Ocean Modelling 142 (2019) 101437

Contents lists available at ScienceDirect

Ocean Modelling

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

Evaluating ecosystem model complexity for the northwest North Atlantic through surrogate-based optimization

Angela M. Kuhn [a,b,*], Katja Fennel [a]

[a] Department of Oceanography, Dalhousie University, Halifax, NS, B3H 4R2, Canada
[b] Scripps Institution of Oceanography, University of California – San Diego. La Jolla, CA, 92093-0218, USA

ARTICLE INFO

Keywords:
Plankton
Ocean model
Model complexity
Optimization
North Atlantic

ABSTRACT

Objectively determining the level of ecosystem model complexity necessary to achieve meaningful representations of biogeochemical cycles at different spatial and temporal scales is an outstanding issue in marine ecosystem modeling. As part of the development of a three-dimensional (3D) Regional Ocean Modelling System (ROMS) application for the northwest North Atlantic Ocean, we compare model results from three alternative ecosystem model versions in which ecological complexity was increased in a step-wise fashion. In order to ensure an objective comparison, the models were optimized to replicate observations of satellite surface chlorophyll, and *in situ* chlorophyll and nitrate profiles. To overcome the high computational cost of optimizing 3D models, we use a surrogate-based optimization method; that is, an ensemble of one-dimensional (1D) models is used as a proxy of the ecosystem model behavior in the 3D setting. The 1D models were configured at locations where *in situ* profiles are available. A total of 17 optimization experiments aim to evaluate different aspects of the comparison between the ecosystem models. We show that for all ecosystem model versions the optimized model performance degrades when the optimization includes all observed variables at all locations instead of individual locations only. Moreover, the choice of parameters to be optimized can significantly affect the behavior of the optimized models and is most noticeable when multiple phytoplankton and zooplankton groups are included. Additionally, evaluation of spatial patterns in optimal parameter values at individual locations allows us to assess geographical model portability. In general, an optimized complex model can achieve lower model-data misfits against assimilated data than simple models, but is also more prone to generating unintended trophic relations. The more complex model also had decreased performance when applied to locations different than those used for calibration (i.e., "portability experiments"), which is discussed in the context of the design of the cost function and selection of parameters to optimize.

1. Introduction

Since the emergence of numerical marine ecology, models have diversified from describing simple prey predator relationships (e.g., Riley, 1965) to representing multiple plankton functional groups and chemical variables, with dependencies on the characteristics of the physical environment (e.g., Dutkiewicz et al., 2015; Kishi et al., 2007). There is an ongoing discussion about the most appropriate level of ecosystem complexity, model structure, and parameterizations of functional relationships. Both simple and complex models have advantages and disadvantages. For instance, the use of simple models under idealized conditions has proven to be valuable in identifying and understanding underlying mechanisms of the marine ecosystem functioning (e.g., Evans and Parslow, 1985; Fasham et al., 1990; Kuhn et al., 2015). However, it is frequently argued that more realistic representations of the plankton community composition and the interrelationships of marine food webs are required to improve forecasting capabilities in regional and global models (e.g., Le Quéré et al., 2005).

Regardless of their complexity, marine biogeochemical models depend on many parameters that describe biological and chemical rates of change such as growth, mortality, and degradation rates, including maximum rates and half-saturation concentrations in nutrient uptake and predation formulations. As models are developed for specific regions or periods, their parameters are typically calibrated to fit observations for those specific conditions. This may lead to overfitting, a loss of model forecasting skill and of portability to different geographic locations (see Friedrichs et al., 2007). In general, the number of parameters increases with the number of state variables in a model (Denman, 2003); thus, complex models are at a higher risk of overfitting. Moreover, most of these parameters are poorly known and wide value ranges are reported in the literature. Studies using systematic calibration methods, known as parameter optimization, have

* Corresponding author at: Scripps Institution of Oceanography, University of California – San Diego. La Jolla, CA, 92093-0218, USA.
E-mail address: angela.kuhn@dal.ca (A.M. Kuhn).

https://doi.org/10.1016/j.ocemod.2019.101437
Received 18 September 2018; Received in revised form 12 August 2019; Accepted 18 August 2019
Available online 20 August 2019
1463-5003/

shown that typically available observational sets are often not sufficient to determine more than a few of these parameters (Fennel et al., 2001; Friedrichs et al., 2007; Ward et al., 2010; Kwon and Primeau, 2006). Parameter values may also change depending on the plankton community composition, and thus it is possible that models need to consider parameter variations with geography (Losa et al., 2004) or time (Mattern et al., 2012).

Following the principle of parsimony, the simplest model able to fit the observations should be favored over more complex ones. Failure of a model to replicate observations suggests that the model structure may be missing key components of the system's behavior. Conversely, the ability of a model to replicate a given set of observations does not unequivocally mean that all processes are properly represented. When models differ not only in their level of ecological complexity, but also in the degree to which they were calibrated and in the model pathways that were affected during calibration, it is tenuous to argue that differences in model performance are due to structural complexity (Friedrichs et al., 2007; Kriest et al., 2010). To better discern the effects of increased ecosystem complexity from differences in a model's response due to its parameter values, it is necessary to calibrate the model versions to comparable levels of performance and through comparable pathways of mass flux (e.g., Xiao and Friedrichs, 2014a; Galbraith et al., 2015; Kriest, 2017).

