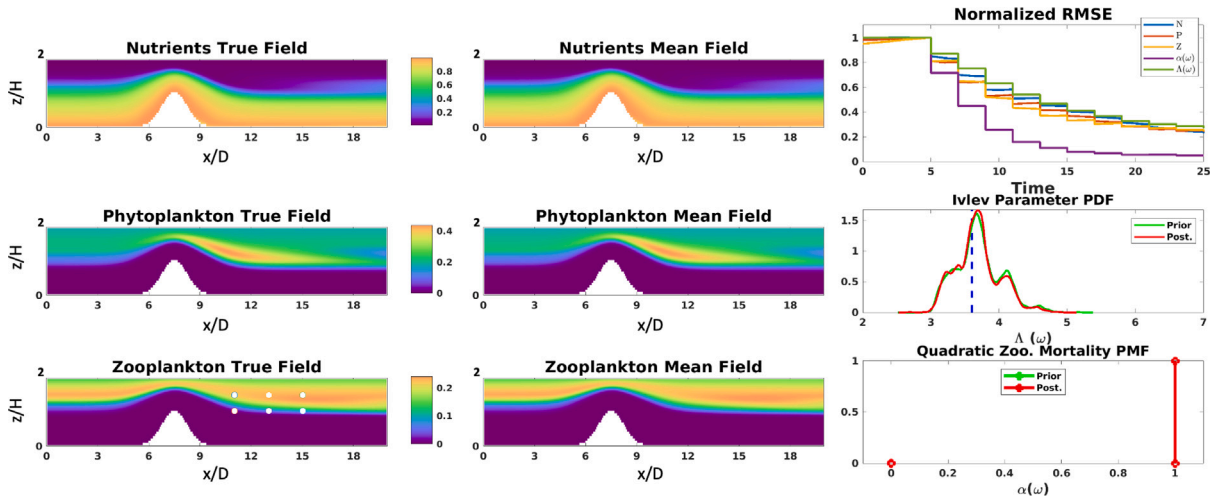



**Fig. 9.** Experiments-1: As Figs. 4, 5, 7 & 8 but for posterior fields and parameters at $t = 25$ (i.e. just after the 11th assimilation). In the first two columns, the axis limits for the state variables have changed so as to follow the bloom, but in the third column, they remain as in Fig. 4 so as to directly highlight the uncertainty evolution.

For the latter case, the multi-model posterior pdfs show that additional observations are needed to sharpen the inference further.

### 5.2. Experiments 2: Discriminating among models of different complexities

In the second set of experiments, the primary goal is to learn the complexity of the biogeochemical model, e.g., its state variables, along with the biogeochemical fields and Ivlev grazing parameter. Two candidate hierarchical model classes, NPZ and NPZD, are considered possible. These NPZ and NPZD models correspond to the candidates $\mathcal{M}_i$'s introduced in Eq. (2). To represent them with a single modeling system, we embed the former into the latter using the stochastic complexity parameter, $\beta(\omega)$. We multiply the detritus state variable ($D$) and other appropriate terms with $\beta(\omega)$, such that, the value of 1 derives the NPZD model, while the value of 0 derives the NPZ model (see Eq. (7)). Thus, the RHS of the general stochastic model which encompasses both NPZ and NPZD models is given by,

$$\begin{aligned} S^N &= -G\frac{PN}{N+K_u} + \Phi D' + \Gamma Z + (1-\beta(\omega))\Xi P \\ S^P &= G\frac{PN}{N+K_u} - \Xi P - R_m Z(1-\exp^{-\Lambda(\omega)P}) \\ S^Z &= R_m(1-\beta(\omega)\gamma)Z(1-\exp^{-\Lambda(\omega)P}) - \Gamma Z \\ S^{D'} &= \beta(\omega)R_m\gamma Z(1-\exp^{-\Lambda(\omega)P}) + \beta(\omega)\Xi P - \Phi D' \\ D' &= \beta(\omega)D\,, \end{aligned} \tag{30}$$

where $D'$ is the modified detritus state. In Experiments-2, the formulation uncertainty is thus within the class of compatible models $\widehat{\mathcal{L}}$ introduced in Eq. (1) and of the compatible-embedded model type defined by Eq. (6) in Section 3.1. Once again, $\Lambda(\omega)$ is sampled from a uniform probability distribution between the non-dimensional values of 3 and 6, and $\beta(\omega)$ is assumed to have 50%–50% probability of being 0 or 1. The stochastic ADR PDEs with the stochastic NPZD′ reactions (Eq. (30)) are coupled with the RANS flow PDEs, and solved with the DO methodology (Sections 4.3–4.5). The other known physical–biogeochemical parameters as well as the hyper-parameters for the DO equations are given in Table 1.

*True solution generation:* The true solution corresponds to the NPZ model with a non-dimensional value of 3.6 for the $\Lambda$ parameter. The state fields are initialized and evolved as described in Section 4.7.

*Observations and learning parameters:* The simulated observations are sparse in both space and time, and again consist of noisy zooplankton measurements at six locations downstream of the ridge, only at every two non-dimensional times, starting at $t = 5$. The non-dimensional $Z$-data error standard deviation is 0.05. Other hyper-parameters related to the GMM-DO filtering are provided in Table 1.

*Learning metrics:* As time advances and sparse noisy data are assimilated, we compare the true fields and parameters to their DO estimates. To quantify performance, we examine the evolution of the normalized RMSEs of state fields and parameters, pdfs of stochastic parameters, and variances of DO coefficients.

#### 5.2.1. Learning results

Fig. 10 shows the state and parameters of the system at $t = 5$, just before the first observational episode. The most distinctive difference is between the true and mean detritus fields. Since the true model is NPZ, the true detritus field is equal to zero, while the mean detritus field is non-zero because half of the realizations correspond to the NPZD model. The RMSEs of all the variables exactly equal 1, because their respective values just before the first assimilation were used for normalization. The pdf of $\Lambda(\omega)$ is uniform in the main range, and $\beta(\omega)$ has 0.5 probability of being 0 or 1. The variances of the top five modes show a rapid decay with mode number, with the top two variances orders of magnitude larger. The variances of modes 3 and 4 differ initially but become similar over time, indicating a potential cross-over at $t = 5$.

In Fig. 11, we directly show the state of the system at time $t = 25$, after eleven GMM-DO data assimilation (six zooplankton values every two non-dimensional times). We find that our Bayesian learning framework is able to learn the true model to be NPZ, along with the posterior pdf of $\Lambda(\omega)$ concentrated around the true value of 3.6. The mean fields also match the true fields, especially the detritus mean field becoming very close to 0 at all the spatial locations. The RMSEs for all the variables decrease over time, up to about $t = 15$. At that time, the RMSE for the phytoplankton field increases due to a mismatch in the strength of the bloom, thus showing that the zooplankton data are not sufficiently informative for the same. The pdf of $\Lambda(\omega)$ features multiple peaks and thus still indicates that competing hypotheses remain for different pairs of parameter values; this was already the case in the intermediate assimilation steps (not shown). The evolution of the variances of the top five modes shows that these variances can increase and cross-over, for example, lower modes become more important as learning progresses. As the bloom develops, more complex nonlinear dynamics are activated, leading to the growth of some uncertainty modes. Results show that our Bayesian filter captures this as well as biases and non-Gaussian behaviors in the pdfs.

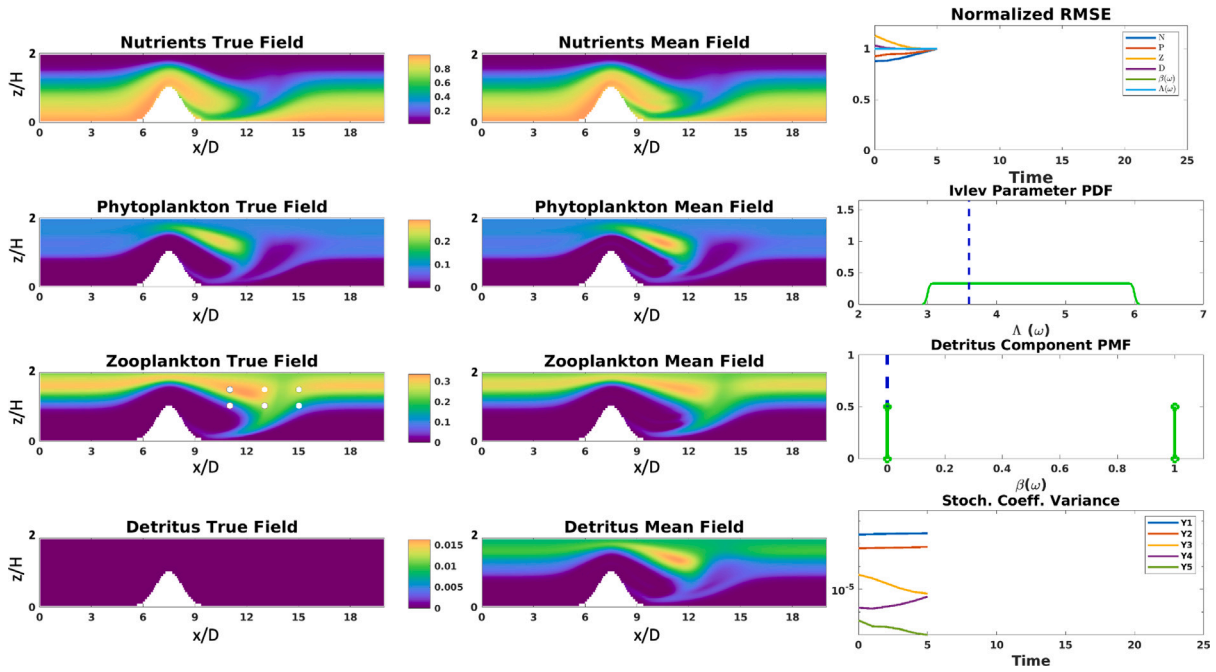



**Fig. 10.** Experiments-2: State of the true and prior estimate NPZD fields and parameters at $t = 5$ (i.e. just before the 1st assimilation). The first two columns consist of the non-dimensionalized true (left) and mean estimate (right) tracer fields of $N$, $P$, $Z$, and $D$. In the third column, the first panel shows the variation of normalized RMSE with time for all the stochastic state variables and parameters. The next two panels contain the pdfs of the non-dimensional $\Lambda(\omega)$ and $\beta(\omega)$ (to learn the complexity, NPZ vs. NPZD), each marked with solid green lines, with the true unknown parameter values marked with blue dotted lines. The last panel shows the evolution with time of the variance (log scale) of the top five modes. The velocity field is deterministic with $Re = 1$. Additionally, the white circles on the zooplankton true field mark the six observation locations.

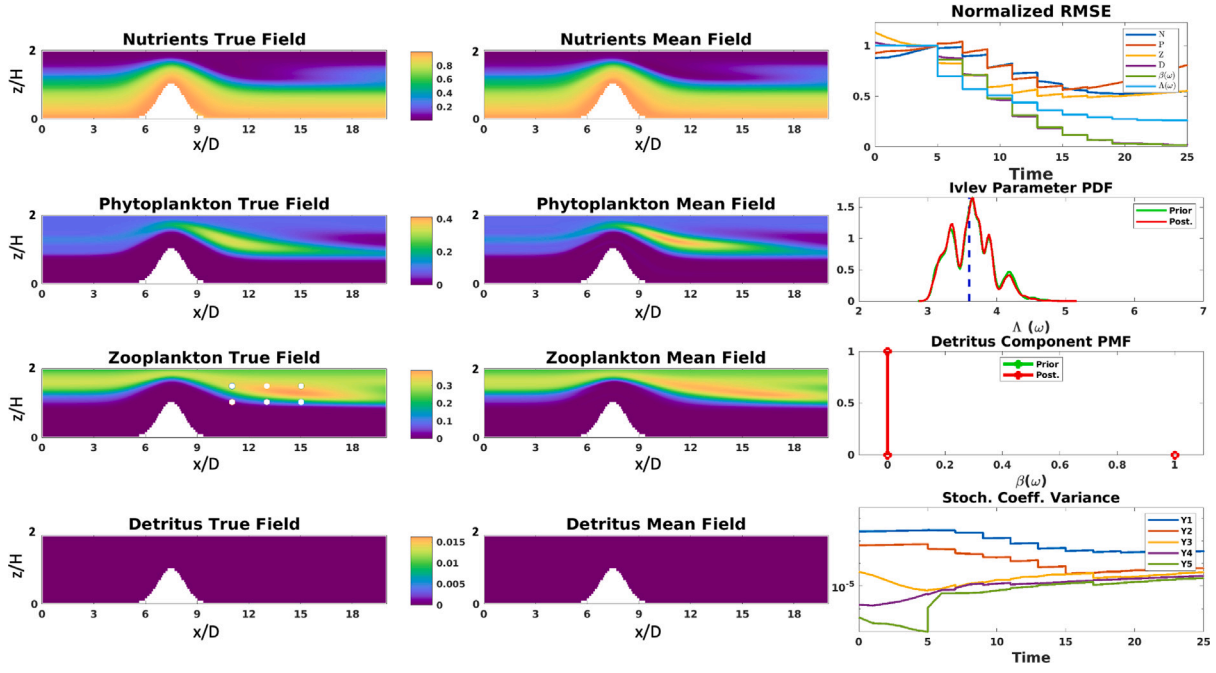



**Fig. 11.** Experiments-2: As Fig. 10 but for posterior fields and parameters at $t = 25$ (i.e. just after the 11th assimilation). In the middle two panels of the third column, the prior pdfs associated with the non-dimensional $\Lambda(\omega)$ and $\beta(\omega)$ at $t = 25$ are marked with solid green lines, while the posterior pdfs are marked with solid red lines. In the first two columns, the axis limits for the state variables have changed so as to follow the bloom, but in the third column, they remain as in Fig. 10 so as to directly highlight the uncertainty evolution.

*Sensitivity Studies.* We performed other experiments with parameter sensitivity studies similar to those of Experiments-1; similar trends were found.

### 5.3. Experiments 3: Learning unknown functional forms

In our third set of experiments, the primary goal is to learn the functional form of the zooplankton mortality without any prior knowledge of candidate forms, along with the uncertain biological tracer fields. This completely unknown zooplankton mortality function corresponds

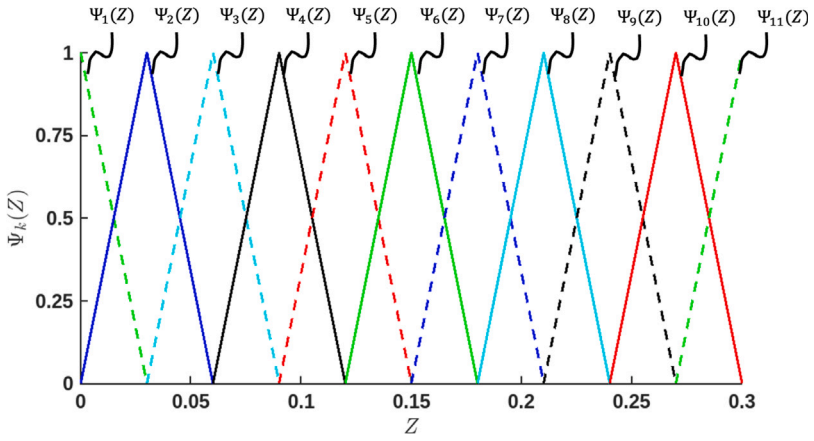



**Fig. 12.** Experiments-3: Linear basis functions, $\{\Psi_1, \ldots, \Psi_{11}\}$ used to represent the unknown zooplankton mortality function, $F(Z;\omega)$, using a rich stochastic space (Eq. (32)).

to the $\widetilde{\mathcal{L}}$ term introduced in Eq. (1). We utilize stochastic piece-wise linear functions to parameterize a large set of possible functional forms within a specified range, as explained in Section 3.2 and Eq. (11). Such a parameterization encompasses many different classes of functions, for example, polynomial, exponential, logarithmic, sinusoidal, etc. The right-hand-side of the stochastic NPZ model with the unknown function is given by,

$$\begin{aligned} S^N &= -G\frac{PN}{N+K_u} + \Xi P + \Gamma Z + \underbrace{F(Z;\omega)}_{\text{Unknown Function}} + R_m\gamma Z(1-\exp^{-\Lambda P}) \\ S^P &= G\frac{PN}{N+K_u} - \Xi P - R_m Z(1-\exp^{-\Lambda P}) \\ S^Z &= R_m(1-\gamma)Z(1-\exp^{-\Lambda P}) - \Gamma Z - \underbrace{F(Z;\omega)}_{\text{Unknown Function}} \end{aligned} \tag{31}$$

From prior knowledge (Newberger et al., 2003), the non-dimensional value of zooplankton is assumed non negative and its maximum value to be 0.3. Thus, $F(Z;\omega)$ is set to be composed of any continuous piece-wise linear segments in the interval $Z \in [0, 0.3]$. Dividing this interval $[0, 0.3]$ into 10 equal non-overlapping sections, such that, $0 = Z_L^1 < Z_R^1 = 0.03 = Z_L^1 < \cdots < Z_R^9 = 0.27 = Z_L^{10} < Z_R^{10} = 0.3$, $F(Z;\omega)$ is thus represented as,

$$F(Z;\omega) = \sum_{k=1}^{11} \gamma_k(\omega)\Psi_k(Z) \tag{32}$$

where the linear basis functions, $\{\Psi_1, \ldots, \Psi_{11}\}$ defined by Eq. (10) are shown in Fig. 12.

