Journal of Marine Systems 96–97 (2012) 32–47

Contents lists available at SciVerse ScienceDirect

Journal of Marine Systems

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

Estimating time-dependent parameters for a biological ocean model using an emulator approach

Jann Paul Mattern [a,b,*], Katja Fennel [b], Michael Dowd [a]

[a] Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada
[b] Department of Oceanography, Dalhousie University, Halifax, Nova Scotia, Canada

ARTICLE INFO

Article history:
Received 30 September 2011
Received in revised form 24 January 2012
Accepted 29 January 2012
Available online 10 February 2012

Keywords:
Statistical emulator
Polynomial chaos
Parameter estimation
Data assimilation
Time-dependent parameters
Biological model
3D ocean model

ABSTRACT

We use a statistical emulator technique, the polynomial chaos expansion, to estimate time-dependent values for two parameters of a 3-dimensional biological ocean model. We obtain values for the phytoplankton carbon-to-chlorophyll ratio and the zooplankton grazing rate by minimizing the misfit between simulated and satellite-based surface chlorophyll. The misfit is measured by a spatially averaged, time-dependent distance function. A cross-validation experiment demonstrates that the influence of outlying satellite data can be diminished by smoothing the distance function in time. The optimal values of the two parameters based on the smoothed distance function exhibit a strong time-dependence with distinct seasonal differences, without overfitting observations. Using these time-dependent parameters, we derive (hindcast) state estimates in two distinct ways: (1) by using the emulator-based interpolation and (2) by performing model runs with time-dependent parameters. Both approaches yield chlorophyll state estimates that agree better with the observations than model estimates with globally optimal, constant parameters. Moreover, the emulator approach provides us with estimates of parameter-induced model state uncertainty, which help determine at what time improvement in the model simulation is possible. The time-dependence of the analyzed parameters can be motivated biologically by naturally occurring seasonal changes in the composition of the plankton community. Our results suggest that the parameter values of typical biological ocean models should be treated as time-dependent and will result in a better representation of plankton dynamics in these models. We further demonstrate that emulator techniques are valuable tools for data assimilation and for analyzing and improving biological ocean models.

1. Introduction

Simple models are often considered advantageous over more complex ones, because they tend to be easier to interpret and to calibrate and less expensive computationally. Their low complexity is typically achieved by combining many properties of the simulated system into single model variables and averaging them in time and space. In the context of biological ocean models, a good example for this is the blending of many plankton species into functional groups or often even into bulk model variables for phytoplankton and zooplankton (so called NPZD-class models). In the bulk variable treatment, each variable represents a large variety of real species with a range of specific physiological characteristics (e.g. different growth and nutrient uptake rates, different carbon-to-chlorophyll ratios). Since the abundance of these species and their relative contribution to the plankton community changes in space and time, so should the physiological characteristics of the bulk variables. In this study, we find evidence for temporal and spatial dependence of the parameters of a biological model that contains just two plankton variables, suggesting that using static parameters is overly simplified and suboptimal. Using an emulator approach in combination with a temporally and spatially dense set of satellite observations we can effectively infer parameter values that evolve in time and space and lead to an improved representation of plankton in the model.

Many studies have employed data assimilation in the context of biological models, often in order to optimize the poorly known parameters but also to update the model state and improve the models' forecast abilities. The techniques used in these studies can be divided into 3 broad categories: (1) variational techniques, such as 3DVAR and 4DVAR (e.g. Lawson et al., 1996; Powell et al., 2008; Smedstad and O'Brien, 1991), (2) Monte-Carlo based techniques which include the ensemble Kalman filter (e.g. Allen et al., 2003; Evensen, 2003; Hu et al., 2012), particle filter methods (Dowd, 2011; Losa et al., 2003; Mattern et al., 2010a) and Markov chain Monte Carlo methods (e.g. Dowd, 2007; Jones et al., 2010), and (3) emulator techniques. Emulators differ from the aforementioned techniques in that they effectively replace computationally expensive model simulations with fast approximations. Emulators require a set of model simulations for

* Corresponding author at: Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada.
E-mail address: Paul.Mattern@dal.ca (J.P. Mattern).

0924-7963/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jmarsys.2012.01.015

specific values of the stochastic inputs (parameters), then approximate unknown model output based on these simulations. The approximation is used in place of the true model output, eliminating the need for additional model simulations. This property makes emulators more efficient than other approaches, especially Monte Carlo techniques which rely on ensemble generation through random sampling and therefore generally require considerably more model simulations (Rougier and Sexton, 2007).

The emulator approach that we use in this study is the polynomial chaos expansion, which was first introduced by Wiener (1938) and later extended (Askey and Wilson, 1985; Wan and Karniadakis, 2006). Polynomial chaos relies on a set of orthogonal polynomial basis functions for the approximation of model results. The method has been applied widely in physical sciences (see Xiu and Karniadakis, 2003, for an overview), with only few applications in an oceanographic context (Lucas and Prinn, 2005; Thacker et al., 2012). Emulators applied in oceanographic contexts include emulators based on Gaussian process models (Scott et al., 2011) and other techniques (Frolov et al., OCT, 2009; Hooten et al., 2011). To our knowledge, emulator approaches have been used in the context of biological ocean models in only one study by Hooten et al. (2011) where 7 biological parameters are estimated. We focus this study on just 2 biological parameters, but employ the emulator to estimate their time-dependence in order to achieve a better representation of plankton dynamics in the model and an enhanced understanding of the biological model dynamics. We further use the emulator to obtain improved state estimates in an efficient manner.

Previously, two approaches have emerged to better represent the diversity of plankton. The first approach is to divide planktonic species into functional groups so that each plankton variable represents a more homogeneous and functionally distinct group of fewer species. While the simpler NPZD-class models contain only one phytoplankton and one zooplankton variable (e.g. Doney et al., 1996; Fennel et al., 2008; Franks and Chen, 1996), many biological models include two or more phytoplankton variables distinguishing, for example, between small and large phytoplankton, diatoms, diazotrophs etc. (e.g. Aumont et al., 2003; Gregg et al., 2003; Lehmann et al., 2009; Moore et al., 2001). One obvious limitation to adding more and more functional groups is that the number of poorly known parameters necessary for describing the biological interactions between functional plankton groups and different pools of other organic and inorganic matter increases dramatically (Denman, 2003) with consequent degradation of predictive skill.

In a recent, alternative approach, Follows et al. (2007) initialized a model with roughly 100 phytoplankton groups with their functional parameters drawn randomly from prescribed probability distributions. This approach allows for spatial and temporal variations in the self-organizing plankton community structure that emerges from local environmental conditions and competition (Goebel et al. (2010), see also review by Follows and Dutkiewicz (2011)) and represents a significant step toward a more flexible and realistic representation of plankton diversity in biological models. One drawback may be the large computational overhead required to carry on the order of 100 state variables.

We propose an alternative approach for the simulation of functional groups in biological models, namely incorporating variability or uncertainty by allowing the plankton parameters to be random variables. The main idea is that a small number of variables can achieve a more flexible representation of the plankton community, if their parameters are not fixed but stochastic properties governed by probability distributions. This approach effectively allows one phytoplankton variable to take on a range of different growth rates, sinking rates, etc. mimicking the behavior of different functional groups at different times. In combination with observations and a data-assimilative framework, the uncertainty in the model can be constrained by limiting the stochastic parameters to ranges that explain the observations best. We accomplish this using the emulator approach described above.

Most studies which combine biological modeling with the estimation of stochastic parameters treat influences such as the varying plankton assemblage as error terms (Dowd, 2011). In these cases one aims to find a static distribution for the stochastic parameter of interest. Stochastic parameters then induce uncertainty into the model state; the mean (or median) model state represents the best estimate of the true state, while its variance (or error estimate) captures the model uncertainty including the variations caused by changing plankton assemblages. Here, our approach is different: using a time-series of observations, we find the parameter values that best explain each observation. That is, parameter values are allowed to change in time and our best state estimate is the model state associated with the series of time-varying parameters.

For this purpose, we use a set of daily chlorophyll satellite images to obtain daily values for 2 parameters of the biological model. We find that there is a strong time-dependence in the optimal parameter values which follow a seasonal cycle. Chlorophyll state estimates derived from the time-varying parameters are significantly closer to observed chlorophyll values than those of a model simulation with optimized fixed parameters. The improvement remains significant in a cross-validation experiment which we performed to avoid overfitting the observations. This is evidence that the introduction of time-varying parameters can achieve a more realistic representation of the biological dynamics in a typical biological ocean model.

2. Methods

2.1. Biological model and parameters of interest

Our model domain is the Middle Atlantic Bight (MAB), a coastal region in the northwest Atlantic that stretches from Cape Cod in the north to Cape Hatteras in the south (Fig. 1). The model is based on the Regional Ocean Modeling System (ROMS; http://myroms.org, Haidvogel et al. (2008)) and consists of a physical model coupled with a biological component. Open boundary conditions for temperature, salinity, sub-tidal frequency velocity and sea level are taken from the larger-scale MAB and Gulf of Maine (MABGOM) regional model described in Chen and He (2010). Further details of the physical model are described in Hu et al. (2012). The biological component is described in Fennel et al. (2006); it simulates a simplified nitrogen cycle and has been employed successfully in various modeling studies (Bianucci et al., 2011; Fennel and Wilkin, 2009; Fennel et al., 2008; Previdi et al., 2009). The model contains one state variable each for phytoplankton and zooplankton, as well as variables for chlorophyll, nitrate, ammonium and small and large detrital nitrogen. Chlorophyll is simulated separately from phytoplankton to account for the effects of photoacclimation which allows phytoplankton species to regulate their chlorophyll content based on the availability of light and nutrients (Geider et al., 1998). Here, all model runs are for 1 year, starting on 1 January 2006 and ending on 31 December 2007. The initial and boundary conditions for the biological variables are taken from a larger scale model of the Northeast North American (NENA) shelf that uses the same biological component (Fennel et al., 2006) as described in Hu et al. (2012).

Despite the relative simplicity of the biological model with only two plankton variables, one for phytoplankton and one for zooplankton, it requires more than 30 physiological parameters for the biological dynamics. Here we focus on only two of these parameters: the maximum ratio of chlorophyll to phytoplankton carbon, and the maximum grazing rate of zooplankton. These two parameters were selected based on a sensitivity study where we compared the effect of variations in several candidate parameters on the chlorophyll field. Specifically, we performed 1-year simulations for a baseline parameter set and for parameter sets where one parameter was doubled and

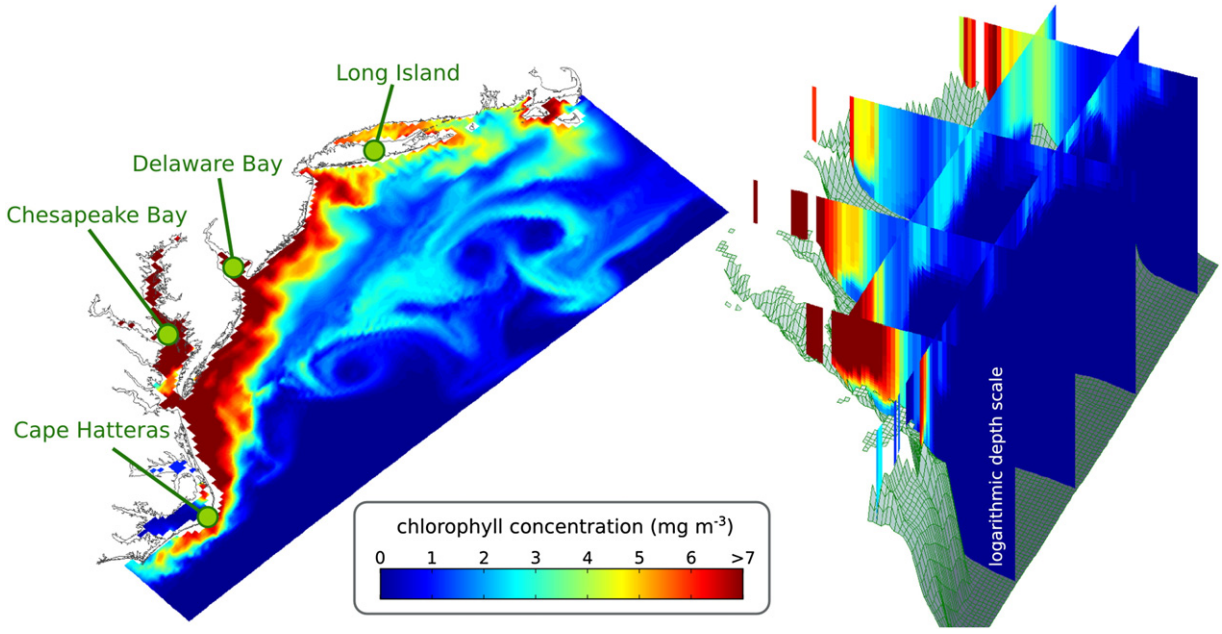

Fig. 1. Snapshot of the chlorophyll variable in the biological model. The left image shows the surface concentration, on the right multiple slices of the 3-dimensional chlorophyll field are placed into the bathymetry of the model region.

halved. The selection criterion is based on chlorophyll sensitivity because we use chlorophyll satellite observations (see Section 2.2 below).