Each set of realizations of $\gamma_k$'s, $k \in \{1, \ldots, 11\}$, are sampled so as to avoid a prior with unnatural highly fluctuating functions. The function range is set within 0 and 0.08; it is non-negative as mortality is negative in the zooplankton equation (Eq. (31)). To initialize the tracer fields, we find equilibrium solutions corresponding to each realization of the zooplankton mortality function. The stochastic ADR PDEs with the stochastic NPZ reactions (Eq. (31)) are coupled with the RANS flow PDEs, and solved with the DO methodology (Sections 4.3–4.5). Table 1 provides the values of other known model and hyper-parameters. The learning objective of these experiments is to learn $F(Z;\omega)$ by estimating $\gamma_k's$ along with the biological tracer fields.

*True solution generation:* The true solution contains quadratic zooplankton mortality, with values of the other parameters provided in Table 1.

*Observations and learning parameters:* The simulated observations remain sparse in time and space, but here they consist of the nutrient field at 8 spatial locations, starting at $t = 1$ and occurring every two non-dimensional times. In these experiments, we start the assimilation at the earlier $t = 1$ time in order to limit the exploding growth of uncertainty in the system, because each function realization leads to very different biological dynamics, several of which would lead to nonphysical biological states. The non-dimensional $N$-data error standard deviation is 0.35. Other hyper-parameters related to the GMM-DO filtering are provided in Table 1.

*Learning metrics:* We compare the true fields and parameters to their DO estimates. To quantify performance, we also examine the evolution of the normalized RMSEs and pdf and realizations of the stochastic piece-wise linear functions.

### 5.3.1. Learning results

Fig. 13 illustrates the prior at $t = 1$. Realizations in the space of the unknown function are provided (third column, bottom panel). Each of the function realizations is colored proportional to the joint probability density of the stochastic expansion parameters ($\gamma_k$'s in Eq. (32)). For the prior, $\gamma_k$'s are considered independent of each other and sampled uniformly, hence each piece-wise linear segment is equiprobable. However, the piece-wise linear segments that constitute unnatural highly fluctuating functions are eliminated. This leads to deviations from each function being equally likely and thus their pdf values are not equal, as seen in the third column, bottom panel, of Fig, 13. In general, mortality being 0 for $Z = 0$ is common knowledge. Otherwise, it could act as a sink for zooplankton and lead to negative tracer values. We let this be discovered by the learning algorithm. The DO biogeochemical mean fields are quite far from the unknown true fields, and the prior function realizations are not similar to the true quadratic mortality.

As the eight noisy $N$-observations are assimilated every two non-dimensional times, nearly all the piece-wise linear function realizations converge to the true quadratic mortality. Results after 13 GMM-DO assimilation in Fig. 14 show this. The posterior function realizations (third column) are constructed by sampling the posterior joint probability density of the stochastic expansion parameters ($\gamma_k$'s in Eq. (32)) and colored proportional to the probability value. We find however that the $N$ data are not as informative about mortality function for $Z$ beyond 0.25. This is in part because the maximum value reached in the true $Z$ field is $\sim$ 0.2, which limits the uncertainty reduction in the larger $Z$ regime. The mean fields also converge to the true fields. The normalized RMSEs of biogeochemical fields decrease at each assimilation. The learned phytoplankton mean field however remains a bit higher than true fields, in part because they were much higher initially. It is also because the observed data (here eight $N$ data) are not equally informative about all the learning objectives. As in Lermusiaux et al. (2017c,d) and Lin (2020), this is confirmed by mutual information fields (not shown).

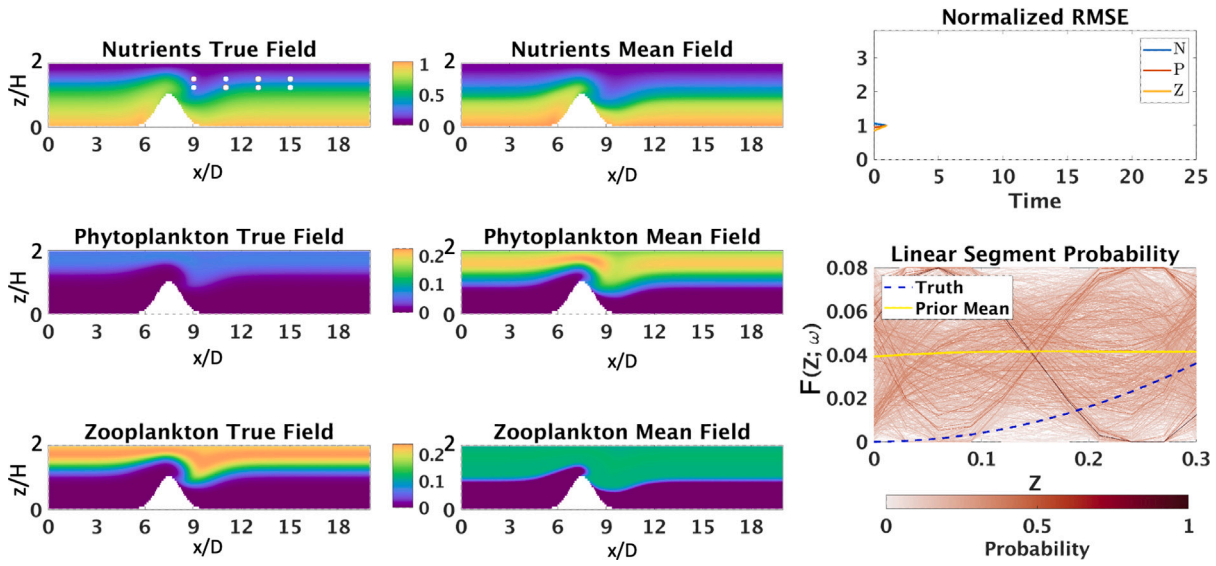



**Fig. 13.** Experiments-3: State of the true and prior estimate NPZ fields and parameters at $t = 1$ (i.e. just before the 1st assimilation). The first two columns consist of the non-dimensionalized true (left) and mean estimate (right) tracer fields of $N$, $P$, and $Z$. In the third column, the first panel shows the evolution of normalized RMSE for all the stochastic state variables. The second panel contains all the realizations of the unknown functional form approximated by piece-wise linear segments. The function realizations are colored according to their respective pdf values (the 0–1 probability colorbar is under the panel). The mean functional form estimate is marked with a solid yellow line, while the true functional form is marked with a dashed blue line. The velocity field is deterministic with $Re = 1$. The white circles on the nutrient true field mark the 8 observation locations.

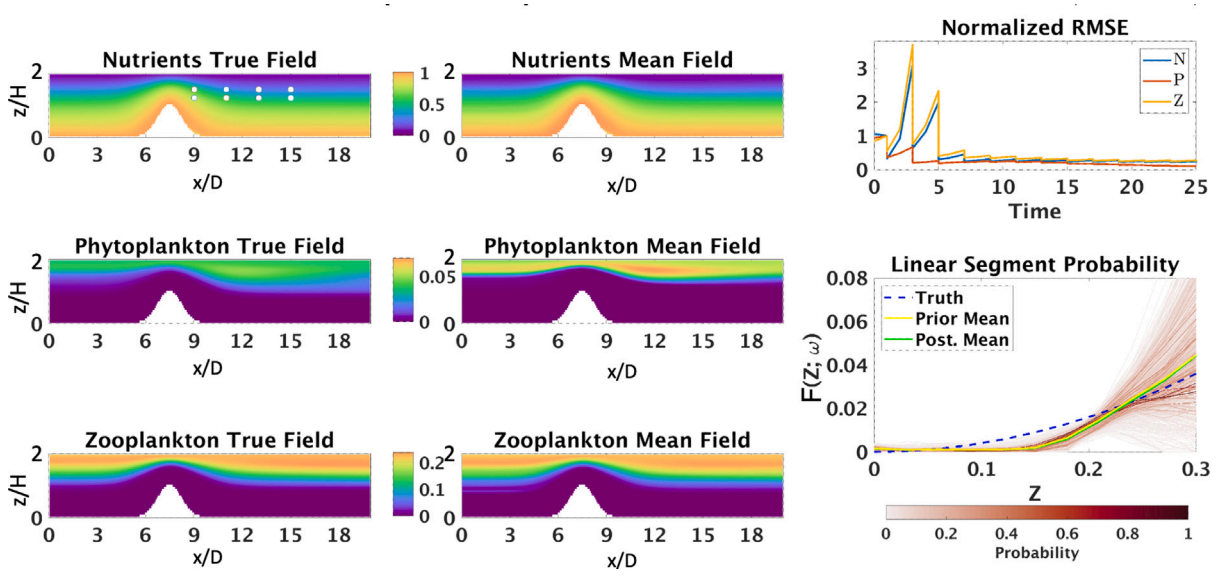



**Fig. 14.** Experiments-3: As Fig. 13 but for posterior fields and function at $t = 25$ (i.e. just after the 13th assimilation). In the second panel of the third column, the prior mean functional form estimate at $t = 25$ is marked with a solid yellow line, while the posterior mean functional form estimate is marked with a solid green line. In the first two columns, the axis limits for the state variables have changed so as to follow the bloom, but in the third column, they remain as in Fig. 13 so as to directly highlight the uncertainty evolution.

*Sensitivity Studies.* Other experiments included studying the effect of incorporating or excluding prior knowledge such as the function value being 0 for $Z = 0$ and using smoothly varying function realizations. For the former, sampling $\gamma_k$'s independent of each other led to highly fluctuating function realizations which completely impaired the learnability of the unknown function. For the latter, enforcing $\gamma_0 = 0$ sets $F(0;\omega) = 0$ for all realizations improved the convergence among the learned function realizations and the true function. Finally, increasing the number of independent observations (more $N$ data, data for $Z$ or $P$ as well, etc.) also improved the sharpness of our GMM-DO inference: in all examples we show, we highlight cases with sparse observations as seen in real ocean applications.

### 5.4. Experiments 4: Learning in chaotic dynamics

In the last set of experiments, in order to robustly test our algorithms, the aim is to learn a five-component NNPZD model with a flow of Reynolds number $Re = 500$. At such high $Re$, vortices start to shed in the wake of the ridge and the flow chaotic. The learning objectives include all 5 biogeochemical fields, the Ivlev grazing parameter ($\Lambda$), the phytoplankton-specific mortality rate ($\Xi$), the zooplankton maximum

grazing rate ($R_m$), the zooplankton specific mortality ($\Gamma$), and the presence or absence of the quadratic zooplankton mortality term. Once again, the ambiguity in the quadratic zooplankton mortality function corresponds to the $\hat{\mathcal{L}}$ term introduced in Eq. (1). The stochastic NNPZD reactions, with all the uncertain parameters explicitly containing $\omega$ as an argument, are given by,

$$
\begin{aligned}
S^{\mathrm{NO_3}} &= \Omega \mathrm{NH_4} - G\left[\frac{\mathrm{NO_3}}{\mathrm{NO_3}+K_u}\exp^{-\Psi_I \mathrm{NH_4}}\right]P\,,\\
S^{\mathrm{NH_4}} &= -\Omega \mathrm{NH_4} + \Phi D + \Gamma(\omega)Z + \alpha(\omega)\underbrace{(\Gamma_2 Z^2)}_{\text{Quad. Z Mort.}}\\
&\quad - G\left[\frac{\mathrm{NH_4}}{\mathrm{NH_4}+K_u}\right]P\,,\\
S^{P} &= G\left[\frac{\mathrm{NO_3}}{\mathrm{NO_3}+K_u}\exp^{-\Psi_I \mathrm{NH_4}} + \frac{\mathrm{NH_4}}{\mathrm{NH_4}+K_u}\right]P - \Xi(\omega)P\,,\\
&\quad - R_m(\omega)Z(1-\exp^{-\Lambda(\omega)P})\,,\\
S^{Z} &= R_m(\omega)(1-\gamma)Z(1-\exp^{-\Lambda(\omega)P}) - \Gamma(\omega)Z + \alpha(\omega)\underbrace{(\Gamma_2 Z^2)}_{\text{Quad. Z Mort.}}\,,\\
S^{D} &= R_m(\omega)\gamma Z(1-\exp^{-\Lambda(\omega)P}) + \Xi(\omega)P - \Phi D\,.
\end{aligned}
\tag{33}
$$

Initially, we assume uniform and independent pdfs for the four uncertain regular parameters and equiprobability for the quadratic zooplankton mortality term to be present or absent. The stochastic ADR PDEs with the stochastic NNPZD reactions (33) are coupled with the deterministic RANS flow PDEs, and solved with the DO methodology (Sections 4.3–4.5). The other known physical–biogeochemical model parameters as well as the hyper-parameters for the DO equations are provided in Table 1.

*True solution generation:* The true solution from which observations are extracted, corresponds to the non-dimensional values, 1.5 for $\Lambda$, 0.04 for $\Xi$, 0.6 for $R_m$, 0.14 for $\Gamma$, and 0 for $\alpha$, i.e. the quadratic mortality term absent. The state fields are initialized and evolved as described in Section 4.7.

*Observations and learning parameters:* The noisy observations remain sparse and univariate, but due to the unstable and fast dynamics of the flow, there is a need for slightly more frequent data than in other experiments. The phytoplankton field is observed at nine locations starting at $t = 2$ and subsequently every one non-dimensional time. In total, we assimilate 24 times, i.e. until $t = 25$, with a non-dimensional $P$-data error standard deviation of 0.04. Other hyper-parameters related to the GMM-DO filtering are given in Table 1.

*Learning metrics:* We compare the true fields and parameters to their DO estimates. To quantify performance, we compute the evolution of the normalized RMSEs for all five biological fields and five stochastic parameters. We also analyze the evolution of pdfs of the regular and formulation parameters, and the variances of DO coefficients.

#### 5.4.1. Learning results

Fig. 15 shows the prior estimates at $t = 2$. The flow has just started to develop. There are significant differences between the true and mean biogeochemical fields. The normalized RMSEs are equal to 1 by construction. The pdfs of all parameters remained as they were initially since no data has been assimilated.

Fig. 16 illustrates the posterior estimates at $t = 2$, just after the first assimilation. Large corrections were made to the mean tracer fields (also visible in their RMSEs that decay by about 15 to 25%), and the GMM-DO learning already predicts the absence of quadratic zooplankton term. These first 9 noisy $P$-observations are not as informative however about the other parameters (their RMSEs only decay by about 4% to 8%).

Fig. 17 shows the estimates at $t = 25$, after 24 GMM-DO assimilation steps. In addition to the mean fields, our augmented filter has been learning the four regular parameters. Their posterior pdfs have become Gaussian which has occurred in intermediate assimilation steps (not shown). We also show the evolution of variance of the top three modes. We find that the total variance on average either decreases or remains similar, while that of individual modes in general decreases but may also increase in accord with the stochastic dynamics. The velocity field being chaotic renders the learning more challenging in this experiment but our framework can still meet all the learning objectives, even with sparse and univariate data.