The most sensitive parameters are the maximum ratio of chlorophyll to phytoplankton carbon and a parameter controlling the grazing rate of zooplankton. From here onward, we will refer to the maximum ratio of chlorophyll to phytoplankton carbon as θ_1 and the maximum grazing rate of zooplankton as θ_2. The physiological parameter θ_1 sets an upper limit on the concentration of chlorophyll relative to phytoplankton biomass. In the model equations (Fennel et al., 2006), the fraction of phytoplankton growth that is devoted to chlorophyll synthesis, ρ_{Chl}, is a function of θ_1:

$$\rho_{\mathrm{Chl}}(\theta_1) = \theta_1 \frac{\mu X_{\mathrm{Phy}}}{\alpha I X_{\mathrm{Chl}}}.$$

Here, X_{Phy} and X_{Chl} are the phytoplankton and chlorophyll variables respectively and $\frac{\mu X_{\mathrm{Phy}}}{\alpha I X_{\mathrm{Chl}}}$ is the ratio of achieved-to-maximum potential photosynthesis (Geider et al., 1997). The parameter θ_2 controls the growth and abundance of zooplankton, which interacts with and is strongly dependent on the concentration of phytoplankton. It scales the zooplankton grazing rate g according to:

$$g(\theta_2) = \theta_2 \frac{X_{\mathrm{Phy}}^2}{k_{\mathrm{P}} + X_{\mathrm{Phy}}^2},$$

where k_{P} is the half-saturation concentration of phytoplankton ingestion. The model equation that contains the sources and sinks of chlorophyll incorporates both $\rho_{\mathrm{Chl}}(\theta_2)$ and $g(\theta_2)$, in its full form it is:

$$\frac{\partial X_{\mathrm{Chl}}}{\partial t} = \underbrace{\rho_{\mathrm{Chl}}(\theta_1)\mu X_{\mathrm{Chl}}}_{\text{growth}} - \underbrace{g(\theta_2) X_{\mathrm{Zoo}} \frac{X_{\mathrm{Chl}}}{X_{\mathrm{Phy}}}}_{\text{grazing}} - \underbrace{m_{\mathrm{P}} X_{\mathrm{Chl}}}_{\text{mortality}} - \underbrace{\tau\left(X_{\mathrm{SDet}} + X_{\mathrm{Phy}}\right) X_{Chl}}_{\text{aggregation}}.$$

Here X_{Zoo} and X_{SDet} are the zooplankton and the small detritus variables respectively; the constants mp and τ are mortality and aggregation parameters. Since both θ_1 and θ_2 directly scale major growth and loss terms, it is not surprising that variations in either parameter have a strong effect on the chlorophyll concentration.

2.2. Chlorophyll observations and model-data comparison

Observations are essential in model calibration, optimization and validation. In all cases model output is compared to observations or made to fit the observations according to a chosen criterion. Thus both the observations and the choice of criterion can affect the results. Mattern et al. (2010b) formulated and analyzed several criteria tailored to model-data comparisons of satellite observations, and suggested a new measure, the "adapted gray block distance" (AGB) as preferable over more commonly used measures such as the root mean square error. For the calculation of AGB, two images are compared at different resolution levels by dividing them into subsequently smaller, square blocks and determining the average intensity value for each block. For each resolution level, from the coarsest resolution where one block encompasses the entire image, up to the finest where each block is made up of a single pixel, the root mean square error is determined, weighted and summed, resulting in the AGB distance value. When comparing an image derived from the model to an observation image, the comparison at multiple resolutions can be advantageous when noise is present in the observations and there are spatial offsets in the images (Mattern et al., 2010b). The AGB is also adapted to deal with missing values in images. Because of these qualities, we make use of AGB for the remainder of this study. Note however, that the methods described in this study do not require the use of AGB, and that any suitable model-data distance measure can be substituted.

The observations used in this study are daily images of surface chlorophyll concentrations derived from the SeaWiFS satellite for the year 2006 (350 images are available). Each image represents a daily average of one or more satellite scenes that have been interpolated onto the model grid. Due to clouds and other effects that impair the view of the optical satellite sensors, large portions of the images may be missing (compare, e.g., the sample satellite images in Fig. 2). In addition, noise is present in our satellite data set and especially evident in coastal regions (see, e.g. the average chlorophyll development of the data in the estuaries in Fig. 7). The same observational

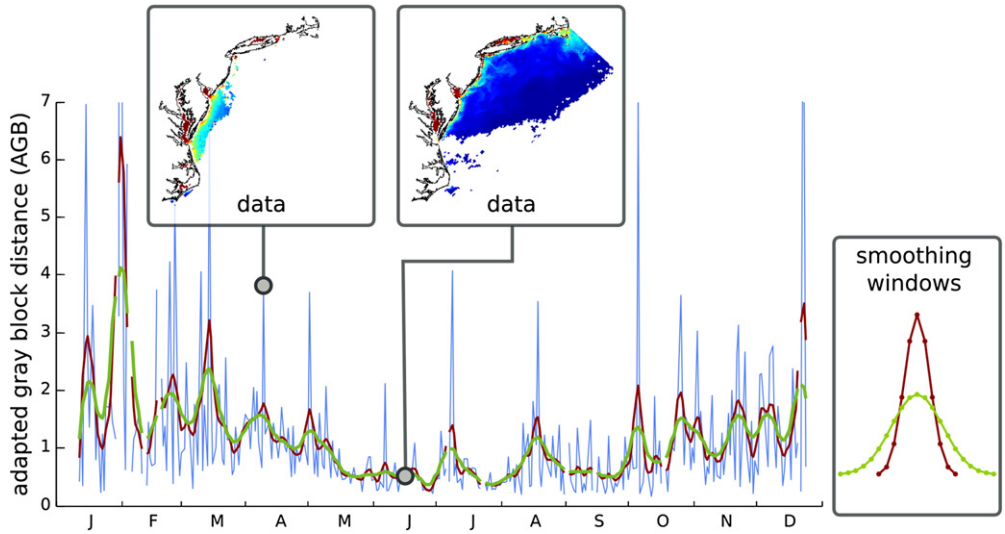

Fig. 2. The time-dependent distance function for the optimal fixed parameter set and two smoothed versions of it. The two smoothed curves correspond to smoothing intensities of 5 and 10 (dark red and green, respectively); the Gaussian windows with which the smoothing was performed are shown on the right. Two data images illustrate one point of the distance function with a high value and many missing values as well as one with a low distance function value and few missing values. High distance values tend to be caused by little available data.

data set used here was also used in Hu et al. (2012) and is described in more detail there.

We denote the distance value of AGB when comparing the satellite image at time index t with the corresponding model chlorophyll field as

$$\mathrm{d}(t,\theta_1,\theta_2) \text{ for } t = 1,2,\ldots,n_{\mathrm{obs}}. \tag{1}$$

Here, n_{obs} is number of (not necessarily equally spaced) time steps with available observations (in our case $n_{\mathrm{obs}=350}$). The dependence of d on the parameters θ_1 and θ_2 stems from the influence of both parameters on model chlorophyll.

2.3. The emulator: the polynomial chaos expansion

Polynomial chaos is an approach to quantifying how uncertainty in model inputs relates to uncertainty in its outputs. Like other emulator approaches it uses deterministic model runs given specific values of the uncertain inputs (i.e. the model's parameters, initial or boundary conditions, all of which will be referred to as parameters in the following). The resulting model output for these specific parameter values is then interpolated in parameter space to obtain approximations of the model output for all parameter values within the considered range. Since each uncertain input has a probability distribution (a prior distribution which must be specified) every model output that is dependent on the uncertain inputs must also have a distribution (induced by the uncertainty in the input). The polynomial chaos expansion provides a framework with which the properties of the distribution of any output value, such as the mean and variance of the distribution, can be approximated easily.

As the name suggests, polynomial chaos performs a polynomial interpolation in parameter space. This feature becomes useful in cases where one wants to obtain an estimate of the model output for a parameter value without performing additional model simulations. Using polynomial chaos, one can estimate any model output for the parameter values of choice. These outputs could range from the phytoplankton concentration in a given grid cell to the entire 3-dimensional chlorophyll field of the model. The interpolation feature of the polynomial chaos expansion can also be used to approximate other functions which depend on the uncertain inputs, e.g. we use it here to approximate the distance function in Eq. (1).

A short introduction to the polynomial chaos theory follows (for recent, more detailed studies see Xiu and Karniadakis (2003) and Marzouk and Najm (2009)). Since our focus is on stochastic parameters we do not discuss other uncertain inputs such as initial or boundary conditions. While we include 2 stochastic parameters, the methodology is described here for only one stochastic parameter θ. By assuming independent stochastic parameters, the theory translates in a straightforward manner into multidimensional parameter space (Xiu and Karniadakis, 2002).

Here, we let the function $f(\mathbf{x},t,\theta)$ be our property of interest, f can represent any model output or a function thereof (e.g. our distance measure in (1) which is a function of the model's chlorophyll output). The function f may be dependent on space $\mathbf{x}$, time t and the uncertain parameter θ. In the polynomial chaos expansion f is approximated by a basis function expansion:

$$f(\mathbf{x},t,\theta) = \sum_{k=0}^{k_{\max}} a_k(\mathbf{x},t)\phi_k(\theta) + \epsilon_{\mathrm{trunc}}(\theta) \tag{2}$$

where $a_k(\mathbf{x}, t)$ are expansion coefficients, independent of the uncertain input θ, and the kth basis function $\phi_k(\theta)$ is a polynomial of order k in the parameter space defined by θ. The parameter $k_{\max}$ is the maximum order of polynomials used in the approximation and determines the quality of the approximation and $\epsilon_{\mathrm{trunc}}$ is the truncation error. Without cutoff, i.e. for $k_{\max}=\infty$, the approximation is exact and $\epsilon_{\mathrm{trunc}}(\theta)=0$. However, the number of required model runs grows with $k_{\max}$, so that computational constraints force us to use relatively small values in typical applications.

The choice of polynomials in Eq. (2) is dependent on the probability density function of the parameter θ which we denote $p(\theta)$. The polynomials are chosen to be orthogonal with respect to p, so that

$$\int_S \phi_k(\theta)\phi_i(\theta)p(\theta)\mathrm{d}\theta = \delta_{k,i}N_k. \tag{3}$$

Here S is the support of p (the region where $p(\theta)>0$); the Kronecker delta function $\delta_{k,i}$ is equal to 1 if $k=i$ and 0 otherwise; $N_k = \int_S \phi_k(\theta)^2 p(\theta)\mathrm{d}\theta$ is a normalization factor specific to the kth polynomial and independent of θ. All common distributions have well known sets of polynomial basis functions (Xiu and Karniadakis, 2002) and polynomial chaos can be generalized further to accommodate arbitrary distributions of θ (Wan and Karniadakis, 2006). For

Table 1
The first 7 Legendre polynomials ϕ_k and their associated normalization factors N_k.

k	$\phi_k(\theta)$ for $\theta \in [-1,1]$	N_k
0	1	1
1	1θ	$\frac{1}{3}$
2	$\frac{1}{2}\left(3\theta^2-1\right)$	$\frac{1}{5}$
3	$\frac{1}{2}\left(5\theta^3-3\theta\right)$	$\frac{1}{7}$
4	$\frac{1}{8}\left(35\theta^4-30\theta^2+3\right)$	$\frac{1}{9}$
5	$\frac{1}{8}\left(63\theta^5-70\theta^3+15\theta\right)$	$\frac{1}{11}$
6	$\frac{1}{16}\left(231\theta^6-315\theta^4+105\theta^2-5\right)$	$\frac{1}{13}$

example, the corresponding set of orthogonal polynomials for a θ with uniform distribution, which we will use in this study (see Section 2.4), are the Legendre polynomials and ϕ_k is the kth Legendre polynomial. The first 7 Legendre polynomials and their associated normalization factors are listed in Table 1.

To perform the basic polynomial chaos approximation in Eq. (2), one needs to compute the coefficients a_k. They are given by

$$a_k(\mathbf{x},t) = \frac{1}{N_k}\int_S f(\mathbf{x},t,\theta)\phi_k(\theta)p(\theta)\mathrm{d}\theta, \tag{4}$$

which is approximated by a Gaussian quadrature as (Xiu and Karniadakis, 2002):

$$a_k(\mathbf{x},t) \approx \frac{1}{N_k}\sum_{i=0}^{k_{\max}} f\left(\mathbf{x},t,\theta^{(i)}\right)\phi_k\left(\theta^{(i)}\right)\omega_i. \tag{5}$$

Here $\theta^{(i)}$ is a quadrature point in parameter space and given by the roots of $\phi_{k_{\max}+1}$ and the scalars ωi are Gaussian quadrature weights (both are dependent on the choice of the distribution of θ and the parameter $k_{\max}$). Table 2 contains the quadrature points and their weights for uniform θ and Gauss–Legendre quadrature with $k_{\max}=6$. From a computational perspective, it is important to note that the computation of the coefficients a_k requires the computation of $f\left(\mathbf{x},t,\theta^{(i)}\right)$ at each quadrature point $\theta^{(i)}$ for $i=0,1,\ldots,k_{\max}$. In other words, $k_{\max+1}$ model runs are needed. Increasing the precision of the approximation by increasing $k_{\max}$ by one, therefore comes at the cost of an additional model run.

One advantage of polynomial chaos lies in the straightforward way in which the uncertainty in the input (the stochastic parameter θ) translates into the output (f). Due to the orthogonality of the polynomials, expected value and variance of f conditional on the distribution of θ are straightforward to calculate once the coefficients a_k have been computed. Conditional expectation and variance are given by

$$\mathbb{E}(f(\mathbf{x},t,\theta)|\theta) = a_0(\mathbf{x},t) \quad \text{and} \quad \mathrm{var}(f(\mathbf{x},t,\theta)|\theta) = \sum_{k=1}^{n} a_k^2(\mathbf{x},t)N_k. \tag{6}$$

They represent the mean and variance of f introduced by the variation of θ. To obtain good estimates of the full (unconditional) variance of f, e.g. for the purpose of creating estimates of model error, it is important to capture all the error of the uncertain inputs and to choose appropriate prior distributions for the inputs.

As mentioned, the above equations feature only one stochastic parameter θ. When expanded to more than one parameter, the computational cost for polynomial chaos increases exponentially with the number of stochastic parameters. For example, when including n_θ stochastic parameters to be approximated using polynomials of order $k_{\max}$, $(k_{\max}+1)^{n\theta}$ model runs are required. Furthermore, if one desires to increase the order of polynomials, the quadrature points change, so that completely new model runs will have to be performed. However, it should be noted that the model simulations are only performed once prior to any attempts at inference.

Table 2
The quadrature points $\theta^{(i)}$ and associated weights ω_i for Gauss Legendre quadrature of maximum order $k_{\max}=6$.

i	1	2	3	4	5	6	7
$\theta^{(i)}$	−0.9491	−0.7415	−0.4058	0	0.4058	0.7415	0.9491
ω_i	0.1295	0.2797	0.3818	0.4180	0.3818	0.2797	0.1295

2.4. Polynomial chaos setup and approximation

When implementing polynomial chaos, the factors that need careful considerations are (1) the choice of uncertain model inputs (parameters), (2) the prior distributions assigned to these inputs, and (3) the highest order of polynomials $k_{\max}$ for each input. In an ideal scenario, one would take a fully Bayesian approach, that is treat all inputs that are not completely known as uncertain and incorporate them into the polynomial chaos procedure. However, complex models such as 3-dimensional ocean models have a large number of inputs that are not fully known, e.g. many parameters, physical forcing, boundary conditions, etc. To incorporate all these sources of uncertainty into the polynomial chaos expansion would necessitate a large number of model runs and prove to be infeasible using current computing resources.

Here, we undertake a targeted study focused on just two biological parameters. Once the uncertain inputs are selected, assigning a prior distribution to the inputs requires careful consideration, as one typically has little knowledge of the uncertainty (or error) of the inputs. Often, and the case in this study, one bases the prior distribution on previous experiments, literature values or educated guesses. Lastly, in the choice of $k_{\max}$, one is again limited by computational resources and faced with a trade-off between precision and number of model runs. The optimal choice is dependent on the problem; in this study we found that the functions and fields we chose to interpolate were well approximated by polynomials of order 6 (see below).

For this study, the two parameters θ_1 and θ_2 (see Section 2.1) are considered to be stochastic. As the prior distribution for θ_1 and θ_2 we used a uniform distribution and set the lower and upper limits of the distribution as 0.25 and 1.75 times the parameters' standard value, respectively. The standard values are taken from Fennel et al. (2006) and turned out to be reasonably close to the optimal (fixed) parameter set for this study (see Fig. 3, Section 3.1.2). We chose the uniform distribution because of its finite support which does not permit negative parameter values, as well as yielding a simple polynomial chaos setup.[1] Finally, we selected the maximum order $k_{\max}=6$ for both parameters. As a result $(k_{\max}+1)^2=49$ model runs had to be performed. The 7×7 grid of quadrature points in parameter space is shown in Fig. 3.

After performing the necessary model runs, polynomial chaos allows for the approximation of any function that is dependent on the stochastic parameters. It can therefore be used to approximate the distance function in (1) for the purpose of model-data comparison. Here, d takes on the role of f in Eq. (2), i.e. we set $f(\mathbf{x},t,\theta_1,\theta_2)=\mathrm{d}(t,\theta_1,\theta_2)$. As described in Section 2.3, we then perform the following steps to approximate d. After the model is run for the parameter values of each quadrature point, the distance

[1] While model uncertainty estimates might benefit from a different parameter distribution, this study relies on the polynomial interpolation aspect of polynomial chaos which is not very sensitive to changes in the distribution. Polynomial interpolation is exact at the quadrature points and a change in distribution affects the layout of the quadrature points in parameter space. Only a drastic change in the quadrature point layout can cause a strong effect on the polynomial interpolation but such a change would need to be caused by an equally drastic change in the parameter distribution, e.g. a strong shift in the range of the uniform distributions.

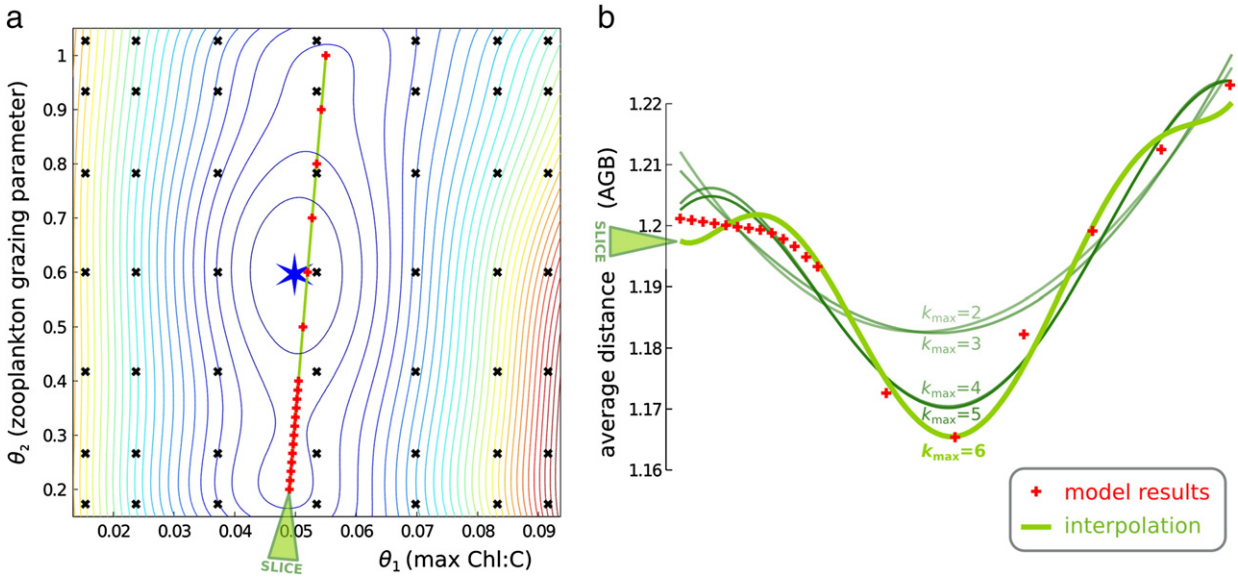

Fig. 3. The interpolated time-averaged distance function. Image (a) shows a contour plot of the distance function with quadrature points (black crosses) and the global minimum (blue star) which corresponds to the optimal fixed parameter set. A slice through the interpolated function in (a) is shown in (b) as a green line in comparison to model results (red pluses). The slice in (b) also illustrates the effect of lowering k_{max}, thereby using fewer basis functions to approximate the average distance function.

function is computed for each of the model runs at all time steps from 1 to n_{obs}. The expansion coefficients $a_k(t)$ specific to the distance function are then computed using Eq. (6). As the distance function d is not dependent on **x** the coefficients a_k do not depend on **x** either. Now we can use the approximation in Eq. (2) to approximate the distance function for each value of θ_1 and θ_2 in their respective ranges.

Approximating multi-dimensional fields such as the surface chlorophyll (as done in our emulation experiment in Section 3.2.1 below) works in a similar way. The surface chlorophyll values in the topmost model layer are extracted for all model runs at all time steps. The extracted output, which is dependent on θ_1, θ_2, t and the two horizontal spatial coordinates contained in **x**, is set equal to $f(\mathbf{x}, t, \theta_1, \theta_2)$. Surface chlorophyll specific coefficients $a_k(\mathbf{x}, t)$ are computed which are, like the surface chlorophyll field, dependent on the spatial coordinates **x**. No recomputation of the polynomials $\phi_k(\theta)$ is necessary to obtain approximate surface chlorophyll values from Eq. (2).

3. Results

We hypothesized that temporal changes in plankton species composition manifest as shifts in the parameter values of our biological model. In other words, we expect that parameter values that shift in time and space will better explain the observations. The polynomial chaos expansion allows us to obtain approximations of model output for any parameter value within prescribed bounds. This property allows us to find optimal parameter values with only a limited number of computationally costly model runs. Specifically, we employed the polynomial chaos expansion to approximate the distance between observed and simulated surface chlorophyll. First, we minimized the distance for the entire data set to obtain global optimal parameters independent of time and space (referred to as *optimal fixed* parameters in the following). Then we minimized the chlorophyll distance for single (daily) observations individually and for different model regions to identify temporal and spatial variations in the optimal parameter values. Both optimizations are based on the same set of 49 model runs and further require only the computationally much less demanding evaluations of the polynomial chaos-based interpolation.

3.1. Interpolating the model-data distance function and parameter estimation

3.1.1. Smoothing the distance function

We obtained estimates of optimal parameter values by interpolating and minimizing the time-dependent distance function d in (1). This function appears to be very noisy and varies considerably from one day to the next (Fig. 2), not necessarily due to bad model output on days with large values of d, but because of outliers in the observations caused by a large number of missing values (Fig. 2).

In order to diminish the influence of outliers in our analysis and to create a more robust distance function, we used a low-pass filter in the form of a Gaussian window to smooth d. From here on, we use the term *smoothing intensity* to describe the amount of smoothing that was applied to the distance function. The smoothing intensity is a positive integer value which increases with the amount of smoothing. More precisely, twice the smoothing intensity plus 1 is the width of the Gaussian smoothing window in days (we only use window widths that are odd), i.e. a smoothing intensity of 0 refers to a window width of $2\times0+1=1$ and therefore no smoothing, while a smoothing intensity of 10 refers to a window width of $2\times10+1=21$. Examples of the smoothed distance function and the corresponding Gaussian windows are shown in Fig. 2. For simplicity, we do not remove any of the data outliers from our analysis, eliminating the need to create an objective criterion for their removal.

The objective of smoothing the distance function is to minimize the impact of outliers, reduce overfitting and to improve the parameter optimization.