As the stochastic states, parameters, and model formulations are estimated jointly, our augmented GMM-DO Bayesian learning can provide interesting insights into the co-dependence, biases, and equifinality of all the quantities being estimated. To showcase this capability, we provide joint distributions for combinations of the four uncertain regular parameters in Fig. 18 at $t = 25$. We find that $\Lambda(\omega)$ and $R_m(\omega)$ are negatively correlated and that $\Gamma(\omega)$ and $R_m(\omega)$ are positively correlated, while $\Xi(\omega)$, $\Lambda(\omega)$, and $\Gamma(\omega)$ are nearly independent of each other. We note that such a negative correlation between $\Lambda(\omega)$ and $R_m(\omega)$ can also be inferred from the zooplankton grazing term in the NNPZD model (Eq. (33)). Similarly, $\Gamma(\Omega)$ and $R_m(\omega)$ need to simultaneously increase or decrease to maintain phytoplankton concentration levels. Thus the joint estimation of all the variables of interest can provide the researcher with an essential tool for additional analysis and discoveries.

*Sensitivity Studies*. Other experiments were performed. As expected, they demonstrated sensitivity to the schedule, type, noise, and quantity of observations. With only nine noisy, sparse, and univariate data, starting them after the chaos sets in, or sampling even less frequently than every one non-dimensional time, led to posterior pdf of some stochastic parameters that were not concentrated around their respective true values. Similar results were found even when less than nine data were collected. Adding other observation types improved the learning. For other sensitivity studies, trends similar to other experiments were found.

### 5.5. Remarks and discussions

*Computational costs and implementation.* Our Bayesian model learning framework can be broadly divided into 3 components, construction and initialization of a general model combining all compatible and compatible-embedded candidate models, probabilistic prediction using the DO differential equations, and multivariate Bayesian filtering to update the augmented state, all as summarized in Fig. 1. The first component ensures that a single stochastic model is evolved, theoretically encompassing infinitely many candidate models. Additionally, our overall framework is general enough to be employed in any existing data-assimilation systems capable of joint state and parameter estimation, without much computational overhead. If only state uncertainty exists in the data-assimilation system, then state augmentation (Appendix D) should be leveraged. Employing our framework with existing stochastic models will involve code modification to implement the derived general model. In the Bayesian filtering step, the stochastic formulation, complexity, and expansion parameters can be treated similarly to other regular stochastic parameters. The number of these additional stochastic parameters is expected to be similar to the number of existing regular stochastic parameters. In the current work, the uncertainty evolution is achieved using the DO methodology, while the non-Gaussian Bayesian filtering uses the GMM-DO filter. The computational cost will depend on the problem and software implementation. In general, the highest cost is the probabilistic DO prediction as it commonly adds costs of the order of the cost of $N_s$ deterministic model simulation. Both DO and GMM-DO have been well studied and rigorously compared to competing algorithms. We refer the reader to Sapsis and Lermusiaux (2009, 2012), Feppon and Lermusiaux (2018b), Sondergaard and Lermusiaux (2013a) and Sondergaard and Lermusiaux (2013b) for more details on efficiency and implementation.

*Extension to 3D in space, high-dimensionality, and smoothing*. As mentioned above, our framework for Bayesian state estimation and model

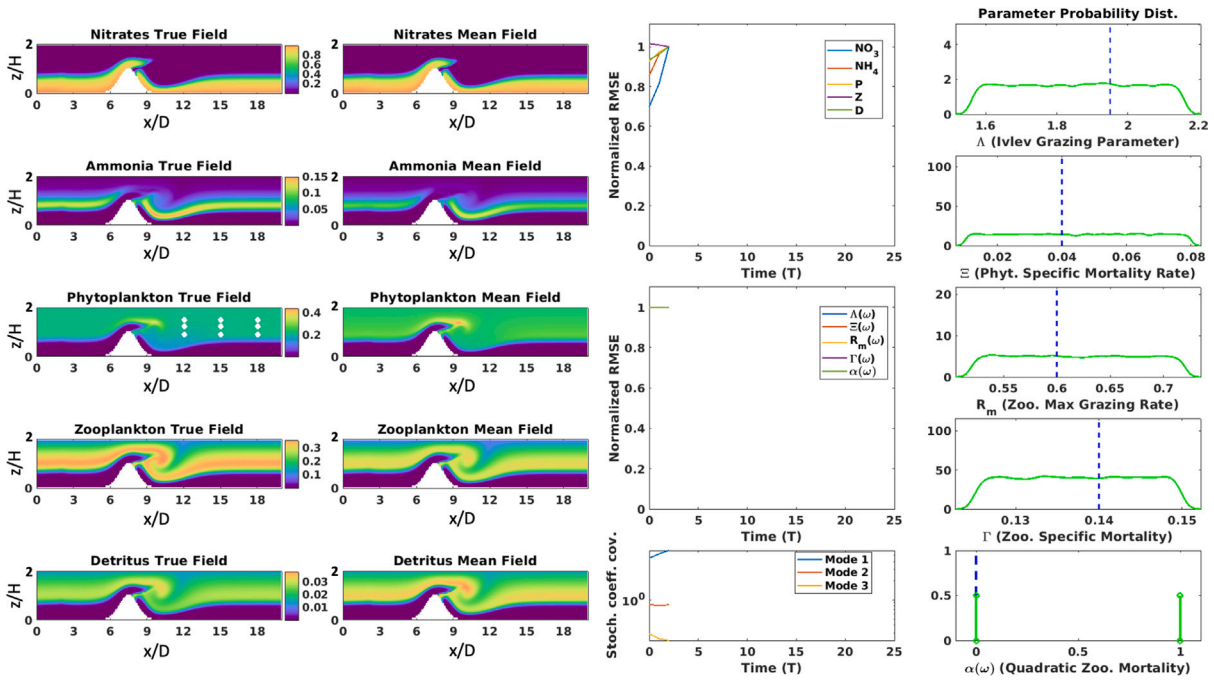



**Fig. 15.** Experiments-4: State of the true and prior estimate NNPZD fields and parameters at $t = 2$ (i.e. just before the 1st assimilation). The first two columns consist of the non-dimensionalized true (left) and mean estimate (right) fields of $NO_3$, $NH_4$, $P$, $Z$, and $D$. In the third column, the first two panels show the evolution of the normalized RMSEs for the five state variables and five parameters. The third panel shows the evolution of variance of the top 3 DO modes. In the fourth column, the pdfs of the non-dimensional $\Lambda(\omega)$, $\Xi(\omega)$, $R_m(\omega)$, $\Gamma(\omega)$, and $\alpha(\omega)$ (learns the presence or absence of quadratic zooplankton mortality) are marked with solid green lines, with the true unknown parameter values marked with blue dotted lines. The velocity field is deterministic with $Re = 500$. Additionally, the white circles on the phytoplankton true field mark the 9 observation locations.

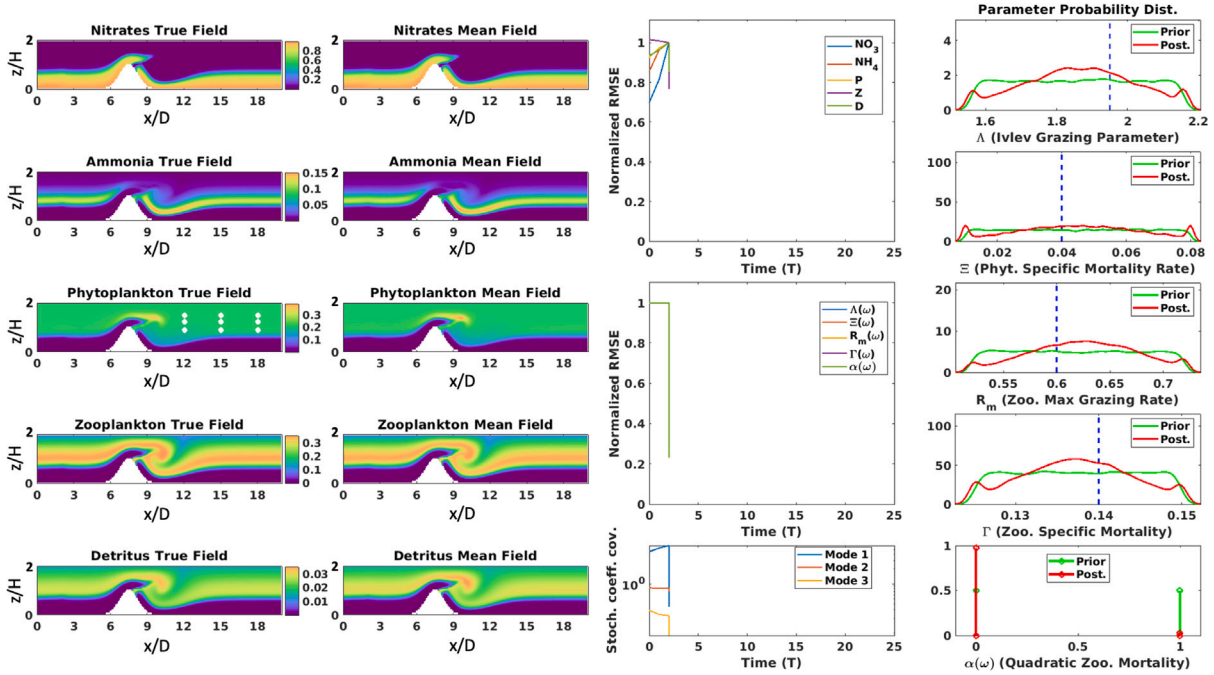



**Fig. 16.** Experiments-4: As Fig. 15, but for posterior fields and parameters at $t = 2$ (i.e. just after the 1st assimilation). In the fourth column, the prior pdfs of the non-dimensional $\Lambda(\omega)$, $\Xi(\omega)$, $R_m(\omega)$, $\Gamma(\omega)$ at $t = 2$ are marked with solid green lines, while the posterior pdfs are marked with solid red lines. In the first two columns, the axis limits for the state variables have changed so as to follow the bloom and decay events, but in the last two columns, they remain as in Fig. 15 so as to directly highlight the uncertainty evolution.

discovery could be used with uncertainty forecasting and data assimilation schemes developed for 3D modeling systems of high numerical dimensions. On the one hand, the DO methodology has been extended to 3D ocean primitive equations with a nonlinear free-surface (Subramani, 2018; Subramani and Lermusiaux, 2023; Gkirgkis, 2021; Gkirgkis and Lermusiaux, 2023). On the other hand, the GMM-DO filter performs the Bayesian update in the DO coefficient subspace with a dimension very small compared to that of the numerical state vector and it has remained efficient for 3D in-space modeling systems. The precursor to the GMM-DO filter, the error subspace statistical estimation (ESSE) scheme (Lermusiaux and Robinson, 1999; Lermusiaux,

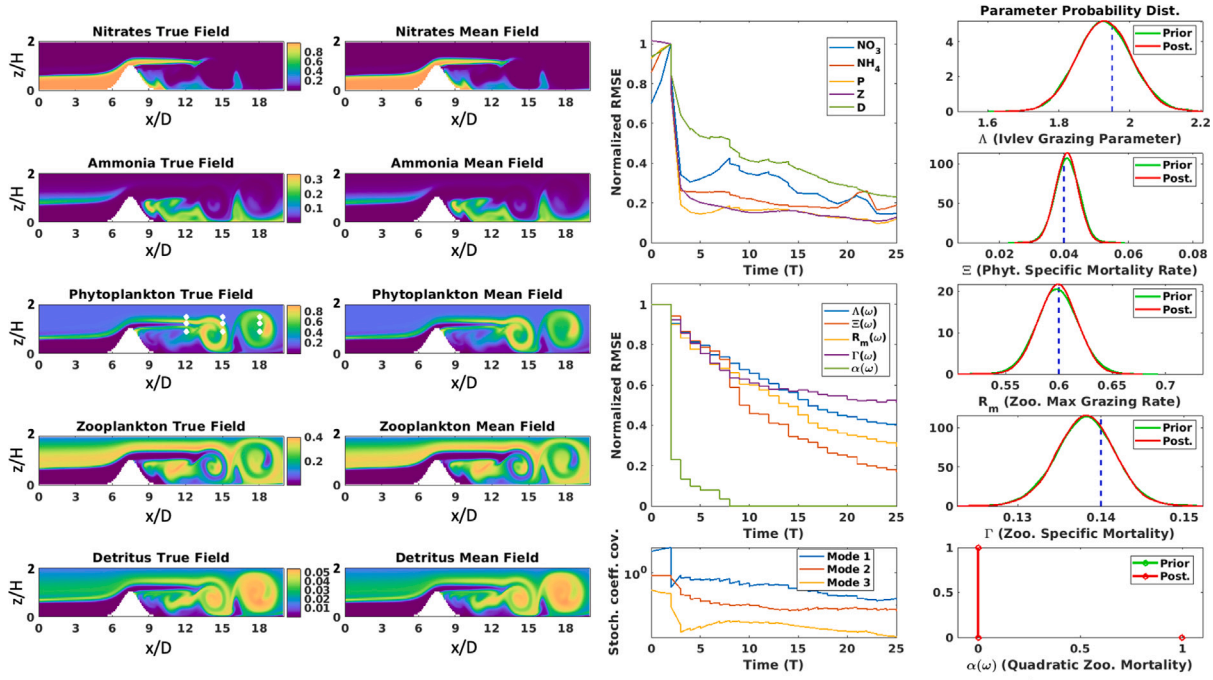



**Fig. 17.** Experiments-4: As Figs. 15 & 16, but for posterior fields and parameters at $t = 25$ (i.e. just after the 24th assimilation). In the first two columns, the axis limits for the state variables have changed so as to follow bloom and decay events, but in the last two columns, they remain as in Fig. 15 so as to directly highlight the uncertainty evolution.

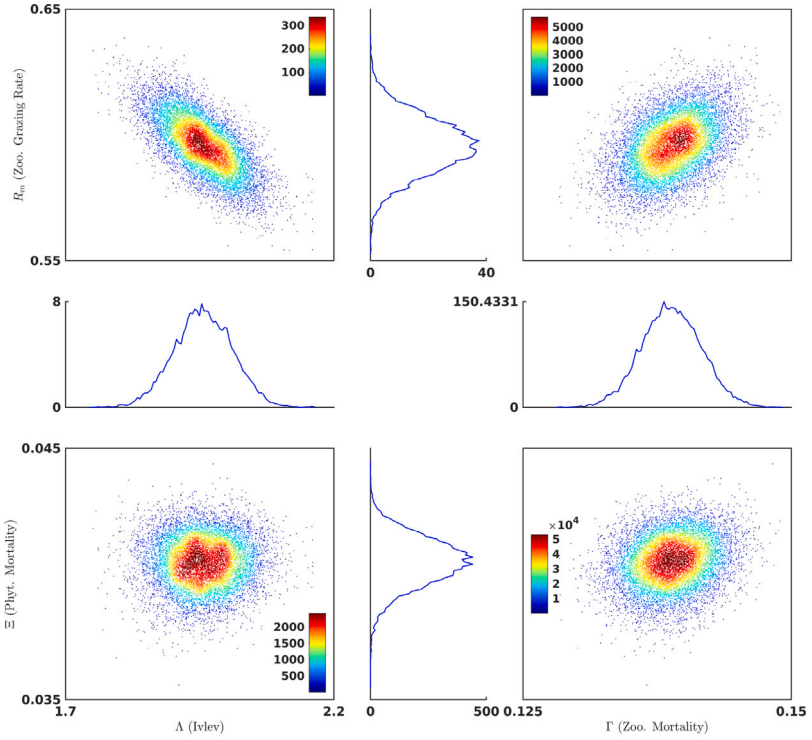



**Fig. 18.** Experiments-4: Posterior joint-double pdf (colored scatter plots) and single pdf (line plots) marginals of the four non-dimensional regular stochastic parameters $\Lambda(\omega)$, $\Xi(\omega)$, $R_m(\omega)$, $\Gamma(\omega)$ at $t = 25$ (i.e. just after the 24th assimilation). The joint-double pdfs highlight non-Gaussian details and dependencies among parameters.