3.1.2. Optimal fixed parameters

Typical parameter optimization studies assume fixed parameter values, and the optimized parameters are determined by minimizing the model-data discrepancy over the full set of available observations. We can do the same using the polynomial chaos expansion: To obtain estimates of the optimal fixed parameter values with respect to the distance function $d(t, \theta_1, \theta_2)$ in Eq. (1), we eliminated the time dependence of d by computing its average in time. We then used the polynomial chaos expansion to approximate the resulting average distance function in parameter space as detailed in Section 2.4. The resulting distance function is smooth and exhibits a clearly defined

global minimum close to the center of the domain defined by the ranges of θ_1 and θ_2 (Fig. 3). Because the average distance function changes more along the θ_1-axis than in the direction orthogonal to it, we can deduce that model chlorophyll is more sensitive to relative changes in the value of the maximum chlorophyll to carbon ratio (θ_1) than the zooplankton grazing parameter (θ_2) and there appears to be little dependency between the parameters.

In order to gauge the quality of the polynomial interpolation, we performed a number of analysis model runs along a slice through the parameter domain (green line in Fig. 3(a)). A comparison of the approximated distances (light green line in Fig. 3(b)) with the exact distances obtained for the analysis runs (red symbols in Fig. 3(b)) reveals that the average distance function is generally well approximated by the interpolation for $k_{\max}$ with only some edge effects typical of polynomial approximations. This leads us to conclude that the position of the global minimum of the average distance function in Fig. 3(a) represents a good approximation of the optimal parameter values with respect to our full data set. In the following we will refer to the parameter pair that minimizes the interpolated average distance function as the optimal fixed parameters.

The analysis model runs can also help us assess the convergence of the polynomial chaos approximation. We chose $k_{\max}=6$ for the approximation in Eq. (2). The effects of truncating the sum at lower orders (smaller values for $k_{\max}$) are shown in Fig. 3(b). The results of the analysis model runs remain fairly well approximated for $k_{\max}\geq 4$, but below that, the approximation becomes considerably worse. Interestingly, the position of the minimum changes relatively little with the addition of higher order polynomials. For our purposes, the position of the minimum of the average distance function is approximated well and choosing a higher $k_{\max}$ at the cost of additional model runs appears unnecessary.

For a different data set or a subset of our data, the average distance function and the position of its minimum is likely to change. It is desirable to gain an understanding of the uncertainty in the position of the global minimum given in Fig. 3. For this purpose, we performed a bootstrapping experiment: We generated subsets of the observations (the bootstraps) by randomly selecting a fixed number of satellite images from the 350 images that make up the complete observational data set. For each bootstrap, we calculate the global minimum of the respective time-averaged distance function. For the relatively large bootstrap size of 200 images, drawn without replacement, we see a tight clustering of minima around the full data minimum (Fig. 4(a)). With a decrease in bootstrap size, the range becomes greater, especially along the θ_2 axis. At the small bootstrap size of 10, the minimum positions are distributed all along our selected range of θ_2 (Fig. 4(d)).

It is apparent that the optimal fixed parameter set is very much dependent on the subset of data used in the optimization exercise and can vary considerably based on its choice. In the following sections, we show that this dependence is mainly due to an underlying time-dependence of the optimal parameters and not primarily due to the noise contained in our data set.

3.1.3. Time-varying parameters

Time-dependence of the optimal values of the physiological parameters θ_1 and θ_2 would hint that there is a signal in the observations that the model cannot account for if the parameter values are fixed. To uncover time-dependence, we return to the polynomial chaos approximation of the distance function. In the previous section we used it to obtain a set of optimal fixed parameters for our entire data set by minimizing the average distance function. Using a very similar procedure, we can approximate the distance function for each daily observation to obtain a set of optimal parameters for each day. In other words, we used the polynomial chaos expansion to interpolate $\mathrm{d}(t,\theta_1,\theta_2)$ in parameter space for $t=1,\ldots,n_{\mathrm{obs}}$ and determine the global minimum of the function for each t. We performed this procedure for the unsmoothed version of d as well as for versions that have been smoothed at different intensities as described in Section 3.1.1. Then we arranged the resulting parameter values for each smoothing intensity into a parameter path in time, as shown in Fig. 5. The path corresponding to the unsmoothed distance function appears very jagged, dominated by high frequency variation and with little structure; as the temporal smoothing increases the paths become more structured and a loop emerges.

The structure of the parameter paths can be interpreted in a straightforward way. With no smoothing, the procedure picks the optimal parameter set to match one satellite image alone, including the noise contained within the image. The distance to the previous or following image is not considered. As the distance function is noisy (see Fig. 2), we expect a high amount of noise in the daily optimal parameter values as well. The high frequency variations in the daily optimal parameters are therefore likely local fits to the noisy data. However, Fig. 5 also shows clear evidence of a low frequency parameter change visible at higher smoothing intensities. This low frequency signal reveals that there is a time-dependence of the optimal parameter values that cannot be explained by the noise in the observations, indicating that the fit between model and observations can be improved by allowing parameters to follow the low frequency signal using cross-validation.

These results also suggest that there is an optimal smoothing intensity, strong enough to filter out the effects of the noise contained in the data, yet not too strong to also remove the low frequency signal we are interested in. In the following section, we show how the chlorophyll output of our model can be improved by using the low frequency parameter paths. Based on a comparison with the observations we also determine the optimal smoothing parameter that best isolates the low frequency signal.

3.2. Emulating surface chlorophyll

3.2.1. Polynomial chaos-based emulation

In the previous section, we described how time-dependent parameter paths can be obtained from the interpolation of the likewise time-dependent distance function. Here we utilize these paths to obtain improved model estimates of surface chlorophyll fields. We use the polynomial chaos expansion as an emulator, i.e. a system that allows us to obtain estimates of the state of the ocean for a parameter combination we did not perform a model run for. In our case, we emulate the full surface chlorophyll field using the polynomial chaos based-interpolation.

As described in Section 2.4, the polynomial chaos expansion can be used to interpolate virtually any model output in parameter space, including the time-dependent chlorophyll concentrations in the surface layer of the model. This feature allows us to efficiently interpolate the chlorophyll values along the parameter paths. We obtained daily pairs of parameter values from one of the time-dependent parameter paths (see Fig. 5). Then, with the help of the polynomial chaos expansion, we estimate the surface chlorophyll fields that correspond to the daily parameter values. This procedure results in an emulated time-dependent surface chlorophyll field, which is dependent on the smoothing intensity that underlies the chosen parameter path. Note that one can use the same procedure to obtain estimates of depth-resolved chlorophyll fields or other biological properties along the parameter paths.

We then compared the interpolated chlorophyll fields to the observations as in previous sections, using the same distance measure d in Eq. (1) but replacing chlorophyll model output with the interpolated model chlorophyll fields. This way, we obtain a distance value for each day which, averaged in time, results in an average distance value. We computed average distance values for different smoothing intensities (blue diamonds in Fig. 6).

The resulting average distance values based on the emulation experiment are smallest for the parameter path without smoothing (Fig. 3(a)), and increase with more smoothing. They are directly

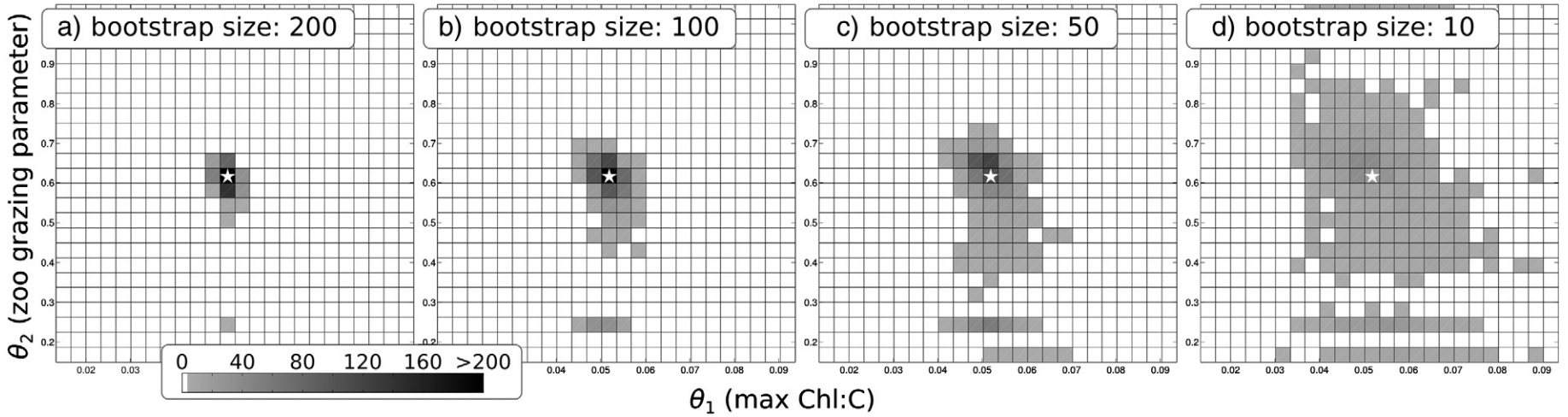

Fig. 4. Two-dimensional histogram of the position of the global minimum of the time-averaged distance function in parameter space (compare Fig. 2) for the bootstrapping experiment with 1000 bootstraps described in Section 3.1.2. With decreasing bootstrap size, the location of the minimum becomes more variable, especially along the axis corresponding to the parameter θ_2. The white star in the center of each image is a reference point.

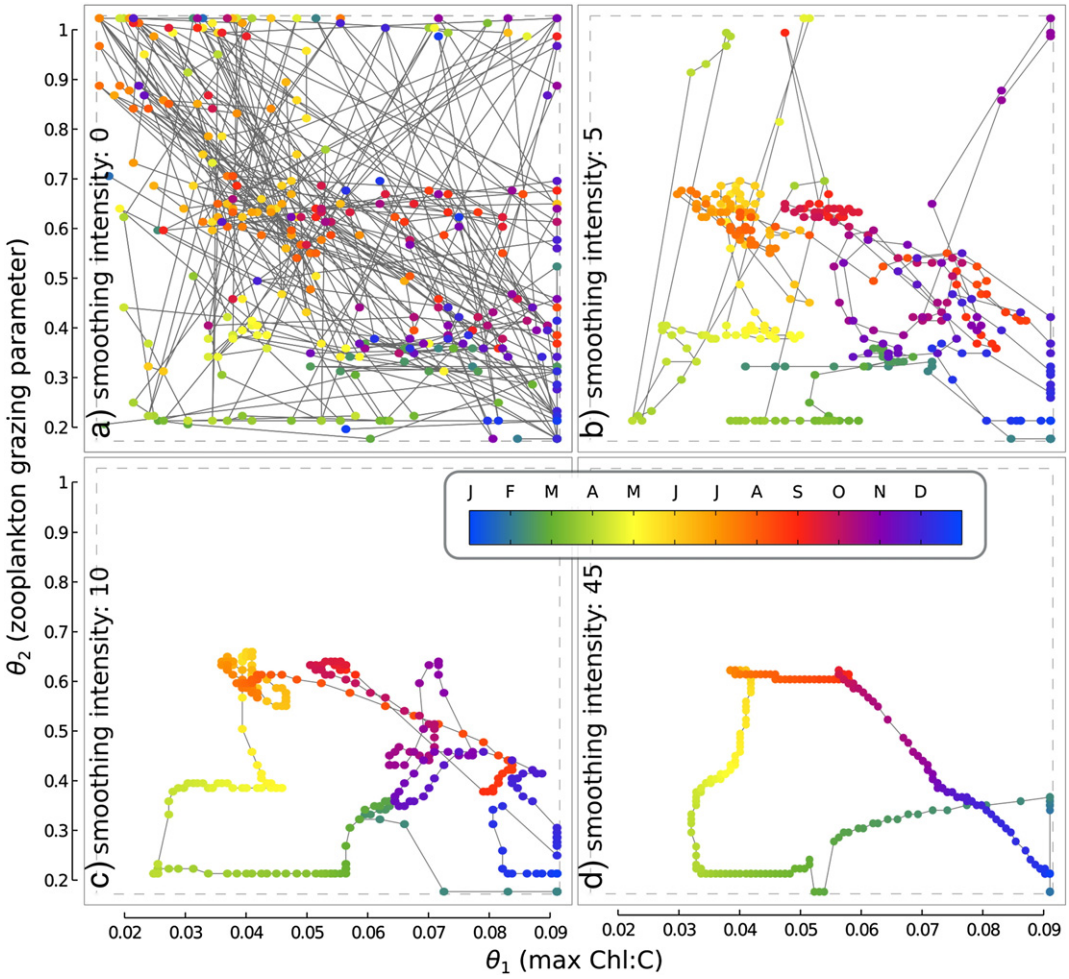

Fig. 5. The parameter paths obtained by minimizing the time-dependent distance function in each time step for 4 different smoothing intensities. The smoothing intensities (0, 5 and 10) in panels 1 to 3 correspond to those shown in Fig. 2.

comparable to the average distance value of the optimal fixed parameter model run from Section 3.1.2. For smoothing intensities up to 45 all average distances are well below the values of the optimal fixed parameter run (indicated by the dashed line in Fig. 6). In other words, the emulated chlorophyll fields are considerably better than those of any model run with fixed parameters, which was to be expected. The fact that the lowest distance is associated with the no-smoothing path, indicates that at least part of the improvement is due to overfitting the data. At low smoothing intensities the emulated values fit even outlying values and noise very well, completely disregarding the model dynamics.

In the following section we perform a cross-validation to address this issue and determine for which smoothing intensities overfitting is not a concern. The cross-validation also allows us to identify the optimal level of smoothing.

3.2.2. Choosing the optimal smoothing parameter in a cross-validation experiment

The jagged nature of the parameter paths at low smoothing intensities (Fig. 5(a)) indicates overfitting of the model to the observations. Cross-validation experiments provide us with a technique to distinguish overfitting from real improvement in model performance. We follow the typical approach where the observational data set is partitioned into two parts, the training set and the validation set. The training set is only used to optimize the model parameters, the quality of the model output is then assessed by a comparison with the validation set. Overfitting the training set will not lead to a better model performance with respect to the validation set.

We performed multiple cross-validation experiments in a bootstrap fashion. In each experiment, the observations were split into training and validation set in the following way. The training set contains the first and last satellite images as well as a number of randomly selected images in between; the validation set consists of the remaining images. We then performed 25 cross-validation experiments for each of five training set sizes (175, 150, 125, 100 and 75) ranging from half of our observational data set to roughly one fifth. In each experiment, we determined the parameter path according to the procedure described in Section 3.1.3, but only using the training data set. This way we obtained optimal parameter values corresponding to the time steps of the training data. We determined the quality of these parameter values with respect to the validation data set in 3 steps: (1) We linearly interpolated the parameter values corresponding to the training set dates in time to obtain the parameter values for the validation set dates. (2) We used the freshly obtained parameter values to interpolate the surface chlorophyll field in parameter space, yielding a surface chlorophyll field for each validation set date. (3) With our standard distance measure, we computed the distances of the surface chlorophyll fields to the validation data and calculated the average distance.

The average distance values obtained through the above procedure are shown as red dots in Fig. 6, and exhibit a clear difference compared to the values of the emulation experiment without cross-validation (Fig. 6, blue diamonds). First of all, the cross-validation distance values

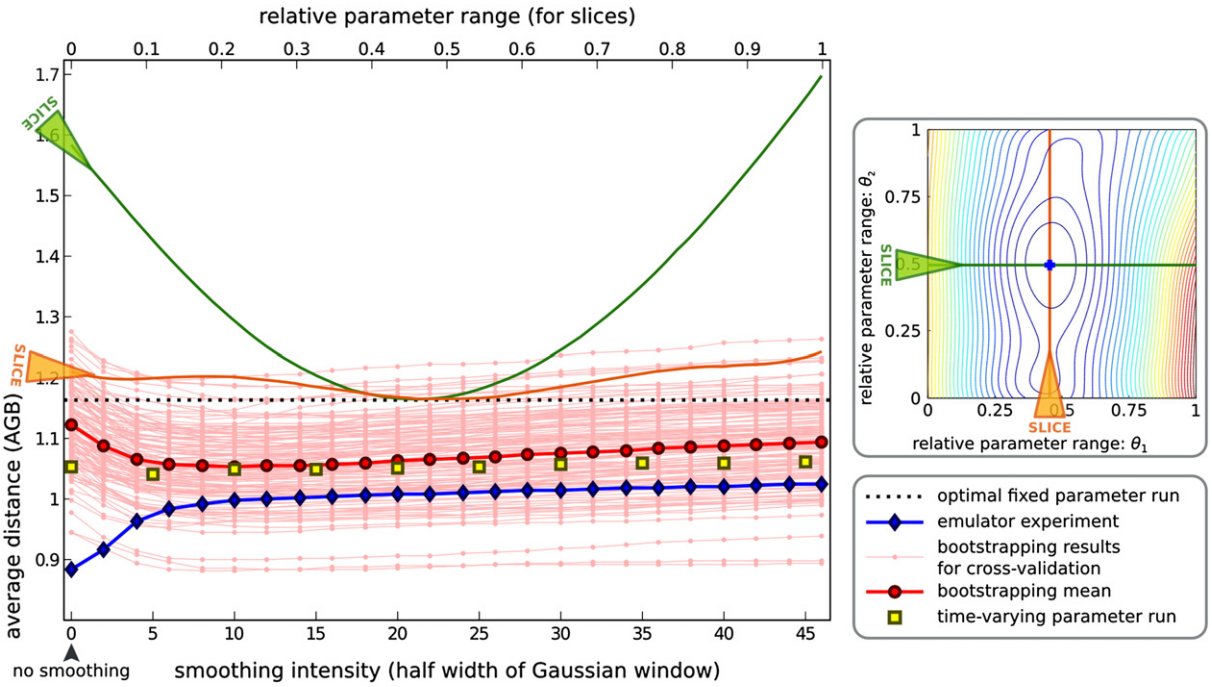

Fig. 6. The average distance values of various experiments: the emulation experiment (Section 3.2.1), the bootstrapping results for the cross-validation experiment (Section 3.2.2) and the time-varying parameter runs (Section 3.3), each dependent on the smoothing intensity. They are compared to the average distance value obtained by the optimal fixed parameter run (Section 3.1.2) which is independent of the smoothing intensity. For scale, two slices of the average distance function for fixed parameters (see Fig. 2) are displayed in the same plot.

are generally higher than those of the emulation experiment. This is to be expected from a cross-validation experiment which uses two separate data sets for optimizing parameters and assessing the fit. More important is another difference: While the setup without cross-validation has the lowest distance at a smoothing intensity of 0 and then increases steadily, the cross-validation mean has a minimum at a smoothing intensity of 10, corresponding to a smoothing window width of 21 days. The minimum is relatively flat toward higher smoothing intensities but shows a sharper incline for intensities lower than 5. This property is strong evidence for the presence of overfitting at low smoothing intensities. For no or little smoothing the jagged parameter path describes the noise in the observations and the parameter values do not generalize well to the validation data set in the cross-validation. As the smoothing is increased, overfitting becomes less of a problem and disappears. When smoothing is increased even further, useful information in the observations is filtered out so that average distances increase again, albeit at a slow rate. We therefore consider a smoothing intensity of 10, the position of the minimum of the cross-validation mean curve, the optimal smoothing intensity for our emulation experiment, and use it as the standard smoothing intensity for the emulation experiment in the following section.

In this section and the previous one, we have shown that the time-dependent parameter paths in combination with state interpolation can be used as an emulation tool that produces state estimates which are considerably better than those of any fixed parameter model run. The improvement is not due to fitting noise in the data, as the smoothing intensity can be adjusted to avoid the problem of overfitting; it is due to the presence of an underlying time-dependence or seasonal cycle in the parameters. In the following, we assess the utility of the parameter paths for obtaining time-dependent parameter values for our biological model.

3.3. Model runs with time-dependent biological parameters

In addition to obtaining improved estimates of chlorophyll by means of a polynomial chaos interpolation, the parameter paths can also be used in a more straightforward way. One can perform biological model runs with time-varying values of θ_1 and θ_2 by plugging parameter paths directly into the model. The values of the two parameters are taken from a specific parameter path and so the results are again dependent on the smoothing intensities used to obtain the path.

To implement time-varying parameters in our model we extended the parameter paths which are defined only for the discrete time steps $t=1,\ldots,n_{\mathrm{obs}}$, to the interval $[1,n_{\mathrm{obs}}]$ by linearly interpolating the paths in time. In the numerical model this was implemented by incorporating a simple lookup table for the parameter values at $t=1,\ldots,n_{\mathrm{obs}}$. At each model time step the model looks up the values of θ_1 and θ_2 that correspond to the two closest points in time and performs the time interpolation. By using different lookup tables, one can perform model runs for different parameter paths or smoothing intensities. We set the initial values of θ_1 and θ_2 to the first value of the parameter path and ran the model, keeping all other settings unchanged. Again, we computed the average distance values for the time-varying parameter runs (Fig. 6, yellow squares).

For the time-dependent parameter runs, the lowest average distance is achieved at a smoothing intensity of 5 (Fig. 6, yellow squares), although there appears to be no strong dependence on the smoothing intensity as all distance values are very closely grouped. Generally, the time-dependent parameter results are in between those corresponding to the emulation experiment described in Section 3.2.1 (Fig. 6, blue diamonds) and the optimal fixed model run results (Fig. 6, dashed black line). One would expect a degraded performance of the time-varying runs in comparison to the emulated results given the memory of the model with time-dependent parameters, an effect that we will reconsider in the discussion in Section 4. However, the improvement in average distance in comparison to the optimal fixed parameter run is still large. In comparison with the optimal fixed values of θ_1 and θ_2, the model creates considerably better chlorophyll output if we allow the values to change in time. As the model reacts relatively slowly to shifts in parameter values and the average distance values of the different time-varying runs are very

similar we can conclude that these runs do not overfit the data. To further assess the improvement we performed a follow up comparison of the estimated chlorophyll values using the 3 estimators: the optimal fixed parameter run, the time-dependent parameter run and the emulation experiment, in the following section.

3.4. Temporal and spatial analyses

3.4.1. Spatial comparison of chlorophyll estimates from model and emulator

In order to assess how the differences in average distance values for our 3 runs, the optimal fixed parameter run (Section 3.1.2), the time-dependent parameter run (Section 3.3) and the polynomial chaos-based emulation experiment (Section 3.2.1), translate into differences in surface chlorophyll we calculated the regional chlorophyll averages for 3 regions of the model domain, the estuaries, the coastal and the open ocean region (Fig. 7). In the estuaries, all model estimates of chlorophyll underestimate the observations (Fig. 7(a)). This result is not unexpected, as the relatively coarse resolution model cannot adequately represent estuarine dynamics. Additionally, satellite chlorophyll estimates might be biased due to high levels of colored dissolved organic matter in the water which are known to enhance the chlorophyll signal in satellite images (Mannino et al., 2008).

The model estimates agree better with the observations in the other two regions, the coastal region and the open ocean. In both regions it is also apparent that the time-varying parameter model run and the emulated state estimates show improvement over the optimal fixed parameter run. A look at the deviations from the data, shown in Fig. 8, reveals that surface chlorophyll estimates are indeed most accurate for the emulated state estimates, followed by the time-varying parameter model run and the fixed parameter model run. Improvement is especially evident in April, during the spring bloom. In a few instances, the fixed parameter model produces the lowest absolute residuals in some regions of the model. These are however offset by higher residuals in other regions (compare, e.g., the June residuals in Fig. 8 across all 3 regions). This demonstrates that there is no uniform improvement across the entire model domain, instead the improvement achieved by time-varying parameters depends on both time and space.

Generally, improvement is more likely where our parameter variation induces the greatest variance into the surface chlorophyll state. This observation follows from a comparison of the absolute residuals with the conditional variance (see Eq. (6); shown as gray area in Fig. 8). Where the conditional variance is high, a change in the parameter values has a large effect on the surface chlorophyll concentration. This, in turn, allows for more effective adjustments of the chlorophyll concentration by means of changing θ_1 and θ_2.

3.4.2. Spatial differences in optimal parameter values

Based on the optimal smoothing intensity found in Section 3.2.2, we now re-evaluate the development of optimal parameter values in time and examine the uncertainty in the model state. Instead of using the minimum as a point estimate for an optimal parameter value, we are interested in a *region* of good parameter values. These values are "good" in the sense that they are associated with low (but not necessarily minimal) distance values. To determine good parameter values, we performed the following steps: (1) For each day with available data, we interpolated the corresponding distance function in parameter space using the polynomial chaos expansion. (2) For each of the distance functions, we determined the region in parameter space that makes up 20% of its lowest values. (3) Finally, we computed the frequency with which a given pair of parameter values is contained within the 20% region. We expect that a good pair of parameter values is contained frequently in the 20% region of lowest distance values. The frequency of occurrence in this region, obtained for all parameter values, therefore provides us a with an estimate of the distribution of good parameter values which can be visualized easily.