1999a) which also completes its assimilation update in a subspace (a Gaussian update), has been successfully used in 3D for multiple ocean regions, e.g., Lermusiaux (1999b), Lermusiaux et al. (2002), Cossarini et al. (2009) and Lermusiaux et al. (2011, 2020). Its non-Gaussian

versions, the GMM-ensemble and GMM-ESSE filtering schemes, have also been very useful, for example to update subsurface fields based on surface observations (e.g., satellite sea surface temperature or color). We confirmed in our sensitivity studies that surface-only observations can update subsurface fields as long as they contain mutual information (Lermusiaux et al., 2017b). The 3D extension of our novel GMM-DO learning of states and models is thus very promising for use with real ocean data (Haley et al., 2020; Gupta, 2022) as well as for the combination with deep learning (Gupta and Lermusiaux, 2023). The extension to Bayesian smoothing for updating the state variable pdfs backward in time is another direct opportunity (Lolla and Lermusiaux, 2017b,a; Gupta, 2022).

*Constraints, biases, and equifinality*. The GMM-DO Bayesian estimation utilizes scientific knowledge in the form of the dynamic joint prior distribution over the states, parameters, and models, and it can accommodate constraints (Sondergaard and Lermusiaux, 2013a; Lolla and Lermusiaux, 2017b; Lu and Lermusiaux, 2021). Indeed, Eq. (4) shows for example that posterior distributions will always be contained within the support of the priors. Thus, constraints incorporated in the priors such as always-positive parameters will be maintained in the posterior. The joint posterior distribution also captures non-Gaussian and nonlinear dynamical relationships between the states, parameters, and models. For example, in Experiments-4, we provided such an analysis by examining the joint posterior of the regular stochastic parameters (Fig. 18). The joint posterior pdfs can also be useful to find different combinations of parameters, model functions, and complexities that lead to the same solution, also known as equifinality (Duda et al., 2006). Biases can be discovered in a similar fashion (Lu and Lermusiaux, 2014, 2021). For example, the posterior multimodal distributions and thus the possible presence of biases or parameter combinations with equifinality were clearly visible in the pdfs of the Ivlev parameter in Experiments-1 and 2, see Figs. 7, 8, 9, and 11.

*Most informative observations for model learning*. What, when, and where to sample for optimal information are critical questions for the efficient use of resource-constrained observation platforms (Lermusiaux et al., 2017b). Our research group and collaborators have developed and employed "adaptive sampling" methods for varied purposes and ocean regions (Evangelinos et al., 2003; Heaney et al., 2016, 2007; Lermusiaux et al., 2007; Lermusiaux, 2007; Ramp et al., 2009; Wang et al., 2009; Petillo et al., 2015; Cococcioni et al., 2015; Rajan et al., 2021). Using the DO methodology, GMM-DO filter, and reachability schemes, adaptive sampling based on mutual information (MI) has been derived and applied to fluid flows and ocean fields and parameters (Lolla, 2016; Lermusiaux et al., 2017a). The often intractable MI computations are made feasible within the DO subspace, while still accounting for both nonlinear dynamics and non-Gaussian prior pdfs. This MI-based scheme was recently extended to adaptive sampling for optimal learning of dynamical models (Lin, 2020), so as to identify the data that best constrain the model formulation. First, the scheme predicts the MI field between the possible measurement types, locations, and times and the adequacy of model formulations. The optimal sampling scheme then utilizes these predicted MI fields to determine where, when, and what to sample for collecting the most information about the adequacy of competing or unknown model formulations (Lin, 2020; Lermusiaux et al., 2017b). The present work does not employ these MI-based optimal sampling schemes, but this can be done, as highlighted in Lermusiaux et al. (2017b).

## 6. Conclusions

Biogeochemical–physical models for the ocean are inherently uncertain due to the inability to capture all the complex marine interactions and processes with a single mathematical model. Uncertainty manifests itself in many different forms including the initial conditions, boundary conditions, parameters, parameterizations, state variables, and the model complexity and equations themselves. In this work, we develop a Bayesian model learning methodology that interpolates in the space of candidate dynamical models and discovers model formulations, all while estimating state variable fields and parameter values, as well as the joint probability distributions of all learned quantities. It employs the GMM-DO filter and state augmentation to predict and update pdfs of high-dimensional and multidisciplinary ecosystem dynamics governed by PDEs. Using noisy, sparse, and indirect univariate observations and Bayes' law, the complete joint probabilities of biogeochemical–physical fields and parameters, and of known, uncertain, and unknown model formulations are updated. Non-Gaussian statistics, ambiguity, and biases are captured. The parameter values, model functional forms, and model complexities that best explain the data are identified. The first crucial innovations are the stochastic formulation and complexity parameters that unify compatible candidate models, possibly of different complexities, into a single general stochastic PDEs system. A second is the use of stochastic expansion parameters within piecewise function approximations that generate dense candidate model spaces. Our new methodology is generalizable and interpretable, and provides marginal pdfs for all learned model quantities. At the cost of a single stochastic DO model simulation with parameter estimation, it seamlessly and rigorously discriminates among many existing models, possibly none of which are accurate, but also extrapolates out of the space of models to discover new ones.

The performance of our Bayesian learning framework was evaluated using a series of twin experiments based on flows past a ridge with compatible and embedded PDEs for the three-component NPZ model (nutrients ($N$), phytoplankton ($P$), and zooplankton ($Z$)), four-component NPZD model ($N$, $P$, $Z$ and Detritus ($D$)), and five-component NNPZD model (ammonia ($NH_4$), nitrate ($NO_3$), $P$, $Z$, and $D$). In the first set of experiments, we use the NPZ model with uncertain initial conditions, unknown Ivlev grazing parameter value, and ambiguity in the presence or absence of the quadratic zooplankton mortality term. Our new Bayesian learning simultaneously estimated the state variables, Ivlev parameter, and unknown functional form, using noisy sparse Z observations in space and time (only six data points every two non-dimensional times). The posterior pdf of the parameter contained secondary peaks, indicating that alternative combinations of parameter values could explain the observed data. This showcased the ability of our framework to capture non-Gaussian statistics including ambiguity and biases. In the second set of experiments, assimilating just eight noisy N-data every two non-dimensional times, we demonstrated the ability to learn the complexity of the model. We identified the true model within NPZ and NPZD, along with the uncertain fields and Ivlev grazing parameter. In the third set of experiments, we assumed no prior knowledge about the functional form of zooplankton mortality and generated a dense function space using stochastic piece-wise linear approximations. Assimilating just eight noisy N-data every two non-dimensional times, our framework then searched in this rich functional space, estimated the fields and regular parameter values, and was shown capable of discovering the mortality function. The last set of experiments involved learning the complex NNPZD model in an unsteady chaotic deterministic flow with vortex shedding. The NNPZD model had uncertainty in all the tracer fields, four parameters, and in the presence or absence of the zooplankton mortality term. All of the learning objectives were achieved simultaneously, using only nine noisy P-data every non-dimensional time. In all cases, we quantified the learning skill, and evaluated convergence and the sensitivity to hyper-parameters.

These four sets of experiments were complementary, allowing us to showcase the features of our PDE Bayesian learning framework. It successfully discriminates among functional forms and model complexities and also learns in the absence of prior knowledge by searching

in dense function spaces. The next steps include applying this framework to more complex ocean applications, especially to realistic ocean models and to real ocean data. Even though we demonstrate our learning framework using biogeochemical models, it is applicable to other domains with model uncertainty, for example, medicine, economics, energy, etc. Our framework can provide scientists not only the ability to choose between competing existing hypotheses but to also discover new ones in a fundamental Bayesian manner.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

No data was used for the research described in the article.

## Acknowledgments

We thank the members of our MSEAS group for insightful discussions, including A. Babu for proofreading as well as C. Mirabito and P.J. Haley Jr. for help with the figures. We also thank the anonymous reviewers and the editor for their useful comments. We are grateful to the Office of Naval Research for partial support under grants N00014-19-1-2693 (IN-BDA) and N00014-20-1-2023 (MURI ML-SCOPE), and to Sea Grant and NOAA for support under grant NA18OAR4170105 (BIOMAPS), all to the Massachusetts Institute of Technology. We also thank MathWorks and the Mechanical Eng. Dept. at MIT for awarding a competitive 2020-2021 MathWorks Mechanical Eng. Fellowship for A.G.

## Appendix A. Notations

We define the notation used throughout the paper in Table D.2, without repeating definitions already given in Table 1.

## Appendix B. Dynamically orthogonal (DO) equations

In this appendix, we derive the dynamically orthogonal (DO) equations (Sapsis and Lermusiaux, 2009, 2012; Feppon and Lermusiaux, 2018a,b) for optimal reduced-order probabilistic evolution of high-dimensional stochastic dynamical systems with regular and new parametric uncertainties for known, uncertain, and unknown model formulations.

The general stochastic nonlinear dynamical system governs the dynamics of $\boldsymbol{\phi}(\boldsymbol{x},t;\omega) : \mathbb{R}^n \times [0,T] \to \mathbb{R}^{N_\phi}$, the stochastic state vector comprising $N_\phi$ physical–biogeochemical fields defined on a spatial domain $D$, where $\omega$ is the realization index belonging to a measurable sample space $\Omega$. It is given by,

$$\begin{aligned}\frac{\partial \boldsymbol{\phi}(\boldsymbol{x},t;\omega)}{\partial t} &= \mathcal{L}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega),\boldsymbol{\theta}(\omega),\boldsymbol{\beta}(\omega),\boldsymbol{x},t;\omega] + \widehat{\mathcal{L}}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega),\boldsymbol{\alpha}(\omega),\boldsymbol{x},t;\omega] \\ &\quad + \widetilde{\mathcal{L}}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega),\boldsymbol{\gamma}(\omega),\boldsymbol{x},t;\omega], \\ &\qquad \boldsymbol{x}\in D,\ t\in[0,T],\ \omega\in\Omega\,, \\ \text{with}\quad \boldsymbol{\phi}(\boldsymbol{x},0;\omega) &= \boldsymbol{\phi}_o(\boldsymbol{x};\omega)\,, \\ \text{and}\quad \mathcal{B}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega)] &= \boldsymbol{b}(\boldsymbol{x},t;\omega),\ \boldsymbol{x}\in\partial D,\ t\in[0,T],\ \omega\in\Omega\,,\end{aligned} \tag{B.1}$$

where $\boldsymbol{\phi}_o(\boldsymbol{x};\omega)$, $\mathcal{B}$, and $\boldsymbol{b}(\boldsymbol{x},t;\omega)$ are the stochastic initial conditions, boundary condition operators, and boundary values respectively. The functional form of the first dynamics term $\mathcal{L}[\bullet]$ is assumed to be known, however it contains $N_\theta$ uncertain regular parameters $\boldsymbol{\theta}(\omega)$. The second term $\widehat{\mathcal{L}}[\bullet]$ is uncertain: it belongs to a family of candidate functions, parameterized using $N_\alpha$ stochastic formulation parameters $\boldsymbol{\alpha}(\omega)$. $\widehat{\mathcal{L}}[\bullet]$ can also contain uncertain regular parameters $\boldsymbol{\theta}(\omega)$. The candidate models of different complexities are combined using $N_\beta$ stochastic complexity parameters $\boldsymbol{\beta}(\omega)$. The $\beta_k(\omega)$'s multiplied with the original state variables (as described in Section 3.1) are absorbed into $\phi_i$'s and not explicitly shown; however, $\beta_k(\omega)$'s can still appear on the right-hand-side (RHS) in $\mathcal{L}[\bullet]$ and $\widehat{\mathcal{L}}[\bullet]$. The third term $\widetilde{\mathcal{L}}[\bullet]$ has a functional form completely unknown, and is parameterized using $N_\gamma$ stochastic expansion parameters $\boldsymbol{\gamma}(\omega)$.

The DO methodology employs a generalized, time-dependent Karhunen-Loéve decomposition of $\boldsymbol{\phi}(\boldsymbol{x},t;\omega)$ into a mean, $\bar{\boldsymbol{\phi}}(\boldsymbol{x},t)\in\mathbb{R}^{N_\phi}$, $N_s$ deterministic modes, $\tilde{\boldsymbol{\phi}}_i(\boldsymbol{x},t)\in\mathbb{R}^{N_\phi}$, and stochastic coefficients, $Y_i(t;\omega)\in\mathbb{R}$,

$$\boldsymbol{\phi}(\boldsymbol{x},t;\omega) = \bar{\boldsymbol{\phi}}(\boldsymbol{x},t) + \sum_{i=1}^{N_s} Y_i(t;\omega)\tilde{\boldsymbol{\phi}}_i(\boldsymbol{x},t)\,. \tag{B.2}$$

We define the stochastic subspace $\boldsymbol{V_S} = span\{\tilde{\boldsymbol{\phi}}_i(\boldsymbol{x},t)\}_{i=1}^{N_s}$ as the linear space spanned by the $N_s$ deterministic modes that evolve to capture the dominant uncertainty in $\boldsymbol{V_S}$. In general, the number of modes $N_s$ is orders of magnitude smaller than the dimension of the discretized state variables or of the domain grid $N_x$, i.e. $N_s \ll N_\phi N_x$. Similarly, uncertain regular and new parameters are split into means and deviations, $\boldsymbol{\theta}(\omega) = \bar{\boldsymbol{\theta}} + \mathfrak{D}^\theta(\omega)$, $\boldsymbol{\alpha}(\omega) = \bar{\boldsymbol{\alpha}} + \mathfrak{D}^\alpha(\omega)$, and $\boldsymbol{\beta}(\omega) = \bar{\boldsymbol{\beta}} + \mathfrak{D}^\beta(\omega)$.

Nonlinear terms on the RHS are handled using local Taylor series expansion around the statistical means of states and parameters. We use first order Taylor series expansion for the $\mathcal{L}[\bullet]$ and $\widehat{\mathcal{L}}[\bullet]$ terms,

$$\begin{aligned}\mathcal{L}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega),\boldsymbol{\theta}(\omega),\boldsymbol{\beta}(\omega),\boldsymbol{x},t;\omega] &\approx \mathcal{L}|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} + \frac{\partial\mathcal{L}}{\partial\boldsymbol{\phi}}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} \sum_{i=1}^{N_s}\tilde{\boldsymbol{\phi}}_i Y_i \\ &\quad + \sum_{i=1}^{N_\theta}\frac{\partial\mathcal{L}}{\partial\theta_i}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}}\mathfrak{D}_i^\theta + \sum_{i=1}^{N_\beta}\frac{\partial\mathcal{L}}{\partial\beta}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}}\mathfrak{D}_i^\beta\,, \\ \widehat{\mathcal{L}}[\boldsymbol{\phi}(\boldsymbol{x},t;\omega),\boldsymbol{\alpha}(\omega),\boldsymbol{x},t;\omega] &\approx \widehat{\mathcal{L}}|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}}} + \frac{\partial\widehat{\mathcal{L}}}{\partial\boldsymbol{\phi}}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}}} \sum_{i=1}^{N_s}\tilde{\boldsymbol{\phi}}_i Y_i \\ &\quad + \sum_{i=1}^{N_\alpha}\frac{\partial\widehat{\mathcal{L}}}{\partial\alpha_i}\Big|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}}}\mathfrak{D}_i^\alpha\,.\end{aligned} \tag{B.3}$$

Using a higher-order polynomial approximation leads to higher accuracy for the DO evolution, but also increases computational costs. For analyses of the scaling of computational costs with the order of polynomial approximation, we refer to Gupta (2016) and Gupta et al. (2016). Handling the $\widetilde{\mathcal{L}}[\bullet]$ term is less straightforward because of the need to evaluate the interval in which each state realization value lies at all discrete times and spatial locations in the domain (see Section 3.2). Thus, for maximum accuracy, we directly evaluate the $\widetilde{\mathcal{L}}[\bullet]$ terms for every state realization in a Monte-Carlo fashion. To increase efficiency without much loss of accuracy, recent techniques such as dynamic clustering (Humara, 2020; Humara et al., 2022; Charous et al., 2021) could also be used.

To derive the DO equations, we substitute the KL decomposition (Eq. (B.2)) into the stochastic system (Eq. (B.1)). To obtain an efficient closed-form dynamical system, without loss of generality, we impose the DO condition (Sapsis and Lermusiaux, 2009): the rate of change of the stochastic subspace is orthogonal to itself,

$$\frac{d\boldsymbol{V_S}}{dt} \perp \boldsymbol{V_S} \Leftrightarrow \left\langle \frac{\partial\tilde{\boldsymbol{\phi}}_i(\boldsymbol{x},t)}{\partial t}, \tilde{\boldsymbol{\phi}}_j(\boldsymbol{x},t)\right\rangle = 0 \quad \forall i,j\in\{1,\dots,N_s\}\,, \tag{B.4}$$

where $\langle \boldsymbol{a},\boldsymbol{b}\rangle = \int_D \sum_i (a^i b^i) dD$ denotes the spatial inner-product of vectors $\boldsymbol{a} = [a^1,a^2,\dots]^T$ and $\boldsymbol{b} = [b^1,b^2,\dots]^T$. Note that the DO condition (B.4) also implies the preservation of orthogonality for the basis $\{\tilde{\boldsymbol{\phi}}_i(\boldsymbol{x},t)\}_{i=1}^{N_s}$ themselves (Ueckermann et al., 2013). Substituting Eq. (B.2) into Eq. (B.1), and using Eq. (B.4) and the above schemes for nonlinear terms, we derive independent evolution equations for the DO mean, modes, and stochastic coefficients. These are the DO evolution

equations (omitting function arguments for brevity),

$$\frac{\partial \bar{\boldsymbol{\phi}}}{\partial t} = \mathcal{L}|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} + \widehat{\mathcal{L}}|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} + \mathbb{E}[\widetilde{\mathcal{L}}]\,,$$

$$\frac{\partial \tilde{\boldsymbol{\phi}}_i}{\partial t} = \boldsymbol{Q}_i - \sum_{j=1}^{N_s} \langle \boldsymbol{Q}_i, \tilde{\boldsymbol{\phi}}_j \rangle \tilde{\boldsymbol{\phi}}_j\,,$$

$$\begin{aligned}
\frac{dY_i}{dt} = {} & \sum_{m=1}^{N_s} \langle \boldsymbol{F}_m, \tilde{\boldsymbol{\phi}}_i \rangle Y_m + \sum_{m=1}^{N_\theta} \left\langle \frac{\partial \mathcal{L}}{\partial \theta_i}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}}, \tilde{\boldsymbol{\phi}}_i \right\rangle \mathfrak{D}_m^\theta \\
& + \sum_{m=1}^{N_\beta} \left\langle \frac{\partial \mathcal{L}}{\partial \beta}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}}, \tilde{\boldsymbol{\phi}}_i \right\rangle \mathfrak{D}_m^\beta \\
& + \sum_{m=1}^{N_\theta} \left\langle \frac{\partial \widehat{\mathcal{L}}}{\partial \theta_i}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}}, \tilde{\boldsymbol{\phi}}_i \right\rangle \mathfrak{D}_m^\theta + \sum_{m=1}^{N_\alpha} \left\langle \frac{\partial \widehat{\mathcal{L}}}{\partial \alpha_i}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}}, \tilde{\boldsymbol{\phi}}_i \right\rangle \mathfrak{D}_m^\alpha \\
& + \sum_{m=1}^{N_\beta} \left\langle \frac{\partial \widehat{\mathcal{L}}}{\partial \beta_i}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}}, \tilde{\boldsymbol{\phi}}_i \right\rangle \mathfrak{D}_m^\beta \\
& + \left\langle \widetilde{\mathcal{L}} - \mathbb{E}[\widetilde{\mathcal{L}}], \tilde{\boldsymbol{\phi}}_i \right\rangle\,,
\end{aligned} \tag{B.5}$$

where,

$$\begin{aligned}
\boldsymbol{Q}_i = {} & \frac{\partial \mathcal{L}}{\partial \boldsymbol{\phi}}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} \tilde{\boldsymbol{\phi}}_i + \sum_{j=1}^{N_s} \sum_{n=1}^{N_\theta} C^{-1}_{Y_i Y_j} C_{\mathfrak{D}_n^\theta Y_j} \frac{\partial \mathcal{L}}{\partial \theta_n}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} \\
& + \sum_{j=1}^{N_s} \sum_{n=1}^{N_\beta} C^{-1}_{Y_i Y_j} C_{\mathfrak{D}_n^\beta Y_j} \frac{\partial \mathcal{L}}{\partial \beta_n}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} \\
& + \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{\phi}}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} \tilde{\boldsymbol{\phi}}_i + \sum_{j=1}^{N_s} \sum_{n=1}^{N_\theta} C^{-1}_{Y_i Y_j} C_{\mathfrak{D}_n^\theta Y_j} \frac{\partial \widehat{\mathcal{L}}}{\partial \theta_n}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} \\
& + \sum_{j=1}^{N_s} \sum_{n=1}^{N_\alpha} C^{-1}_{Y_i Y_j} C_{\mathfrak{D}_n^\alpha Y_j} \frac{\partial \widehat{\mathcal{L}}}{\partial \alpha_n}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} \\
& + \sum_{j=1}^{N_s} \sum_{n=1}^{N_\beta} C^{-1}_{Y_i Y_j} C_{\mathfrak{D}_n^\beta Y_j} \frac{\partial \widehat{\mathcal{L}}}{\partial \beta_n}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} + \sum_{j=1}^{N_s} C^{-1}_{Y_i Y_j} \mathbb{E}[Y_j \widetilde{\mathcal{L}}]\,,
\end{aligned} \tag{B.6}$$

$$\boldsymbol{F}_m = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\phi}}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} \tilde{\boldsymbol{\phi}}_m + \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{\phi}}\bigg|_{\substack{\boldsymbol{\phi}=\bar{\boldsymbol{\phi}},\\ \boldsymbol{\theta}=\bar{\boldsymbol{\theta}},\\ \boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}},\\ \boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}} \tilde{\boldsymbol{\phi}}_m\,,$$

and $\mathbb{E}[\bullet]$ represents the expectation operator, $C^{-1}_{Y_i Y_j}$ the inverse of the cross-covariance between the $i$th and $j$th stochastic coefficients, and $C_{Y_i Y_j}$ is given by,

$$C_{Y_i, Y_j} = \mathbb{E}[Y_i(t;\omega) Y_j(t;\omega)]\,. \tag{B.7}$$

As discussed in Gupta (2016), Gupta et al. (2016) and Gupta (2022), the boundary conditions are also obtained by inserting DO decompositions in Eq. (B.1). This yields for the mean fields,

$$\mathcal{B}[\bar{\boldsymbol{\phi}}(\boldsymbol{x},t)]|_{\boldsymbol{x}\in\partial D} = \mathbb{E}[\boldsymbol{b}(\boldsymbol{x},t;\omega)]\,, \tag{B.8}$$

and for the modes fields,

$$\mathcal{B}[\tilde{\boldsymbol{\phi}}_i(\boldsymbol{x},t)]|_{\boldsymbol{x}\in\partial D} = \sum_{j=1}^{N_s} \mathbb{E}[Y_j(t;\omega)\boldsymbol{b}(\boldsymbol{x},t;\omega)] C^{-1}_{Y_i Y_j}\,. \tag{B.9}$$

Similarly, the initial conditions in Eq. (B.2) are approximated by using the DO decomposition of the initial stochastic fields $\boldsymbol{\phi}_o(\boldsymbol{x};\omega)$.

Finally, the stochastic dynamical system (Eq. (B.1)) is multivariate and we normalize the spatial inner-product operator using appropriate scaling, so as to account for the different uncertainty magnitudes of state variables (Lermusiaux, 1999a; Lermusiaux et al., 2000; Lermusiaux, 2002; Subramani, 2018; Subramani and Lermusiaux, 2023). For the present DO modes $\tilde{\boldsymbol{\phi}}_i(\boldsymbol{x},t) = [\tilde{\phi}_i^1(\boldsymbol{x},t), \ldots, \tilde{\phi}_i^{N_\phi}(\boldsymbol{x},t)]$, the normalized spatial inner-product is,

$$\langle \tilde{\boldsymbol{\phi}}_i(\boldsymbol{x},t), \tilde{\boldsymbol{\phi}}_j(\boldsymbol{x},t) \rangle = \frac{1}{|D|} \int_D \sum_{k=1}^{N_\phi} \left( \frac{1}{\sigma^2_{nd,k}} \tilde{\phi}_i^k \tilde{\phi}_j^k \right) dD\,, \tag{B.10}$$

where $|D|$ is the volume (area) of the domain and $\sigma_{nd,k}$ is the expected volume-averaged standard deviations of state variable $k$. These $\sigma_{nd,k}$'s normalize the relative weights given to state variables in the inner-product.

## Appendix C. Gaussian mixture model (GMM)-DO filter

The GMM-DO filter (Sondergaard and Lermusiaux, 2013a,b) consists of a recursive succession in time of two steps: a forecast DO step (Appendix B) and a Bayesian update step. Using the affine transformation between stochastic coefficients and state variables (Eq. (B.2)), the GMM-DO filter obtains the Bayesian update of the state variable distribution through an equivalent update of the stochastic coefficient distribution. The result is an efficient reduced-dimension Bayesian state variable inference (Sondergaard and Lermusiaux, 2013a). Next, we assume the DO coefficients of the discrete state variables are augmented with the regular and new parameters (see Appendix D).

For the Bayesian update, the GMM-DO filter first represents the prior probability distribution of the stochastic coefficients in the DO subspace using a GMM,

$$p_{\boldsymbol{Y}^f}(Y^f) \approx \sum_{j=1}^{N_{\text{GMM}}} \pi^f_{\boldsymbol{Y},j} \times \mathcal{N}(Y^f; \boldsymbol{\mu}^f_{\boldsymbol{Y},j}, \boldsymbol{\Sigma}^f_{\boldsymbol{Y},j}) \qquad \forall Y^f \in \mathbb{R}^{N_s}\,, \tag{C.1}$$

where $N_{\text{GMM}}$ is the to-be-determined number of GMM components, $\pi^f_{\boldsymbol{Y},j} \in [0,1]$ the $j$th component weight (also $\sum_{j=1}^{N_{\text{GMM}}} \pi^f_{\boldsymbol{Y},j} = 1$), $\boldsymbol{\mu}^f_{\boldsymbol{Y},j}$ the $j$th component mean vector, and $\boldsymbol{\Sigma}^f_{\boldsymbol{Y},j}$ the $j$th component covariance matrix. This approximation is found by performing a semiparametric fit to the Monte-Carlo samples used to numerically evolve the stochastic coefficients. Specifically, the expectation–maximization (EM) algorithm for GMMs (Bilmes et al., 1998) is used to find maximum likelihood estimate for the GMM parameters $\pi^f_{\boldsymbol{Y},j}$, $\boldsymbol{\mu}^f_{\boldsymbol{Y},j}$ and $\boldsymbol{\Sigma}^f_{\boldsymbol{Y},j}$, while the selection of the number of GMM components ($N_{\text{GMM}}$) is determined by the Bayesian Information Criterion (BIC) (Stoica and Selen, 2004) by successively fitting GMMs of varying complexity (e.g. GMM = 1, 2, 3, ...) until a minimum of the BIC is obtained.

Using the Gaussian observation model (Eq. (3)), the GMM for the prior stochastic coefficients is updated by Bayesian update, using conjugacy (Sondergaard and Lermusiaux, 2013a). The resulting GMM of the posterior stochastic coefficients is,

$$p_{\boldsymbol{Y}^a}(Y^a) \approx \sum_{j=1}^{N_{\text{GMM}}} \pi^a_{\boldsymbol{Y},j} \times \mathcal{N}(Y^a; \boldsymbol{\mu}^a_{\boldsymbol{Y},j}, \boldsymbol{\Sigma}^a_{\boldsymbol{Y},j})\,, \qquad \forall Y^a \in \mathbb{R}^{N_s}\,, \tag{C.2}$$

where,

$$\begin{aligned}
\pi^a_{\boldsymbol{Y},j} &= \frac{\pi^f_{\boldsymbol{Y},j} \times \mathcal{N}(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{H}}\boldsymbol{\mu}^f_{\boldsymbol{Y},j}, \tilde{\boldsymbol{H}}\boldsymbol{\Sigma}^f_{\boldsymbol{Y},j}\tilde{\boldsymbol{H}}^T + \boldsymbol{R})}{\sum_{m=1}^{N_{\text{GMM}}} \pi^f_{\boldsymbol{Y},m} \times \mathcal{N}(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{H}}\boldsymbol{\mu}^f_{\boldsymbol{Y},m}, \tilde{\boldsymbol{H}}\boldsymbol{\Sigma}^f_{\boldsymbol{Y},m}\tilde{\boldsymbol{H}}^T + \boldsymbol{R})}\,, \qquad \forall j \in \{1, \ldots, N_{\text{GMM}}\}\,, \\
\boldsymbol{\mu}^a_{\boldsymbol{Y},j} &= \hat{\boldsymbol{\mu}}^a_{\boldsymbol{Y},j} - \sum_{m=1}^{N_{\text{GMM}}} \pi^a_{\boldsymbol{Y},m} \times \hat{\boldsymbol{\mu}}^a_{\boldsymbol{Y},m}\,, \qquad \forall j \in \{1, \ldots, N_{\text{GMM}}\}\,, \\
\boldsymbol{\Sigma}^a_{\boldsymbol{Y},j} &= (\boldsymbol{I} - \tilde{\boldsymbol{K}}_j \tilde{\boldsymbol{H}}) \boldsymbol{\Sigma}^f_{\boldsymbol{Y},j}\,, \qquad \forall j \in \{1, \ldots, N_{\text{GMM}}\}\,,
\end{aligned} \tag{C.3}$$