Estimates of the parameter distribution for each season (Fig. 9(a)), obtained by the procedure described above, correspond well to the parameter paths in Fig. 5, yet the distribution additionally reveals features hidden in the point estimates. For example during the spring (AMJ; corresponding to April, May and June) there appears to be

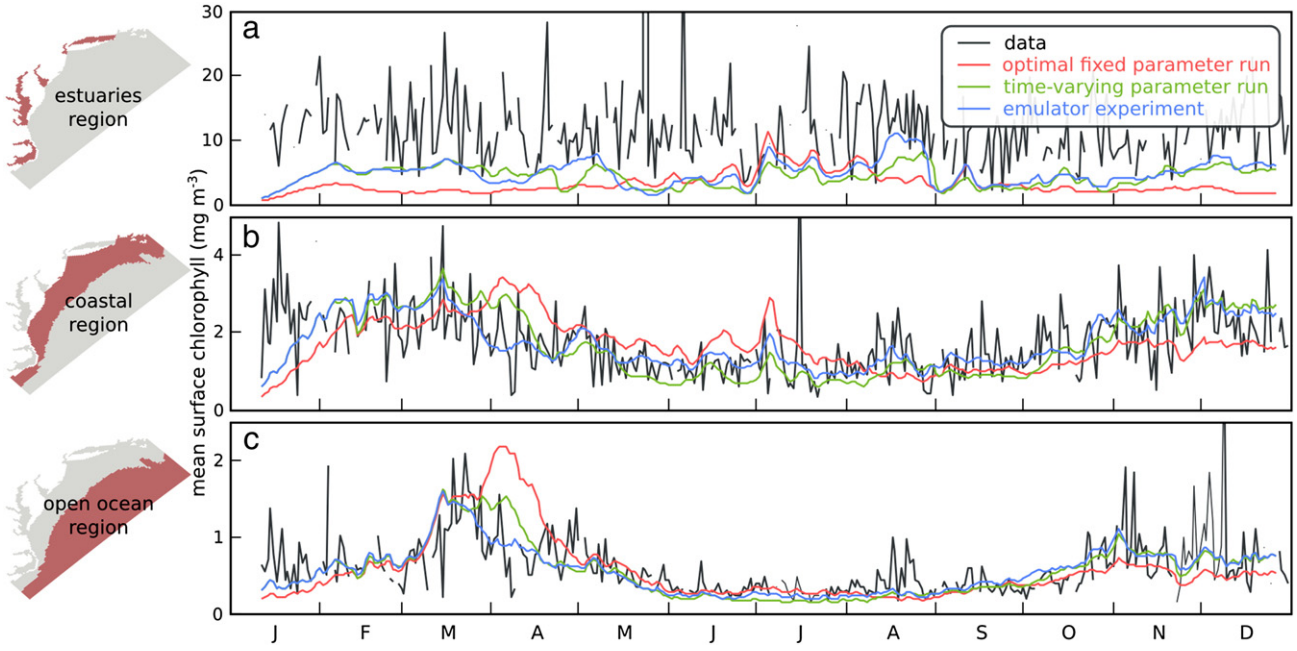

Fig. 7. The development of the average concentration of surface chlorophyll for the model, the optimal fixed parameter run, the time-varying parameter run (with a smoothing intensity of 5, corresponding to the best average distance result in Fig. 6) and the emulation experiment (with a smoothing intensity of 10, corresponding to the lowest average distance in the cross-validation experiment in Fig. 6). The analysis is divided into 3 model regions which are displayed in the left panel. The corresponding absolute residuals are shown in Fig. 8.

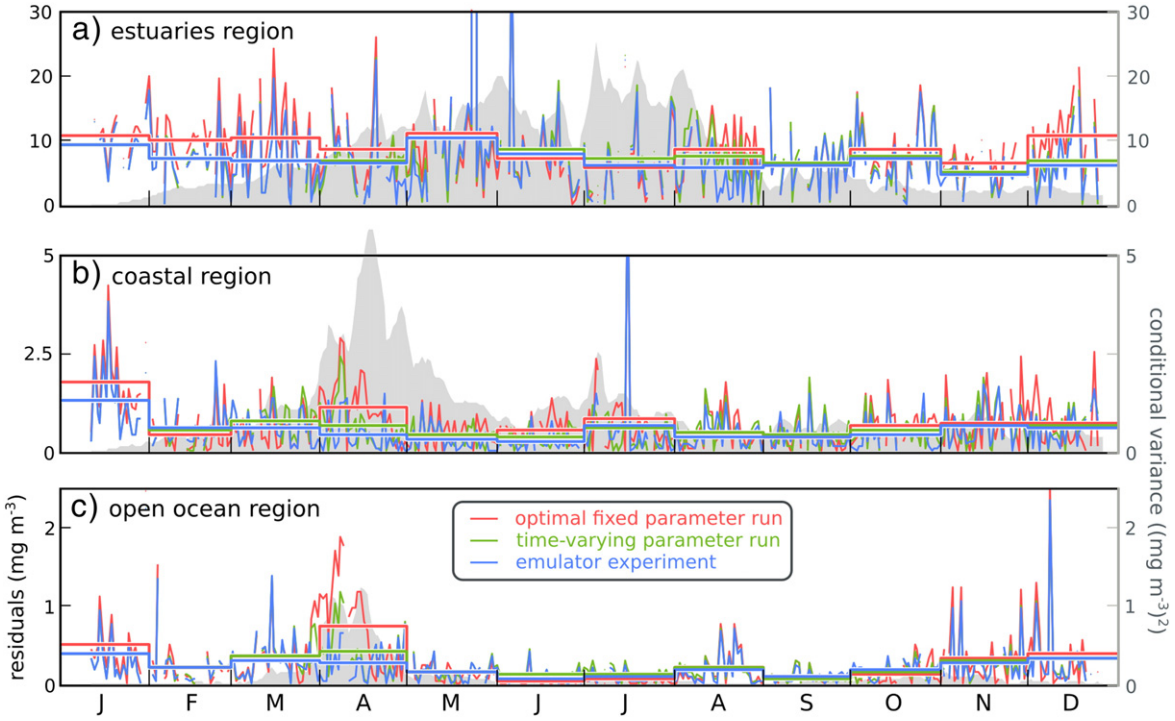

Fig. 8. The absolute residuals of the surface chlorophyll content shown in Fig. 7; monthly averages of the absolute residuals are displayed as thick lines. The gray area in the background is the conditional variance of the surface chlorophyll content based on Eq. (6).

very little sensitivity to changes in θ_2, the zooplankton grazing parameter, and good parameter values are distributed all along the θ_2 axis. In summer (JAS; July, August, September), the distribution changes in this respect, as low values of θ_2 become less probable as good parameter values. Seasonal differences are generally apparent, strengthening our previous observations that optimal parameter values change in time.

So far, we focused mostly on the change of parameters in time, but we can also use the same methodology for an analysis of spatial differences. For a spatial analysis, we use the 3 model regions introduced in Section 3.4.1 and shown in Fig. 7. All previous results were based on the distance function introduced in Eq. (1) which uses the full data set to compute distance values. By including observations from within one of the 3 regions only, the distance values can be recomputed and we can gain an understanding of suitable parameter values for that region. In order to detect spatial differences in good parameters we performed the distribution estimation for the 3 regions again (Fig. 9(b,c,d)).

Differences between model regions are apparent. In the estuaries, where chlorophyll is always underestimated, good parameter combinations tend to increase chlorophyll by combining high values of the chlorophyll-to-carbon ratio with low values of the zooplankton grazing rate throughout the whole year. More temporal variation is evident in the other two regions. In the shelf region, seasonal changes are most apparent and values of the zooplankton grazing rate tend to be generally high, especially in spring and summer. This result corresponds well to the tendency of the optimal fixed parameter run to overestimate chlorophyll during those months. In the outer ocean region which exhibits the lowest chlorophyll values, the model tends to be most insensitive to changes in the zooplankton grazing parameter whereas a very narrow range of θ_1 is preferred. The low amount of chlorophyll combined with relatively little chlorophyll variability sustains only a small population of zooplankton, thus the grazing parameter of zooplankton has a low impact.

Taken together, the results for the 3 model regions account for the full domain result presented above. The distance measure that was used (AGB) has no knowledge about the regions, thus the influence of the regions on the general result is mainly determined by their size. Hence, the large coastal and open ocean regions far outweigh the influence of the small estuaries region. Due to their different parameter preferences, the fit between data and model remains relatively poor for the estuaries region (compare Fig. 8). Despite being small, the estuaries exert a constant influence on the parameter estimation to raise chlorophyll levels. Here the polynomial chaos based interpolation shows its strength as a model analysis tool.

4. Discussion

In this study we obtained improved surface chlorophyll estimates from a biological ocean model by treating two of its parameters as stochastic. This was achieved through the approximation of the model by a low dimensional emulator, the polynomial chaos expansion. Using the polynomial chaos expansion in combination with a model-data distance function we found that the values of two biological parameters have a clear time dependence and follow a seasonal path through parameter space (Fig. 5).

At two points in this study we encountered high frequency variations; they appeared in the distance function (Fig. 2) and the derived parameter paths (Fig. 5). In the case of the distance function we attribute the high frequency signal to noise and missing values in the chlorophyll images. The high frequency changes in the parameter paths indicate that the same noise is overfitted by our optimization procedure. We confirmed this inference in a cross-validation experiment (Section 3.2.2), where we observed a strong increase in the average distance value for low smoothing intensities while the best results were achieved at medium smoothing intensities (Fig. 6). This result is evidence that the improvement of our time-varying parameter state estimates is based on an actual signal in the observations that is not captured in the fixed parameter run.

By treating only two biological parameters as stochastic and by adjusting them to fit the observations, we do not account for the fact that model-data discrepancies are also caused by other sources

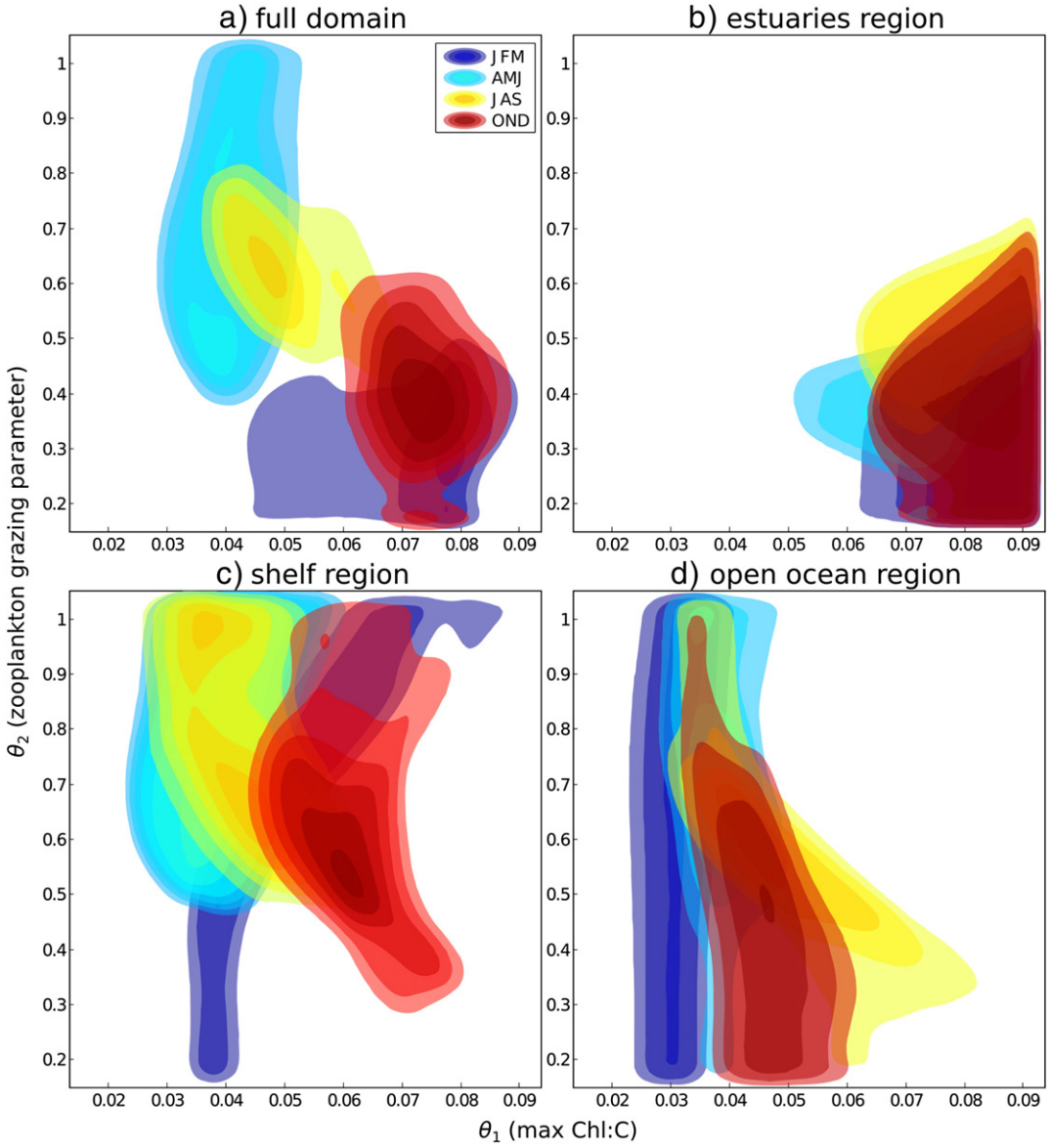

Fig. 9. The distribution of good parameter values in parameter space (compare Section 3.4.2) depending on season. The first panel shows a contour plot of the distributions for the entire model domain, the other panels contain the distributions for the 3 regions shown in Fig. 7.

of model error, such as the other biological parameters, parameters of the underlying physical model, physical forcing, boundary and initial conditions as well as the functional form of the equations themselves. For example, the selection of the maximum chlorophyll-to-carbon ratio (θ_1) as a stochastic parameter and its optimization may adjust for errors in the phytoplankton growth rate and errors in the model's nutrient supply. A fully Bayesian approach, which would incorporate all sources of model uncertainty, is computationally infeasible. We chose to focus this study on one obvious shortcoming of the model, the representation of phytoplankton and zooplankton as homogenous groups. Within this much more limited scope, we selected the two parameters that have the strongest influence on the model's chlorophyll concentration. Here, our motivation is simply that the most sensitive parameters will likely be identifiable using chlorophyll data and yield the biggest improvement in chlorophyll estimates.