**Table D.2**
Notation compendium.

| | | |
|---|---|---|
| **General** | | |
| $\boldsymbol{x}$ | $\in \mathbb{R}^n$ | Spatial coordinate vector |
| $t$ | $\in \mathbb{R}$ | Temporal coordinate |
| $T$ | $\in \mathbb{R}$ | Total simulation time |
| $\mathcal{D}$ | | Simulation domain |
| $\partial D$ | | Simulation domain boundary |
| $\omega$ | | Realization index |
| $\Omega$ | | Measurable sample space |
| $N_x$ | $\in \mathbb{N}$ | Size of the discretized domain |
| $\boldsymbol{\phi}$ | $\in \mathbb{R}^{N_v}$ or $\mathbb{R}^{N_\phi}$ | General state vector or biological tracer fields |
| $\boldsymbol{\phi}_o$ | $\in \mathbb{R}^{N_v}$ or $\mathbb{R}^{N_\phi}$ | Initial condition of $\boldsymbol{\phi}$ |
| $N_v$ | $\in \mathbb{N}$ | Number of state variables |
| $N_v(i)$ | $\in \mathbb{N}$ | Number of state variables in the $i$th candidate model of different complexity |
| $\{\phi^i_1, \ldots, \phi^i_{N_v(i)}\}$ | $\in \mathbb{R}^{N_v(i)}$ | State variables for the $i$th candidate model of different complexity |
| $\boldsymbol{\Phi}$ | $\in \mathbb{R}^{N_v N_x}$ | Discretized state vector of $\boldsymbol{\phi}$ |
| $\boldsymbol{b}$ | | Boundary values |
| $\mathcal{M}$ | | Candidate model |
| $N_m$ | $\in \mathbb{N}$ | Number of candidate models |
| $\boldsymbol{\theta}$ | $\in \mathbb{R}^{N_\theta}$ | Uncertain regular parameters |
| $N_\theta$ | $\in \mathbb{N}$ | Number of uncertain regular parameters |
| $\boldsymbol{u}$ | $\in \mathbb{R}^{N_2}$ | Velocity field |
| $N_\phi$ | $\in \mathbb{N}$ | Number of biological tracers |
| $p$ | $\in \mathbb{R}^2$ | Pressure field |
| $u$ and $v$ | $\in \mathbb{R}$ | Horizontal and vertical velocity |
| $x$ and $z$ | $\in \mathbb{R}$ | Horizontal and vertical direction |
| $i$, $j$, $m$, and $n$ | $\in \mathbb{N}$ | Miscellaneous index |
| **Bayesian model learning** | | |
| $\boldsymbol{\alpha}$ | $\in \mathbb{R}^{N_\alpha}$ | Stochastic formulation parameters for combining candidate models with different functional forms |
| $N_\alpha$ | $\in \mathbb{N}$ | Number of stochastic formulation parameters $\alpha_k$'s |
| $\boldsymbol{\beta}$ | $\in \mathbb{R}^{N_\beta}$ | Stochastic complexity parameters for combining candidate models of different complexities |
| $N_\beta$ | $\in \mathbb{N}$ | Number of stochastic complexity parameters $\beta_k$'s |
| $\mathcal{H}$ | | Range of values taken by the state variable |
| $I_i$ | | Interval with non-zero measure |
| $N_I$ | $\mathbb{N}$ | Number of intervals |
| $\boldsymbol{\gamma}$ | $\in \mathbb{R}^{N_\gamma}$ | Stochastic expansion parameters |
| $N_\gamma$ | $\in \mathbb{N}$ | Number of stochastic expansion parameters $\gamma_k$'s |
| $k$ | $\in \mathbb{N}$ | Index for uncertain regular parameters and stochastic formulation, complexity, and expansion parameters |
| **DO evolution equations** | | |
| $N_s$ | $\in \mathbb{N}$ | Number of DO modes |
| $N_r$ | $\in \mathbb{N}$ | Number of Monte-Carlo samples |
| $\bar{\boldsymbol{\phi}}$ | $\in \mathbb{R}^{N_\phi}$ | Biological tracer DO mean |
| $\bar{\boldsymbol{\Phi}}$ | $\in \mathbb{R}^{N_\phi N_x}$ | Discretized biological tracer DO mean $\bar{\boldsymbol{\phi}}$ |
| $\tilde{\boldsymbol{\phi}}_i$ | $\in \mathbb{R}^{N_\phi}$ | $i$th biological tracer DO mode |
| $\tilde{\boldsymbol{\Phi}}_i$ | $\in \mathbb{R}^{N_\phi N_x \times N_s}$ | Discretized biological tracer DO modes matrix |
| $Y_i$ | $\in \mathbb{R}^{N_s}$ | $i$th DO stochastic coefficient |
| $\boldsymbol{Y}$ | $\in \mathbb{R}^{N_r \times N_s}$ | DO stochastic coefficient matrix |
| $\bar{\boldsymbol{\theta}}$, $\bar{\boldsymbol{\alpha}}$, $\bar{\boldsymbol{\beta}}$, and $\bar{\boldsymbol{\gamma}}$ | $\in \mathbb{R}^{N_\theta}$, $\mathbb{R}^{N_\alpha}$, $\mathbb{R}^{N_\beta}$, and $\mathbb{R}^{N_\gamma}$ resp. | Mean vectors of uncertain parameters |
| $\mathfrak{D}^\theta$, $\mathfrak{D}^\alpha$, $\mathfrak{D}^\beta$, and $\mathfrak{D}^\gamma$ | $\in \mathbb{R}^{N_\theta}$, $\mathbb{R}^{N_\alpha}$, $\mathbb{R}^{N_\beta}$, and $\mathbb{R}^{N_\gamma}$ resp. | Mean removed (deviation) part of uncertain parameters |
| $\sigma_{nd,\bullet}$ | $\in \mathbb{R}$ | Weight of different state variables in inner-product computation |
| **GMM-DO filter** | | |
| $\boldsymbol{\mathcal{Y}}$ | $\in \mathbb{R}^{N_y}$ | Observation vector |
| $\boldsymbol{y}$ | $\in \mathbb{R}^{N_y}$ | Observation vector realization |
| $N_y$ | $\in \mathbb{N}$ | Number of observations |
| $\boldsymbol{H}$ | $\in \mathbb{R}^{N_y \times N_v N_x}$ or $\mathbb{R}^{N_y \times N_\phi N_x}$ | Linear observation matrix |
| $\boldsymbol{V}$ | $\in \mathbb{R}^{N_y}$ | Measurement noise vector |
| $\boldsymbol{R}$ | $\in \mathbb{R}^{N_y \times N_y}$ | Covariance matrix of measurement noise |
| $N_{\text{GMM}}$ | $\in \mathbb{N}$ | Number of Gaussian mixture model (GMM) components |
| $\pi^\bullet_{\bullet,j}$ | $\in [0,1]$ | $j$th GMM component weight |
| $\boldsymbol{\mu}^\bullet_{\bullet,j}$ | $\in \mathbb{R}^{N_s}$ or $\mathbb{R}^{N_\phi N_x + N_\theta + N_\alpha + N_\beta + N_\gamma}$ | $j$th GMM component mean vector |
| $\boldsymbol{\Sigma}^{\bullet\bullet}_{\bullet,j}$ | $\in \mathbb{R}^{N_s \times N_s}$ or $\mathbb{R}^{\left(N_\phi N_x + N_\theta + N_\alpha + N_\beta + N_\gamma\right) \times \left(N_\phi N_x + N_\theta + N_\alpha + N_\beta + N_\gamma\right)}$ | $j$th GMM component covariance matrix |
| $\boldsymbol{K}_j$ | $\in \mathbb{R}^{N_v N_x \times N_y}$ or $\mathbb{R}^{N_\phi N_x \times N_y}$ or $\mathbb{R}^{\left(N_\phi N_x + N_\theta + N_\alpha + N_\beta + N_\gamma\right) \times N_y}$ | $j$th Kalman gain matrix |
| $p_\bullet(\bullet)$ | $\in \mathbb{R}$ | Probability distribution value |
| $\tilde{\boldsymbol{y}}$ | $\in \mathbb{R}^{N_y}$ | Transformed observation vector realization |
| $\tilde{\boldsymbol{H}}$ | $\in \mathbb{R}^{N_y \times N_s}$ | Transformed linear observation matrix |
| $\tilde{\boldsymbol{K}}_j$ | $\in \mathbb{R}^{N_s \times N_y}$ | Transformed $j$th Kalman gain matrix |
| $\hat{\boldsymbol{\mu}}^\bullet_{\bullet,j}$ | $\in \mathbb{R}^{N_s}$ | Intermediate $j$th GMM component mean vector |

(*continued on next page*)

**Table D.2** (*continued*).

| | | |
|---|---|---|
| General | | |
| State augmentation | | |
| $\boldsymbol{\Phi}_{\text{aug}}$ | $\in \mathbb{R}^{N_\phi N_x+N_\theta+N_\alpha+N_\beta+N_\gamma}$ | Augmented discretized state vector |
| $\bar{\boldsymbol{\Phi}}_{\text{aug}}$ | $\in \mathbb{R}^{N_\phi N_x+N_\theta+N_\alpha+N_\beta+N_\gamma}$ | Augmented discretized DO mean vector |
| $\tilde{\boldsymbol{\Phi}}_{\text{aug}}$ | $\in \mathbb{R}^{N_\phi N_x+N_\theta+N_\alpha+N_\beta+N_\gamma}$ | Augmented discretized DO modes matrix |
| $\boldsymbol{DY}$ | $\in \mathbb{R}^{N_s+N_\theta+N_\alpha+N_\beta+N_\gamma}$ | Augmented DO stochastic coefficient matrix |
| $\boldsymbol{H}_{\text{aug}}$ | $\in \mathbb{R}^{N_y\times(N_\phi N_x+N_\theta+N_\alpha+N_\beta+N_\gamma)}$ | Augmented linear observation matrix |
| $\boldsymbol{I}$ | $\in \mathbb{R}^{(N_\theta+N_\alpha+N_\beta+N_\gamma)\times(N_\theta+N_\alpha+N_\beta+N_\gamma)}$ | Identity matrix |
| $\boldsymbol{0}$ | | Matrix of zeros of appropriate size |
| Operators, functions, and indicators | | |
| $\mathcal{L}[\bullet]$ | | Functional form of known dynamics |
| $\widehat{\mathcal{L}}[\bullet]$ | | Unknown dynamics belonging to a set of candidate functional forms |
| $\{\widehat{\mathcal{L}}_1[\bullet],\dots,\widehat{\mathcal{L}}_{N_m}[\bullet]\}$ | | Set of candidate functional forms |
| $\widetilde{\mathcal{L}}[\bullet]$ | | Functional form of completely unknown dynamics |
| $\{\mathcal{L}^i_1[\bullet],\dots,\mathcal{L}^i_{N_c(i)}[\bullet]\}$ | | Right-hand-sides of $i$th candidate model of different complexity |
| $\mathcal{N}(\bullet,\bullet)$ | | Multivariate Gaussian distribution |
| $\Psi_k(\bullet)$ | | $k$th Linear function |
| $\{S^{\phi^1}(\bullet),\dots,S^{\phi^{N_\phi}}(\bullet)\}$ | | Biological reaction terms |
| $\widehat{S}^{\phi}(\bullet)$ | | Unknown biological reaction terms belonging to a set of candidate functional forms |
| $\widetilde{S}^{\phi}(\bullet)$ | | Completely unknown biological reaction terms |
| $\nabla_\cdot(\bullet)$ | | Gradient operator |
| $\nabla^2(\bullet)$ | | Diffusion operator |
| $\langle\bullet,\bullet\rangle$ | | $i$th Spatial inner-product |
| $\mathbb{E}[\bullet]$ | | Expectation |
| $C_{\bullet,\bullet}$ | | Cross-covariance |
| $\mathcal{B}[\bullet]$ | | Boundary condition operator |
| $\boldsymbol{V}_S$ | | DO Subspace |
| $(\bullet)^f$ | | Prior |
| $(\bullet)^a$ | | Posterior |

with the following definitions,

$$\begin{aligned}
&\tilde{\boldsymbol{H}} = \boldsymbol{H}\tilde{\boldsymbol{\Phi}}\,,\\
&\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{\Phi}}^f\,,\\
&\hat{\boldsymbol{\mu}}^a_{\boldsymbol{Y},j} = \boldsymbol{\mu}^f_{\boldsymbol{Y},j} + \tilde{\boldsymbol{K}}_j(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{H}}\boldsymbol{\mu}^f_{\boldsymbol{Y},j})\,, \qquad \forall j\in\{1,\dots,N_{\text{GMM}}\}\,,\\
&\tilde{\boldsymbol{K}}_j = \boldsymbol{\Sigma}^f_{\boldsymbol{Y},j}\tilde{\boldsymbol{H}}^T(\tilde{\boldsymbol{H}}\boldsymbol{\Sigma}^f_{\boldsymbol{Y},j}\tilde{\boldsymbol{H}}^T + \boldsymbol{R})^{-1} \equiv \tilde{\boldsymbol{\Phi}}^T\boldsymbol{K}_j\,, \qquad \forall j\in\{1,\dots,N_{\text{GMM}}\}\,.
\end{aligned} \tag{C.4}$$

The posterior GMM state space distribution is obtained from Eq. (C.2) by updating the state vector mean,

$$\bar{\boldsymbol{\Phi}}^a = \bar{\boldsymbol{\Phi}}^f + \tilde{\boldsymbol{\Phi}}\sum_{j=1}^{N_{\text{GMM}}}\pi^a_{\boldsymbol{Y},j}\times\hat{\boldsymbol{\mu}}^a_{\boldsymbol{Y},j}\,. \tag{C.5}$$

In the GMM-DO update step, no matrices of size larger than $N_\phi N_x\times S \ll (N_\phi N_x)^2$ are manipulated. The GMM-DO filter is thus computationally feasible for high-dimensional multivariate PDE systems (Eq. (B.1)).

At last, new Monte-Carlo samples are drawn from the posterior GMM (Eq. (C.2)) and dynamically evolved using the DO evolution Eqs. (B.5) until new observations come in and the filtering process is repeated.

## Appendix D. State augmentation

To simultaneously estimate the uncertain parameters and states, we employ state augmentation (Gelb, 1974). We start by decomposing the stochastic regular parameters ($\boldsymbol{\theta}(\omega)\in\mathbb{R}^{N_\theta}$), formulation and complexity parameters ($\boldsymbol{\alpha}(\omega)\in\mathbb{R}^{N_\alpha}$ and $\boldsymbol{\beta}(\omega)\in\mathbb{R}^{N_\beta}$), and expansion parameters ($\boldsymbol{\gamma}(\omega)\in\mathbb{R}^{N_\gamma}$) into their means and uncertain parts,

$$\begin{aligned}
\boldsymbol{\theta}(\omega) &= \bar{\boldsymbol{\theta}} + \mathfrak{D}^{\boldsymbol{\theta}}(\omega)\,,\\
\boldsymbol{\alpha}(\omega) &= \bar{\boldsymbol{\alpha}} + \mathfrak{D}^{\boldsymbol{\alpha}}(\omega)\,,\\
\boldsymbol{\beta}(\omega) &= \bar{\boldsymbol{\beta}} + \mathfrak{D}^{\boldsymbol{\beta}}(\omega)\,,\\
\boldsymbol{\gamma}(\omega) &= \bar{\boldsymbol{\gamma}} + \mathfrak{D}^{\boldsymbol{\gamma}}(\omega)\,.
\end{aligned} \tag{D.1}$$

The augmented state vector can be written as,

$$\boldsymbol{\Phi}_{\text{aug}}(t;\omega) = \begin{bmatrix}\boldsymbol{\theta}(\omega)\\ \boldsymbol{\alpha}(\omega)\\ \boldsymbol{\beta}(\omega)\\ \boldsymbol{\gamma}(\omega)\\ \boldsymbol{\Phi}(t;\omega)\end{bmatrix} \in \mathbb{R}^{N_\phi N_x+N_\theta+N_\alpha+N_\beta+N_\gamma}\,. \tag{D.2}$$

Now, let us write the DO decomposition for this new augmented system. We define a new coefficient matrix in which each parameter uncertainty amounts to an additional scalar stochastic coefficient,

$$\boldsymbol{DY}(t;\omega) = \left[\mathfrak{D}^{\boldsymbol{\theta}}(\omega)|\mathfrak{D}^{\boldsymbol{\alpha}}(\omega)|\mathfrak{D}^{\boldsymbol{\beta}}(\omega)|\mathfrak{D}^{\boldsymbol{\gamma}}(\omega)|\boldsymbol{Y}(t;\omega)\right] \in \mathbb{R}^{N_s+N_\theta+N_\alpha+N_\beta+N_\gamma}\,, \tag{D.3}$$

a new modes matrix with parameters having unit modes,

$$\tilde{\boldsymbol{\Phi}}_{\text{aug}}(t) = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \tilde{\boldsymbol{\Phi}}(t)\end{bmatrix} \in \mathbb{R}^{(N_\phi N_x+N_\theta+N_\alpha+N_\beta+N_\gamma)\times(N_s+N_\theta+N_\alpha+N_\beta+N_\gamma)}\,, \tag{D.4}$$

and a new augmented mean vector,

$$\bar{\boldsymbol{\Phi}}_{\text{aug}}(t) = \begin{bmatrix}\bar{\boldsymbol{\theta}}\\ \bar{\boldsymbol{\alpha}}\\ \bar{\boldsymbol{\beta}}\\ \bar{\boldsymbol{\gamma}}\\ \bar{\boldsymbol{\Phi}}(t)\end{bmatrix} \in \mathbb{R}^{N_\phi N_x+N_\theta+N_\alpha+N_\beta+N_\gamma}\,. \tag{D.5}$$

Thus, the DO decomposition of the augmented state is given by,

$$\begin{aligned}
\boldsymbol{\Phi}_{\text{aug}}(t;\omega) &= \bar{\boldsymbol{\Phi}}_{\text{aug}}(t) + \sum_{i=1}^{\substack{N_s+N_\theta+N_\alpha\\+N_\beta+N_\gamma}}\tilde{\boldsymbol{\Phi}}_{\text{aug},i}(t)DY_i(t;\omega)\\
&= \bar{\boldsymbol{\Phi}}_{\text{aug}}(t) + \tilde{\boldsymbol{\Phi}}_{\text{aug}}(t)\boldsymbol{DY}(t;\omega)\,.
\end{aligned} \tag{D.6}$$

We can also define the augmented observation model as,

$$\begin{aligned}
\boldsymbol{\mathcal{Y}} &= \begin{bmatrix}\boldsymbol{0} & \boldsymbol{H}\end{bmatrix}\boldsymbol{\Phi}_{\text{aug}} + \boldsymbol{V}, \qquad \boldsymbol{V}\sim\mathcal{N}(0,\boldsymbol{R})\\
&= \boldsymbol{H}_{\text{aug}}\boldsymbol{\Phi}_{\text{aug}} + \boldsymbol{V}\,,
\end{aligned} \tag{D.7}$$

where $\boldsymbol{H}$ is the original observation matrix, and $\boldsymbol{H}_{\text{aug}} \in \mathbb{R}^{N_y \times (N_\phi N_x + N_\theta + N_\alpha + N_\beta + N_\gamma)}$ the augmented observation matrix, while $\boldsymbol{\Phi}_{\text{aug}}$ is the augmented state ensemble.

We can consider the above augmented state vector as forecast for time $t_k$, and employ the GMM-DO filter (Appendix C) to obtain joint posterior distributions of all parameters and state variables. The GMM fit is completed jointly for the (normalized) parameter realizations and DO stochastic coefficients realizations of the discrete state variables. If the observations, commonly of state variables, are informative about some parameter values, the pdf of these parameters will be updated by the Bayesian update, jointly with the pdf of the DO stochastic coefficients of the discrete state variables.

## References


Allen, J.I., Eknes, M., Evensen, G., 2003. An Ensemble Kalman Filter with a complex marine ecosystem model: hindcasting phytoplankton in the Cretan Sea. In: Annales Geophysicae, Vol. 21. pp. 399–411.

Baretta, J.W., 1997. Preface to the European Regional Seas Ecosystem Model II. J. Sea Res. 38 (3), 169–171.

Baretta, J.W., Ebenhöh, W., Ruardij, P., 1995. The European Regional Seas Ecosystem Model, a complex marine ecosystem model. Neth. J. Sea Res. 33 (3), 233–246.

Bassenne, M., Lozano-Durán, A., 2019. Computational model discovery with reinforcement learning. arXiv preprint arXiv:2001.00008.

Bayes, M., Price, M., 1763. An essay towards solving a problem in the doctrine of chances. By the late rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S.. Philos. Trans. (1683-1775) 53, 370–418. http://dx.doi.org/10.1098/rstl.1763.0053.

Beşiktepe, Ş.T., Lermusiaux, P.F.J., Robinson, A.R., 2003. Coupled physical and biogeochemical data-driven simulations of Massachusetts Bay in late summer: Real-time and post-cruise data assimilation. J. Mar. Syst. 40–41, 171–212. http://dx.doi.org/10.1016/S0924-7963(03)00018-6.

Bengtsson, L., Ghil, M., Källén, E., 1981. Dynamic Meteorology: Data Assimilation Methods. Springer.

Bertsekas, D., Tsitsiklis, J.N., 2008. Introduction to Probability, Vol. 1. Athena Scientific.

Bilmes, J.A., et al., 1998. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Vol. 4. International Computer Science Institute, p. 126, (510).

Blackford, J.C., Allen, J.I., Gilbert, F.J., 2004. Ecosystem dynamics at six contrasting sites: a generic modelling study. J. Mar. Syst. 52 (1), 191–215.

Branicki, M., Majda, A.J., 2013. Fundamental limitations of polynomial chaos for uncertainty quantification in systems with intermittent instabilities. Commun. Math. Sci. 11 (1), 55–103.

Brunton, S.L., Proctor, J.L., Kutz, J.N., 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. USA 113 (15), 3932–3937.

Casella, G., Berger, R.L., 2021. Statistical Inference. Cengage Learning.

Charous, A., Humara, M.J., Ali, W.H., Bhabra, M.S., Gupta, A., Lermusiaux, P.F., 2021. Dynamically orthogonal ray equations with adaptive reclustering. J. Acoust. Soc. Am. 150 (4), A209. http://dx.doi.org/10.1121/10.0008139.

Cococcioni, M., Lazzerini, B., Lermusiaux, P., 2015. Adaptive sampling using fleets of underwater gliders in the presence of fixed buoys using a constrained clustering algorithm. In: Proceedings of IEEE OCEANS'15 Conference. IEEE, Genoa, http://dx.doi.org/10.1109/oceans-genova.2015.7271446.

Cossarini, G., Lermusiaux, P.F.J., Solidoro, C., 2009. Lagoon of venice ecosystem: Seasonal dynamics and environmental guidance with uncertainty analyses and error subspace data assimilation. J. Geophys. Res.: Oceans 114 (C6), http://dx.doi.org/10.1029/2008JC005080.

Davis, C.S., Steele, J.H., 1994. Biological/Physical Modeling of Upper Ocean Processes. Tech. rep., Woods Hole Oc. Inst..

Denman, K.L., 2003. Modelling planktonic ecosystems: parameterizing complexity. Prog. Oceanogr. 57 (3–4), 429–452.

Doron, M., Brasseur, P., Brankart, J.-M., 2011. Stochastic estimation of biogeochemical parameters of a 3D ocean coupled physical–biogeochemical model: Twin experiments. J. Mar. Syst. 87 (3), 194–207.

Dowd, M., Jones, E., Parslow, J., 2014. A statistical overview and perspectives on data assimilation for marine biogeochemical models. Environmetrics 25 (4), 203–213.

Duda, R.O., Hart, P.E., et al., 2006. Pattern Classification. John Wiley & Sons.

Evangelinos, C., Chang, R., Lermusiaux, P.F.J., Patrikalakis, N.M., 2003. Rapid real-time interdisciplinary ocean forecasting using adaptive sampling and adaptive modeling and legacy codes: Component encapsulation using XML. In: Computational Science–ICCS 2003. Springer, pp. 375–384. http://dx.doi.org/10.1007/3-540-44864-0_39.

Fasham, M.J.R., Ducklow, H.W., McKelvie, S.M., 1990. A nitrogen-based model of plankton dynamics in the oceanic mixed layer. J. Mar. Res. 48 (3), 591–639.

Fennel, K., Mattern, J.P., Doney, S.C., Bopp, L., Moore, A.M., Wang, B., Yu, L., 2022. Ocean biogeochemical modelling. Nat. Rev. Methods Primers 2 (1), 76.

Fennel, W., Neumann, T., 2014. Introduction to the Modelling of Marine Ecosystems:(with MATLAB Programs on Accompanying CD-ROM). In: Oceanography, vol. 72, Elsevier.

Feppon, F., Lermusiaux, P.F.J., 2018a. Dynamically orthogonal numerical schemes for efficient stochastic advection and Lagrangian transport. SIAM Rev. 60 (3), 595–625. http://dx.doi.org/10.1137/16M1109394.

Feppon, F., Lermusiaux, P.F.J., 2018b. A geometric approach to dynamical model-order reduction. SIAM J. Matrix Anal. Appl. 39 (1), 510–538. http://dx.doi.org/10.1137/16M1095202.

Ferziger, J.H., Perić, M., Street, R.L., 2002. Computational Methods for Fluid Dynamics, Vol. 3. Springer.

Flierl, G., McGillicuddy, D.J., 2002. Mesoscale and submesoscale physical-biological interactions. In: The Sea, Vol. 12. Wiley, New York, pp. 113–185.

Franks, P.J.S., 2002. NPZ models of plankton dynamics: their construction, coupling to physics, and application. J. Oceanogr. 58 (2), 379–387.

Franks, P.J.S., Wroblewski, J.S., Flierl, G.R., 1986. Behavior of a simple plankton model with food-level acclimation by herbivores. Mar. Biol. 91 (1), 121–129.

Friedrichs, M.A.M., et al., 2007. Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups. J. Geophys. Res.: Oceans 112 (C8).

Gelb, A., 1974. Applied Optimal Estimation. MIT Press.

Giricheva, E., 2015. Aggregation in ecosystem models and model stability. Prog. Oceanogr. 134, 190–196.

Gkirgkis, K.A., 2021. Stochastic Ocean Forecasting with the Dynamically Orthogonal Primitive Equations (Master's thesis). Massachusetts Institute of Technology, Department of Mechanical Engineering, Cambridge, Massachusetts.

Gkirgkis, K.A., Lermusiaux, P.F.J., 2023. Massive probabilistic forecasts for the Gulf of Mexico: Dynamically-orthogonal primitive equations. in preparation.

Gupta, A., 2016. Bayesian Inference of Obstacle Systems and Coupled Biogeochemical-Physical Models (Master's thesis). Indian Institute of Technology Kanpur, Kanpur, India.

Gupta, A., 2022. Scientific Machine Learning for Dynamical Systems: Theory and Applications to Fluid Flow and Ocean Ecosystem Modeling (Ph.D. thesis). Massachusetts Institute of Technology, Department of Mechanical Engineering, Cambridge, Massachusetts.

Gupta, A., Ali, W.H., Lermusiaux, P.F.J., 2016. Boundary Conditions for Stochastic DO Equations. MSEAS Report, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Gupta, A., Haley, P.J., Subramani, D.N., Lermusiaux, P.F.J., 2019. Fish modeling and Bayesian learning for the Lakshadweep Islands. In: OCEANS 2019 MTS/IEEE SEATTLE. IEEE, Seattle, pp. 1–10. http://dx.doi.org/10.23919/OCEANS40490.2019.8962892.

Gupta, A., Lermusiaux, P.F.J., 2021. Neural closure models for dynamical systems. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 477 (2252), 1–29. http://dx.doi.org/10.1098/rspa.2020.1004.

Gupta, A., Lermusiaux, P.F.J., 2023. Generalized neural closure models with interpretability. Sci. Rep. http://dx.doi.org/10.48550/arXiv.2301.06198, Sub-judice.

Haley, Jr., P.J., Agarwal, A., Lermusiaux, P.F.J., 2015. Optimizing velocities and transports for complex coastal regions and archipelagos. Ocean Model. 89, 1–28. http://dx.doi.org/10.1016/j.ocemod.2015.02.005.

Haley, Jr., P.J., Gupta, A., Mirabito, C., Lermusiaux, P.F.J., 2020. Towards Bayesian ocean physical-biogeochemical-acidification prediction and learning systems for massachusetts bay. In: OCEANS 2020 IEEE/MTS. IEEE, pp. 1–9. http://dx.doi.org/10.1109/IEEECONF38699.2020.9389210.

Haley, Jr., P.J., Lermusiaux, P.F.J., 2010. Multiscale two-way embedding schemes for free-surface primitive equations in the "Multidisciplinary Simulation, Estimation and Assimilation System". Ocean Dyn. 60 (6), 1497–1537. http://dx.doi.org/10.1007/s10236-010-0349-4.

Hart, P.E., Stork, D.G., Duda, R.O., 2000. Pattern Classification. Wiley Hoboken.

Heaney, K.D., Gawarkiewicz, G., Duda, T.F., Lermusiaux, P.F.J., 2007. Nonlinear optimization of autonomous undersea vehicle sampling strategies for oceanographic data-assimilation. J. Field Robotics 24 (6), 437–448. http://dx.doi.org/10.1002/rob.20183.

Heaney, K.D., Lermusiaux, P.F.J., Duda, T.F., Haley, Jr., P.J., 2016. Validation of genetic algorithm based optimal sampling for ocean data assimilation. Ocean Dyn. 66, 1209–1229. http://dx.doi.org/10.1007/s10236-016-0976-5.

Hecht, M.W., Hasumi, H., 2013. Ocean Modeling in an Eddying Regime, Vol. 177. John Wiley & Sons.

Hofmann, E., Friedrichs, M.A., 2002. Predictive modeling for marine ecosystems. In: The Sea, Vol. 12. John Wiley & Sons, New York, pp. 537–565.

Hu, J., Fennel, K., Mattern, J.P., Wilkin, J., 2012. Data assimilation with a local Ensemble Kalman Filter applied to a three-dimensional biological model of the Middle Atlantic Bight. J. Mar. Syst. 94, 145–156.

Humara, M.J., 2020. Stochastic Acoustic Ray Tracing with Dynamically Orthogonal Equations (Master's thesis). Massachusetts Institute of Technology, Joint Program in Applied Ocean Science and Engineering, Cambridge, Massachusetts.

Humara, M.J., Ali, W.H., Charous, A., Bhabra, M., Lermusiaux, P.F.J., 2022. Stochastic acoustic ray tracing with dynamically orthogonal differential equations. In: OCEANS 2022 IEEE/MTS. IEEE, Hampton Roads, VA, pp. 1–10. http://dx.doi.org/10.1109/OCEANS47191.2022.9977252.

Ide, K., Ghil, M., 1998a. Extended Kalman filtering for vortex systems. Part 1: Methodology and point vortices. Dyn. Atmos. Oceans 27 (1–4), 301–332.

Ide, K., Ghil, M., 1998b. Extended Kalman filtering for vortex systems. Part II: Rankine vortices and observing-system design. Dyn. Atmos. Oceans 27 (1–4), 333–350.

Janjić, T., Bormann, N., Bocquet, M., Carton, J., Cohn, S., Dance, S.L., Losa, S., Nichols, N.K., Potthast, R., Waller, J.A., et al., 2018. On the representation error in data assimilation. Q. J. R. Meteorol. Soc. 144 (713), 1257–1278.

Jones, E., Parslow, J., Murray, L., 2010. A Bayesian approach to state and parameter estimation in a Phytoplankton-Zooplankton model. Aust. Meteorol. Oceanogr. J. 59 (SP), 7–16.

Kulkarni, C.S., Gupta, A., Lermusiaux, P.F.J., 2020. Sparse regression and adaptive feature generation for the discovery of dynamical systems. In: Darema, F., Blasch, E., Ravela, S., Aved, A. (Eds.), Dynamic Data Driven Application Systems. DDDAS 2020. In: Lecture Notes in Computer Science, vol. 12312, Springer, Cham, pp. 208–216. http://dx.doi.org/10.1007/978-3-030-61725-7_25.

Kulkarni, C.S., Lermusiaux, P.F.J., 2019. Advection without compounding errors through flow map composition. J. Comput. Phys. 398, 108859. http://dx.doi.org/10.1016/j.jcp.2019.108859.

Lalli, C., Parsons, T.R., 1997. Biological Oceanography: An Introduction. Elsevier Butterworth-Heinemann.

Lambers, J., 2023. MAT 772 fall semester 2010-11 lecture 17 notes. Available at https://www.math.usm.edu/lambers/mat772/fall10/lecture17.pdf.

Lermusiaux, P.F.J., 1999a. Data assimilation via error subspace statistical estimation, Part II: Mid-atlantic bight shelfbreak front simulations, and ESSE validation. Mon. Weather Rev. 127 (7), 1408–1432. http://dx.doi.org/10.1175/1520-0493(1999)127<1408:DAVESS>2.0.CO;2.

Lermusiaux, P.F.J., 1999b. Estimation and study of mesoscale variability in the strait of sicily. Dyn. Atmos. Oceans 29 (2), 255–303. http://dx.doi.org/10.1016/S0377-0265(99)00008-1.

Lermusiaux, P.F.J., 2001. Evolving the subspace of the three-dimensional multiscale ocean variability: Massachusetts Bay. J. Mar. Syst. 29 (1), 385–422. http://dx.doi.org/10.1016/S0924-7963(01)00025-2.

Lermusiaux, P.F.J., 2002. On the mapping of multivariate geophysical fields: Sensitivities to size, scales, and dynamics. J. Atmos. Ocean. Technol. 19 (10), 1602–1637. http://dx.doi.org/10.1175/1520-0426(2002)019<1602:OTMOMG>2.0.CO;2.

Lermusiaux, P.F.J., 2007. Adaptive modeling, adaptive data assimilation and adaptive sampling. Physica D 230 (1), 172–196. http://dx.doi.org/10.1016/j.physd.2007.02.014.