Although we have no detailed information on the phytoplankton species succession and seasonal changes in grazing rate, we can attempt a qualitative comparison of the development of θ_1 and θ_2 with typical seasonal changes in the plankton composition of the Middle Atlantic Bight. In our model run with time-varying parameters, there is a positive correlation between the inverse of the maximum chlorophyll-to-carbon ratio ($\frac{1}{\theta_1}$) and the achieved phytoplankton carbon-to-chlorophyll ratio (C:Chl) in the surface (Fig. 10(a)). In comparison to the model with optimal fixed parameters, the time-varying parameters lead to an increase in C:Chl in the summer months following the phytoplankton spring bloom. In the Middle Atlantic Bight dinoflagellates typically dominate the phytoplankton community in the shelf region during summer (Marra et al., 1990) while diatoms are the dominant phytoplankton group during the spring bloom (Barlow et al., 1993). Due to a significantly lower C:Chl in diatoms in comparison to dinoflagellates (Chan, 1980), we would expect a lower C:Chl during the spring bloom and a higher C:Chl in summer. While the optimal fixed parameter run shows no marked increase in C:Chl as the bloom subsides, there is a notable increase in the C:Chl induced by the time-varying parameters, consistent with our expected C:Chl development (Fig. 10(a)). This improved correlation does not imply causation, as we have pointed out in the previous paragraph, yet it is consistent with the hypothesis that variations in C:Chl are significantly affected by shifts in the phytoplankton composition.

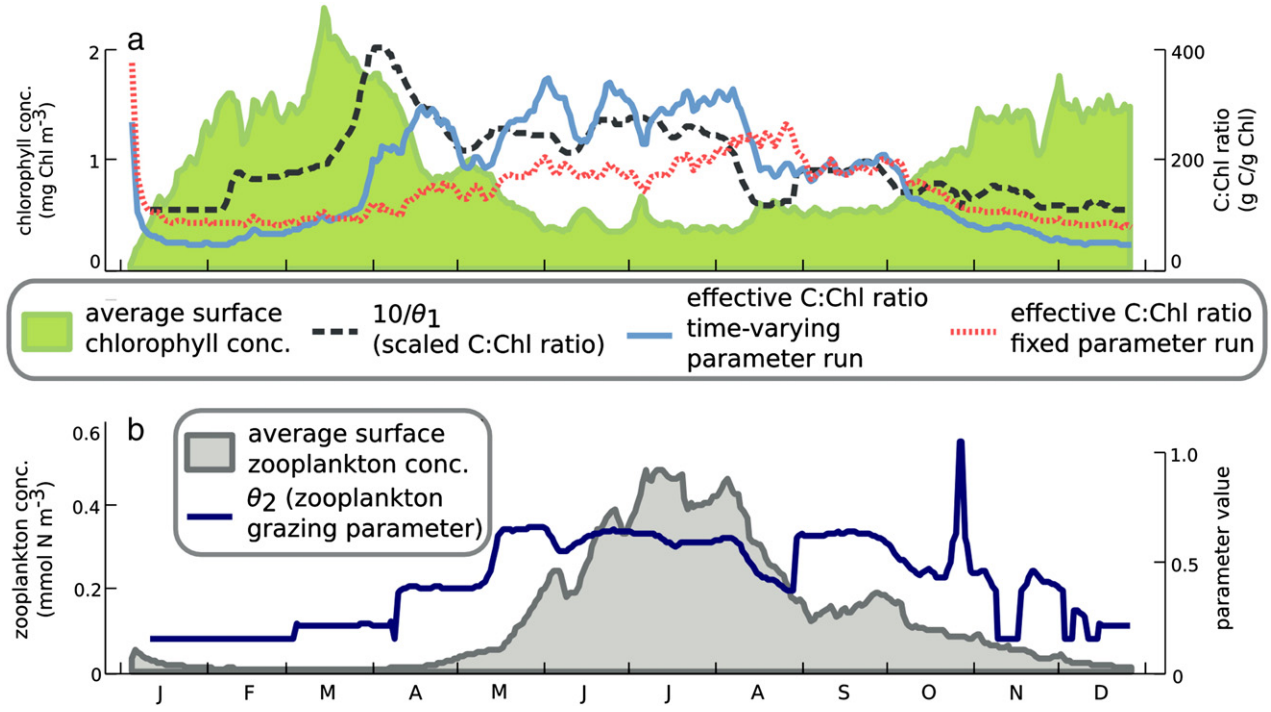

Fig. 10. The development of the time-varying values of θ_1 and θ_2 (smoothing intensity of 10, compare Fig. 5) in relation to the development of the surface chlorophyll content. In addition to θ_1, panel (a) also shows the resulting carbon to chlorophyll (C:Chl) ratio in the surface time-varying parameter run and the corresponding C:Chl ratio for the optimal fixed parameter model run. For better comparison, θ_1 is transformed to $\frac{1}{\theta_1}$ which is also a carbon to chlorophyll ratio and then scaled by a factor of 10.

Evaluating the development of θ_2, the zooplankton grazing parameter, is more difficult. In the model, the zooplankton maximum occurs in the summer, is preceded by a notable increase in θ_2 in April (Fig. 10(b)) and remains high for several months until November. This pattern may reflect a correction of the seasonal cycle of zooplankton. The increase of θ_2 in April enhances zooplankton grazing and hints that the effect of grazing in the model is too low at that time of the year. Kane (2005) found that the zooplankton species *Calanus finmarchicus*, an important part of the zooplankton population, shows a sharp increase in abundance in early spring and suggest that it is a consequence of an import of zooplankton into the Middle Atlantic Bight from neighboring regions. Such a process is unaccounted for in the biological model and could explain the development of θ_2 in spring. Lack of import causes an underestimation of zooplankton abundance and grazing in the model, which is counteracted by an increased zooplankton grazing parameter. It should be noted, however, that it may be difficult to constrain the zooplankton parameter using chlorophyll observations, given the indirect effect of changes in zooplankton grazing on chlorophyll. In addition, zooplankton dynamics are known to be highly variable from one year to the next, even under similar phytoplankton conditions (Flagg et al., 1994).

Given our self-imposed restriction of optimizing two parameters, the improvement in surface chlorophyll estimates is considerable. By using the parameter paths as time-varying parameter values, more improvement can be achieved than by changing θ_2 (the zooplankton grazing parameter and one of the models' most sensitive parameters) from its most disadvantageous value in our broadly selected parameter range, to its optimum value (Fig. 3). The results of our emulation experiment tend to be better than those of the biological model simulation with time-varying parameters; the main reason for this is that the emulation experiment is not bound by the model dynamics and changes in parameters become effective immediately. In contrast, the time-varying parameter model run has a memory of accumulated (or lost) chlorophyll and a change in parameter value needs some time to translate into a changed surface state. We expect that the level of improvement to be gained from time-varying parameters, will in general depend on the model's memory, where properties with fast response will be more prone to improvements.

We varied only two of the biological parameters and decided to keep the general setup simple, e.g. by using the entire data set without excluding outliers. The distance function we interpolated appears to be smooth and well approximated by the polynomial interpolation (Fig. 3). Consequently, we can still expect good results for fewer quadrature points in parameter space, which have the benefit of decreasing the number of necessary model runs. Yet even after a reduction of quadrature points it would be computationally expensive to extend our analysis to more than a few parameters. Other emulator approaches can sample parameter space in a more efficient, non-grid based manner (e.g. Latin hypercube sampling introduced by McKay et al. (1979) or a free selection of parameter values as in Hooten et al. (2011)) and may be better suited for parameter estimation in higher dimensional spaces. One advantage of the polynomial chaos technique is that it offers a straightforward way to obtain model uncertainty estimates (see Section 2.3), which do not require an additional analysis step. In contrast to other emulators that do not utilize basis functions, there is no need to run Monte Carlo-based sampling techniques within the emulator framework to obtain approximates of model uncertainty (integrals of interest can be evaluated directly with the help of the polynomial basis functions).

We focused this study on one specific data type, satellite images of chlorophyll, in conjunction with one specific model-data distance measure, the AGB. The approach we took to estimate optimal parameters and further obtain improved state estimates is very flexible and allows for the use of other model-data distance measures (such as RMSE), other data types (such as in-situ measurements) and combinations of different observations. Any model-data distance measure suitable for the comparison of the data type of choice or a (weighted) sum of multiple such distance measures would have to be substituted for the distance function d in Eq. (1). The polynomial chaos expansion can then operate on the new distance values without any further

changes. In fact, the use of one or more new data sets or new distance measures does not require new model simulations.

A great advantage of the polynomial chaos expansion is the amount of postprocessing and analysis options. Once the necessary model simulations have been performed (in our case 49 runs) various different analyses, from distance function interpolation, to spatial analyses and chlorophyll surface state interpolations, can be performed without any further model simulations. In addition to direct estimates of the model output, the polynomial chaos expansion also provides us with estimates of the conditional variance (Eq. (6)) of the output. While it is not a measure for the full model error, knowledge of the conditional variance can be useful for analyzing the model output, for example, to gauge the impact of the parameter variation on a specific model region or time. In our analysis, the conditional variance (Fig. 8) gives a good indication where in space and time model improvement is possible by means of parameter optimization. Given these advantages, we consider the polynomial chaos expansion a useful tool for model analysis and the introduction of uncertainty into biological models.

Our study offers some insights into general parameter optimization issues. The average model-data distance function is well behaved and contains a clearly defined global minimum (Fig. 3), which even simple parameter optimization techniques will find easily. Yet its smoothness hides the fact that the optimal parameters for individual observations are widely scattered in parameter space (Fig. 5). Part of the reason for the wide spread of optimal parameter values is the strong underlying time dependence. Were we to optimize our model with fixed parameters using only satellite data from spring months, we would get significantly different results than by using fall data (Fig. 9). By optimizing the model with a full year's worth of observations the fixed parameter values fall somewhere in between the optimal seasonal values. For this study, only one year of daily satellite observations was used. One of our next steps will be to analyze if the parameter paths generalize well for other years.

5. Conclusions

The model-data fit of a typical biological ocean model can be greatly improved by allowing its biological parameters to vary in time. We obtained the parameter values of two biological parameters by minimizing a time-dependent distance function using an emulator-based approximation. State estimates that are based on the time varying parameters fit observations much better than those gained from the optimal fixed parameter run. This improvement is not due to overfitting the data, instead there is a low frequency variation present in the parameter values: the two biological parameters analyzed here appear to follow a seasonal cycle in parameter space. The development of at least one of the parameters matches patterns observed in plankton dynamics in the Middle Atlantic Bight.

Beside temporal differences, we also detect spatial differences of optimal parameter values for selected model regions. The estuaries and coastal and open ocean regions in our model domain show clear preferences for distinct parameter values. The polynomial chaos expansion can help identify spatial differences, detect model regions with a generally bad fit to the data and assess their influence on optimal parameter values.