Lermusiaux, P.F.J., Anderson, D.G.M., Lozano, C.J., 2000. On the mapping of multivariate geophysical fields: Error and variability subspace estimates. Q. J. R. Meteorol. Soc. 126 (565), 1387–1429. http://dx.doi.org/10.1256/smsqj.56509.

Lermusiaux, P.F.J., Chiu, C.-S., Robinson, A.R., 2002. Modeling uncertainties in the prediction of the acoustic wavefield in a shelfbreak environment. In: Shang, E.-C., Li, Q., Gao, T.F. (Eds.), Proceedings of the 5th International Conference on Theoretical and Computational Acoustics. World Scientific Publishing Co., pp. 191–200. http://dx.doi.org/10.1142/9789812777362_0020, Refereed invited manuscript.

Lermusiaux, P.F.J., Evangelinos, C., Tian, R., Haley, Jr., P.J., McCarthy, J.J., Patrikalakis, N.M., Robinson, A.R., Schmidt, H., 2004. Adaptive coupled physical and biogeochemical ocean predictions: A conceptual basis. In: Computational Science - ICCS 2004. In: Lecture Notes in Computer Science, vol. 3038, Springer Berlin Heidelberg, pp. 685–692. http://dx.doi.org/10.1007/978-3-540-24688-6_89.

Lermusiaux, P.F.J., Haley, P.J., Leslie, W.G., Agarwal, A., Logutov, O., Burton, L.J., 2011. Multiscale physical and biological dynamics in the philippine archipelago: Predictions and processes. Oceanography 24 (1), 70–89. http://dx.doi.org/10.5670/oceanog.2011.05, Special Issue on the Philippine Straits Dynamics Experiment.

Lermusiaux, P.F.J., Haley, Jr., P.J., Jana, S., Gupta, A., Kulkarni, C.S., Mirabito, C., Ali, W.H., Subramani, D.N., Dutt, A., Lin, J., Shcherbina, A., Lee, C., Gangopadhyay, A., 2017a. Optimal planning and sampling predictions for autonomous and Lagrangian platforms and sensors in the Northern Arabian Sea. Oceanography 30 (2), 172–185. http://dx.doi.org/10.5670/oceanog.2017.242, Special issue on Autonomous and Lagrangian Platforms and Sensors (ALPS).

Lermusiaux, P.F.J., Haley, Jr., P.J., Yilmaz, N.K., 2007. Environmental prediction, path planning and adaptive sampling: Sensing and modeling for efficient ocean monitoring, management and pollution control. Sea Technol. 48 (9), 35–38.

Lermusiaux, P.F.J., Mirabito, C., Haley, Jr., P.J., Ali, W.H., Gupta, A., Jana, S., Dorfman, E., Laferriere, A., Kofford, A., Shepard, G., Goldsmith, M., Heaney, K., Coelho, E., Boyle, J., Murray, J., Freitag, L., Morozov, A., 2020. Real-time probabilistic coupled ocean physics-acoustics forecasting and data assimilation for underwater GPS. In: OCEANS 2020 IEEE/MTS. IEEE, pp. 1–9. http://dx.doi.org/10.1109/IEEECONF38699.2020.9389003.

Lermusiaux, P.F.J., Robinson, A.R., 1999. Data assimilation via error subspace statistical estimation, Part I: Theory and schemes. Mon. Weather Rev. 127 (7), 1385–1407. http://dx.doi.org/10.1175/1520-0493(1999)127<1385:DAVESS>2.0.CO;2.

Lermusiaux, P.F.J., Subramani, D.N., Lin, J., Kulkarni, C.S., Gupta, A., Dutt, A., Lolla, T., Haley, Jr., P.J., Ali, W.H., Mirabito, C., Jana, S., 2017b. A future for intelligent autonomous ocean observing systems. J. Mar. Res. 75 (6), 765–813. http://dx.doi.org/10.1357/002224017823524035, The Sea. Volume 17, The Science of Ocean Prediction, Part 2..

Lermusiaux, P.F.J., et al., 2017c. A future for intelligent autonomous ocean observing systems. J. Mar. Res. 75 (6), 765–813. http://dx.doi.org/10.1357/002224017823524035, The Sea. Volume 17, The Science of Ocean Prediction, Part 2..

Lermusiaux, P.F.J., et al., 2017d. Optimal planning and sampling predictions for autonomous and Lagrangian platforms and sensors in the Northern Arabian Sea. Oceanography 30 (2), 172–185. http://dx.doi.org/10.5670/oceanog.2017.242, Special issue on Autonomous and Lagrangian Platforms and Sensors (ALPS).

Lin, J., 2020. Bayesian Learning for High-Dimensional Nonlinear Systems: Methodologies, Numerics and Applications to Fluid Flows (Ph.D. thesis). Massachusetts Institute of Technology, Department of Mechanical Engineering, Cambridge, Massachusetts.

Lolla, S.V.T., 2016. Path Planning and Adaptive Sampling in the Coastal Ocean (Ph.D. thesis). Massachusetts Institute of Technology, Department of Mechanical Engineering, Cambridge, Massachusetts.

Lolla, T., Haley, Jr., P.J., Lermusiaux, P.F.J., 2014. Time-optimal path planning in dynamic flows using level set equations: Realistic applications. Ocean Dyn. 64 (10), 1399–1417. http://dx.doi.org/10.1007/s10236-014-0760-3.

Lolla, T., Lermusiaux, P.F.J., 2017a. A Gaussian mixture model smoother for continuous nonlinear stochastic dynamical systems: Applications. Mon. Weather Rev. 145, 2763–2790. http://dx.doi.org/10.1175/MWR-D-16-0065.1.

Lolla, T., Lermusiaux, P.F.J., 2017b. A Gaussian mixture model smoother for continuous nonlinear stochastic dynamical systems: Theory and scheme. Mon. Weather Rev. 145, 2743–2761. http://dx.doi.org/10.1175/MWR-D-16-0064.1.

Losa, S.N., Kivman, G.A., Ryabchenko, V.A., 2004. Weak constraint parameter estimation for a simple ocean ecosystem model: what can we learn about the model and data? J. Mar. Syst. 45 (1), 1–20.

Lu, P.G.Y., Lermusiaux, P.F.J., 2014. PDE-Based Bayesian Inference of High-Dimensional Dynamical Models. MSEAS Report 19, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.

Lu, P., Lermusiaux, P.F.J., 2021. Bayesian learning of stochastic dynamical models. Physica D 427, 133003. http://dx.doi.org/10.1016/j.physd.2021.133003.

Maslyaev, M., Hvatov, A., Kalyuzhnaya, A., 2019. Data-driven PDE discovery with evolutionary approach. arXiv preprint arXiv:1903.08011.

MathWorks, 2023. Solve system of nonlinear equations - MATLAB. Accessed on January 21, 2023. URL https://www.mathworks.com/help/optim/ug/fsolve.html.

Mattern, J.P., Dowd, M., Fennel, K., 2010. Sequential data assimilation applied to a physical–biological model for the Bermuda Atlantic time series station. J. Mar. Syst. 79 (1), 144–156.

Mattern, J.P., Dowd, M., Fennel, K., 2013. Particle filter-based data assimilation for a three-dimensional biological ocean model and satellite observations. J. Geophys. Res.: Oceans 118 (5), 2746–2760.

Mattern, J.P., Fennel, K., Dowd, M., 2012. Estimating time-dependent parameters for a biological ocean model using an emulator approach. J. Mar. Syst. 96, 32–47.

McGillicuddy, D., Lynch, D., Moore, A., Gentleman, W., Davis, C., Meise, C., 1998. An adjoint data assimilation approach to diagnosis of physical and biological controls on Pseudocalanus spp. in the Gulf of Maine–Georges Bank region. Fisheries Oceanography 7 (3–4), 205–218.

McWilliams, J.C., 2008. The nature and consequences of oceanic eddies. In: Ocean Model in an Eddying Regime, Vol. 177. pp. 5–15.

Messenger, D.A., Bortz, D.M., 2021. Weak SINDy for partial differential equations. J. Comput. Phys. 110525.

Natvik, L.J., Evensen, G., 2003. Assimilation of ocean colour data into a biochemical model of the North Atlantic: Part 1. Data assimilation experiments. J. Mar. Syst. 40, 127–153.

Newberger, P.A., Allen, J.S., Spitz, Y.H., 2003. Analysis and comparison of three ecosystem models. J. Geophys. Res.: Oceans (1978–2012) 108 (C3).

Niven, R.K., Mohammad-Djafari, A., Cordier, L., Abel, M., Quade, M., 2020. Bayesian identification of dynamical systems. In: Multidisciplinary Digital Publishing Institute Proceedings, Vol. 33. p. 33.

Novati, G., de Laroussilhe, H.L., Koumoutsakos, P., 2021. Automating turbulence modelling by multi-agent reinforcement learning. Nat. Mach. Intell. 3 (1), 87–96.

Pershing, A., et al., 2019. Temperature and Circulation Conditions in the Gulf of Maine in 2050 and their Expected Impacts. Scientific scenario paper, Gulf of Maine 2050 International Symposium.

Petillo, S., Schmidt, H., Lermusiaux, P.F.J., Yoerger, D., Balasuriya, A., 2015. Autonomous & adaptive oceanographic front tracking on board autonomous underwater vehicles. In: Proceedings of IEEE OCEANS'15 Conference. IEEE, Genoa, http://dx.doi.org/10.1109/oceans-genova.2015.7271616.

Pineda, J., Starczak, V., da Silva, J.C., Helfrich, K., Thompson, M., Wiley, D., 2015. Whales and waves: Humpback whale foraging response and the shoaling of internal waves at Stellwagen Bank. J. Geophys. Res.: Oceans 120 (4), 2555–2570.

Raissi, M., Karniadakis, G.E., 2018. Hidden physics models: Machine learning of nonlinear partial differential equations. J. Comput. Phys. 357, 125–141.

Rajan, K., Aguado, F., Lermusiaux, P., de Sousa, J.B., Subramaniam, A., Tintore, J., 2021. METEOR: A Mobile (Portable) ocEan roboTic ObsErvatORy. Mar. Technol. Soc. J. 55 (3), 74–75. http://dx.doi.org/10.4031/MTSJ.55.3.42.

Ramp, S.R., Davis, R.E., Leonard, N.E., Shulman, I., Chao, Y., Robinson, A.R., Marsden, J., Lermusiaux, P.F.J., Fratantoni, D.M., Paduan, J.D., Chavez, F.P., Bahr, F.L., Liang, S., Leslie, W., Li, Z., 2009. Preparing to predict: The Second Autonomous Ocean Sampling Network (AOSN-II) experiment in the Monterey Bay. Deep Sea Res. II 56 (3–5), 68–86. http://dx.doi.org/10.1016/j.dsr2.2008.08.013.

Robinson, A.R., Lermusiaux, P.F.J., 2002. Data assimilation for modeling and predicting coupled physical–biological interactions in the sea. In: Robinson, A.R., McCarthy, J.J., Rothschild, B.J. (Eds.), Biological-Physical Interactions in the Sea. In: The Sea, vol. 12, John Wiley and Sons, New York, pp. 475–536, chapter 12.

Rudy, S., Alla, A., Brunton, S.L., Kutz, J.N., 2019. Data-driven identification of parametric partial differential equations. SIAM J. Appl. Dyn. Syst. 18 (2), 643–660.

Sapsis, T.P., Lermusiaux, P.F.J., 2009. Dynamically orthogonal field equations for continuous stochastic dynamical systems. Physica D 238 (23–24), 2347–2360. http://dx.doi.org/10.1016/j.physd.2009.09.017.

Sapsis, T.P., Lermusiaux, P.F.J., 2012. Dynamical criteria for the evolution of the stochastic dimensionality in flows with uncertainty. Physica D 241 (1), 60–76. http://dx.doi.org/10.1016/j.physd.2011.10.001.

Silva, T.L., 2021. State of the Science Report: An Addendum to the Stellwagen Bank National Marine Sanctuary 2020 Condition Report.

Sondergaard, T., Lermusiaux, P.F.J., 2013a. Data assimilation with Gaussian Mixture Models using the Dynamically Orthogonal field equations. Part I: Theory and scheme. Mon. Weather Rev. 141 (6), 1737–1760. http://dx.doi.org/10.1175/MWR-D-11-00295.1.

Sondergaard, T., Lermusiaux, P.F.J., 2013b. Data assimilation with Gaussian Mixture Models using the Dynamically Orthogonal field equations. Part II: Applications. Mon. Weather Rev. 141 (6), 1761–1785. http://dx.doi.org/10.1175/MWR-D-11-00296.1.

Stoica, P., Selen, Y., 2004. Model-order selection: a review of information criterion rules. IEEE Signal Process. Mag. 21 (4), 36–47.

Subramani, D.N., 2018. Probabilistic Regional Ocean Predictions: Stochastic Fields and Optimal Planning (Ph.D. thesis). Massachusetts Institute of Technology, Department of Mechanical Engineering, Cambridge, Massachusetts.

Subramani, D.N., Lermusiaux, P.F.J., 2016. Energy-optimal path planning by stochastic dynamically orthogonal level-set optimization. Ocean Model. 100, 57–77. http://dx.doi.org/10.1016/j.ocemod.2016.01.006.

Subramani, D., Lermusiaux, P.F.J., 2023. Probabilistic ocean predictions with dynamically-orthogonal primitive equations. in preparation.

Subramani, D.N., Wei, Q.J., Lermusiaux, P.F.J., 2018. Stochastic time-optimal path-planning in uncertain, strong, and dynamic flows. Comput. Methods Appl. Mech. Engrg. 333, 218–237. http://dx.doi.org/10.1016/j.cma.2018.01.004.

Tian, R., Chen, C., Qi, J., Ji, R., Beardsley, R.C., Davis, C., 2015. Model study of nutrient and phytoplankton dynamics in the Gulf of Maine: patterns and drivers for seasonal and interannual variability. ICES J. Mar. Sci. 72 (2), 388–402.

Tian, R.C., Lermusiaux, P.F.J., McCarthy, J.J., Robinson, A.R., 2004. A Generalized Prognostic Model of Marine Biogeochemical-Ecosystem Dynamics: Structure, Parameterization and Adaptive Modeling. Harvard Reports in Physical/Interdisciplinary Ocean Science 67, Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA.

Toyoda, T., et al., 2013. Improved state estimations of lower trophic ecosystems in the global ocean based on a Green's function approach. Prog. Oceanogr. 119, 90–107.

Trefethen, L.N., 2019. Approximation Theory and Approximation Practice, Extended Edition. SIAM.

Ueckermann, M.P., Lermusiaux, P.F.J., 2012. 2.29 Finite Volume MATLAB Framework Documentation. MSEAS Report 14, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, URL http://mseas.mit.edu/?p=2567.

Ueckermann, M.P., Lermusiaux, P.F.J., Sapsis, T.P., 2013. Numerical schemes for dynamically orthogonal equations of stochastic fluid and ocean flows. J. Comput. Phys. 233, 272–294. http://dx.doi.org/10.1016/j.jcp.2012.08.041.

Van Leer, B., 1977. Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. J. Comput. Phys. 23 (3), 276–299.

Wang, D., Lermusiaux, P.F.J., Haley, Jr., P.J., Eickstedt, D., Leslie, W.G., Schmidt, H., 2009. Acoustically focused adaptive sampling and on-board routing for marine rapid environmental assessment. J. Mar. Syst. 78 (Supplement), S393–S407. http://dx.doi.org/10.1016/j.jmarsys.2009.01.037.

Wang, Y., Shen, Z., Long, Z., Dong, B., 2019. Learning to discretize: solving 1D scalar conservation laws via deep reinforcement learning. arXiv preprint arXiv:1905.11079.

Ward, B.A., Friedrichs, M.A.M., Anderson, T.R., Oschlies, A., 2010. Parameter optimisation techniques and the problem of underdetermination in marine biogeochemical models. J. Mar. Syst. 81 (1), 34–43.

Ward, B.A., et al., 2013. When is a biogeochemical model too complex? Objective model reduction and selection for North Atlantic time-series sites. Prog. Oceanogr. 116, 49–65.

Wornell, G., 2016. Inference and Information. Lecture Notes for MIT Course 6.437 in Spring 2013. MIT.