The polynomial chaos expansion proved to be a versatile tool for the optimization and analysis of our biological model. While computational cost limits the number of parameters one can analyze jointly to just a few, we achieved large gains by analyzing only two parameters that the model is sensitive to. The number of postprocessing options we gained after performing the necessary model runs is great: model uncertainty estimates can be obtained directly and multiple parameter estimations with different data sets can be performed efficiently without the requirement for any additional model runs.

Acknowledgments

This work was supported by the ONR MURI grant N00014-06-1-0739 to KF. KF is also acknowledging support from ACEnet, NSERC and CFI. MD acknowledges support from NSERC. We thank Carlisle Thacker for many constructive comments on an earlier version of this manuscript. We also thank an anonymous reviewer whose comments led to substantial improvements.

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. Ann. Geophys. 21 (1), 399–411. doi:10.5194/angeo-21-399-2003.

Askey, R., Wilson, J.A., 1985. Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials. Memoirs of the American Mathematical Society, Vol. 319.

Aumont, O., Maier-Reimer, E., Blain, S., Monfray, P., 2003. An ecosystem model of the global ocean including Fe, Si, P colimitations. Global Biogeochem. Cycles 17 (2). doi:10.1029/2001GB001745.

Barlow, R.G., Mantoura, R.F.C., Gough, M.A., Fileman, T.W., 1993. Pigment signatures of the phytoplankton composition in the northeastern Atlantic during the 1990 spring bloom. Deep-Sea Res. II Top. Stud. Oceanogr. 40 (1–2), 459–477. doi:10.1016/0967-0645(93)90027.

Bianucci, L., Denman, K.L., Ianson, D., 2011. Low oxygen and high inorganic carbon on the Vancouver Island Shelf. J. Geophys. Res. 116 (C7), C07011. doi:10.1029/2010JC006720.

Chan, A.T., 1980. Comparative physiological study of marine diatoms and dinoflagellates in relation to irradiance and cell size. II. Relationship between photosynthesis, growth, and carbon/chlorophyll a ratio. J. Phycol. 16, 428–432. doi:10.1111/j.1529-8817.1978.tb02458.x.

Chen, K., He, R., 2010. Numerical investigation of the Middle Atlantic Bight shelfbreak frontal circulation using a high-resolution ocean hindcast model. J. Phys. Oceanogr. 40 (5), 949–964. doi:10.1175/2009JPO4262.1.

Denman, K.L., 2003. Modelling planktonic ecosystems: parameterizing complexity. Prog. Oceanogr. 57 (3–4), 429–452. doi:10.1016/S0079-6611(03)00109-5.

Doney, S.C., Glover, D.M., Najjar, R.G., 1996. A new coupled, one-dimensional biological–physical model for the upper ocean: applications to the JGOFS Bermuda Atlantic time-series study (BATS) site. Deep-Sea Res. II Top. Stud. Oceanogr. 43 (2–3), 591–624. doi:10.1016/0967-0645(95)00104-2.

Dowd, M., 2007. Bayesian statistical data assimilation for ecosystem models using Markov Chain Monte Carlo. J. Mar. Syst. 68 (3–4), 439–456. doi:10.1016/j.jmarsys.2007.01.007.

Dowd, M., 2011. Estimating parameters for a stochastic dynamic marine ecological system. Environmetrics 22 (4), 501–515. doi:10.1002/env.1083.

Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dyn. 53 (4), 343–367. doi:10.1007/s10236-003-0036-9.

Fennel, K., Wilkin, J., 2009. Quantifying biological carbon export for the northwest North Atlantic continental shelves. Geophys. Res. Lett. 36, L18605. doi:10.1029/2009GL039818.

Fennel, K., Wilkin, J., Levin, J., Moisan, J., Haidvogel, D., 2006. Nitrogen cycling in the Middle Atlantic Bight: results from a three-dimensional model and implications for the North Atlantic nitrogen budget. Global Biogeochem. Cycles 20 (3), GB3007. doi:10.1029/2005GB002456.

Fennel, K., Wilkin, J., Previdi, M., Najjar, R., 2008. Denitrification effects on air–sea CO_2 flux in the coastal ocean: simulations for the Northwest North Atlantic. Geophys. Res. Lett. 35, L24608. doi:10.1029/2008GL036147.

Flagg, C.N., Wirick, C.D., Smith, S.L., 1994. The interaction of phytoplankton, zooplankton and currents from 15 months of continuous data in the Mid-Atlantic Bight. Deep-Sea Res. II Top. Stud. Oceanogr. 41 (2–3), 411–435. doi:10.1016/0967-0645(94)90030-2.

Follows, M.J., Dutkiewicz, S., 2011. Modeling diverse communities of marine microbes. Ann. Rev. Mar. Sci. 3 (1), 427–451. doi:10.1146/annurev-marine-120709-142848.

Follows, M.J., Dutkiewicz, S., Grant, S., Chisholm, S.W., 2007. Emergent biogeography of microbial communities in a model ocean. Science 315 (5820), 1843–1846. doi:10.1126/science.1138544.

Franks, P.J.S., Chen, C.S., 1996. Plankton production in tidal fronts: a model of Georges Bank in summer. J. Mar. Res. 54 (4), 631–651. doi:10.1357/0022240963213718.

Frolov, S., Baptista, A.M., Leen, T.K., Lu, Z., van der Merwe, R., OCT 2009. Fast data assimilation using a nonlinear Kalman filter and a model surrogate: an application to the Columbia River estuary. Dyn. Atmos. Oceans 48 (1–3), 16–45. doi:10.1016/j.dynatmoce.2008.10.004.

Geider, R.J., MacIntyre, H.L., Kana, T.M., 1997. Dynamic model of phytoplankton growth and acclimation: responses of the balanced growth rate and the chlorophyll a:carbon ratio to light, nutrient-limitation and temperature. Mar. Ecol. Prog. Ser. 148, 187–200. doi:10.3354/meps148187.

Geider, R.J., MacIntyre, H.L., Kana, T.M., 1998. A dynamic regulatory model of phytoplanktonic acclimation to light, nutrients, and temperature. Limnol. Oceanogr. 43 (4), 679–694. doi:10.4319/lo.1998.43.4.0679.

Goebel, N.L., Edwards, C.A., Zehr, J.P., Follows, M.J., 2010. An emergent community ecosystem model applied to the California Current System. J. Mar. Syst. 83 (3–4), 221–241. doi:10.1016/j.jmarsys.2010.05.002.

Gregg, W.W., Ginoux, P., Schopf, P.S., Casey, N.W., 2003. Phytoplankton and iron: validation of a global three-dimensional ocean biogeochemical model. Deep-Sea Res. II Top. Stud. Oceanogr. 50 (22–26), 3143–3169. doi:10.1016/j.dsr2.2003.07.013.

Haidvogel, D.B., Arango, H., Budgell, W.P., Cornuelle, B.D., Curchitser, E., Di Lorenzo, E., Fennel, K., Geyer, W.R., Hermann, A.J., Lanerolle, L., et al., 2008. Ocean forecasting in terrain-following coordinates: formulation and skill assessment of the regional ocean modeling system. J. Comput. Phys. 227 (7), 3595–3624. doi:10.1016/j.jcp.2007.06.016.

Hooten, M., Leeds, W., Fiechter, J., Wikle, C., 2011. Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models. J. Agric. Biol. Environ. Stat. 16, 475–494. doi:10.1007/s13253-011-0073-7.

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. doi:10.1016/j.jmarsys.2011.11.016.

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, 7–16.

Kane, J., 2005. The demography of *Calanus finmarchicus* (Copepoda: Calanoida) in the Middle Atlantic Bight, USA, 1977–2001. J. Plankton Res. 27 (5), 401–414. doi:10.1093/plankt/fbi009.

Lawson, L.M., Hofmann, E.E., Spitz, Y.H., 1996. Time series sampling and data assimilation in a simple marine ecosystem model. Deep-Sea Res. II Top. Stud. Oceanogr. 43 (2), 625–651. doi:10.1016/0967-0645(95)00096-8.

Lehmann, M.K., Fennel, K., He, R., 2009. Statistical validation of a 3-D bio-physical model of the western North Atlantic. Biogeosciences 6 (10), 1961–1974. doi:10.5194/bg-6-1961-2009.

Losa, S.N., Kivman, G.A., Schröter, J., Wenzel, M., 2003. Sequential weak constraint parameter estimation in an ecosystem model. J. Mar. Syst. 43 (1–2), 31–49. doi:10.1016/j.jmarsys.2003.06.001.

Lucas, D.D., Prinn, R.G., 2005. Parametric sensitivity and uncertainty analysis of dimethylsulfide oxidation in the clear-sky remote marine boundary layer. Atmos. Chem. Phys. 5 (6), 1505–1525. doi:10.5194/acp-5-1505-2005.

Mannino, A., Russ, M.E., Hooker, S.B., 2008. Algorithm development and validation for satellite-derived distributions of DOC and CDOM in the US Middle Atlantic Bight. J. Geophys. Res. 113 (C7), C07051. doi:10.1029/2007JC004493.

Marra, J., Houghton, R.W., Garside, C., 1990. Phytoplankton growth at the shelf-break front in the Middle Atlantic Bight. J. Mar. Res. 48 (4), 851–868.

Marzouk, Y.M., Najm, H.N., 2009. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. J. Comput. Phys. 228 (6), 1862–1902. doi:10.1016/j.jcp. 2008.11.024.

Mattern, J.P., Dowd, M., Fennel, K., 2010a. Sequential data assimilation applied to a physical–biological model for the Bermuda Atlantic time series station. J. Mar. Syst. 79 (1–2), 144–156. doi:10.1016/j.jmarsys.2009.08.004.

Mattern, J.P., Fennel, K., Dowd, M., 2010b. Introduction and assessment of measures for quantitative model-data comparison using satellite images. Remote Sens. 2 (3), 794–818. doi:10.3390/rs2030794.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245. doi:10.2307/1268522.

Moore, J.K., Doney, S.C., Kleypas, J.A., Glover, D.M., Fung, I.Y., 2001. An intermediate complexity marine ecosystem model for the global domain. Deep-Sea Res. II Top. Stud. Oceanogr. 49 (1–3), 403–462. doi:10.1016/S0967-0645(01)00108-4.

Powell, B.S., Arango, H.G., Moore, A.M., Di Lorenzo, E., Milliff, R.F., Foley, D., 2008. 4DVAR data assimilation in the intra-Americas sea with the Regional Ocean Modeling System (ROMS). Ocean Modell. 23 (3–4), 130–145. doi:10.1016/j.ocemod.2008.04.008.

Previdi, M., Fennel, K., Wilkin, J., Haidvogel, D., 2009. Interannual variability in atmospheric CO_2 uptake on the northeast U.S. continental shelf. J. Geophys. Res. 114 (G4), G04003. doi:10.1029/2008JG000881.

Rougier, J., Sexton, D.M.H., 2007. Inference in ensemble experiments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 365 (1857), 2133–2143. doi:10.1098/rsta.2007.2071.

Scott, V., Kettle, H., Merchant, C.J., 2011. Sensitivity analysis of an ocean carbon cycle model in the North Atlantic: an investigation of parameters affecting the air–sea CO_2 flux, primary production and export of detritus. Ocean Sci. 7 (3), 405–419. doi:10.5194/os-7-405-2011.

Smedstad, O., O'Brien, J.J., 1991. Variational data assimilation and parameter estimation in an equatorial Pacific Ocean model. Prog. Oceanogr. 26 (2), 179–241. doi:10.1016/0079-6611(91)90002-4.

Thacker, W.C., Srinivasan, A., Iskandarani, M., Knio, O.M., Le Hénaff, M., 2012. Propagating boundary uncertainties using polynomial expansions. Ocean Modell. 43–44, 52–63. doi:10.1016/j.ocemod.2011.11.011.

Wan, X.L., Karniadakis, G.E., 2006. Beyond Wiener–Askey expansions: handling arbitrary PDFs. J. Sci. Comput. 27 (1–3), 455–464. doi:10.1007/s10915-005-9038-8.

Wiener, N., 1938. The homogeneous chaos. Am. J. Math. 60 (4), 897–936. doi:10.2307/2371268.

Xiu, D.B., Karniadakis, G.E., 2002. The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. 24 (2), 619–644. doi:10.1137/S1064827501387826.

Xiu, D., Karniadakis, G.E., 2003. Modeling uncertainty in flow simulations via generalized polynomial chaos. J. Comput. Phys. 187 (1), 137–167. doi:10.1016/S0021-9991(03)00092-5.