During the 90s, the calibration of marine ecosystem models to a specific study region was predominantly subjective (Arhonditsis and Brett, 2004). This approach is inefficient, increases the risk of overlooking structural inadequacies in the models, and is complicated by the number of parameters in play and their co-dependencies. In recent decades, this problem has been increasingly addressed with the use of parameter optimization techniques (Fennel et al., 2001; Kwon and Primeau, 2006, 2008; Friedrichs et al., 2007; Bagniewski et al., 2011; Schartau et al., 2017; Kriest et al., 2017). Parameter optimization provides a more objective framework for comparing models with different degrees of trophic complexity, but optimization experiments require a large number of model runs and thus, their direct application to computationally expensive 3D models is difficult.

An alternative, which we choose here, is to perform the optimization using a simplified faster model that replicates the results of the computationally more expensive 3D model. The computationally efficient model is referred to as a model surrogate or emulator. The surrogate can be a statistical model, a coarser resolution model, or a reduced order model that allows one to perform a large number of simulations required for parameter sensitivity analyses and model calibration. Different techniques for the construction of statistical emulators of 3D biogeochemical models have been tested in recent years (Hooten et al., 2011; Leeds et al., 2012; Mattern et al., 2012). Other tested approaches include reduced temporal resolution (Prießet al., 2013a,b), and reduced physical dimensionality (Hemmings and Challenor, 2012; Hemmings et al., 2015). Our methodology resembles the latter reduced dimensionality studies in that our model surrogate is a mechanistic emulator constructed with an ensemble of 1D models, located at points where *in situ* chlorophyll-a and nitrate profiles are available. The surrogate (1D models) and the target model (3D model) share the same ecosystem model. Therefore, in comparison with statistical and reduced process-resolution surrogates, the reduced dimensionality approach provides insight into the ecosystem responses most affected by the physical dynamics. Features that are well replicated in 1D are likely controlled by the ecosystem model itself (structure, equations and parameter values), whereas biases between 1D and 3D are a consequence of the simplified physical framework.

In this manuscript, we focus on the methodology used to bring models with different ecological structures to a comparable level of calibration by analyzing results from several optimization experiments. Our study region is the northwest North Atlantic continental shelf, which involves areas with contrasting oceanographic conditions. Our overarching goal is to better understand the variability of phytoplankton and primary production in the region, while addressing the unresolved question of how much ecological complexity is needed to represent it. We specifically aim to answer: 1. Does the ecosystem model structure and/or its dependency on temperature affect the surrogate-based optimization performance?, and 2. How does the number of observed variables compared, the number of parameters optimized, and the location the model is optimized for affect each model optimization? Aside from gaining insight about the answers to these questions, we demonstrate that the surrogate approach is effective as a calibration tool despite its simplicity. An in-depth comparison of chlorophyll and primary production patterns obtained by the 3D ocean model after the optimization will be explored in a subsequent study.

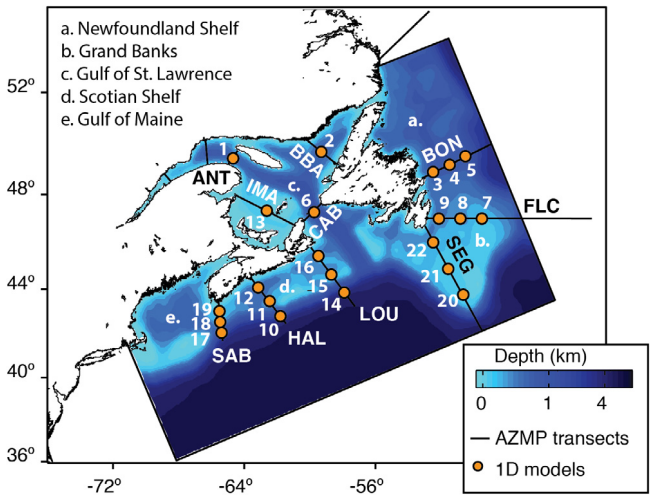

Fig. 1. Domain of the 3D ocean model with bathymetry of the study area (color background), and sampling transects (black lines) of the Atlantic Zone Monitoring Program (AZMP). The selected locations of the surrogate 1D models are represented as orange circles along the sampling tracks. A single 1D location is included for transects inside the Gulf of St. Lawrence, and their observational counterparts correspond to all observations within a 0.5° ratio from the reference center coordinates detailed in Table 1. Locations outside the Gulf of St. Lawrence are selected by dividing the length of the transect (within the 3D model domain) in three non-overlapping sections. A 1D location is placed at the center of each transect segment. Observational counterparts correspond to all observations within a 0.5° ratio from the 1D location center coordinates or within the length of the segment, whichever is shorter.

2. Study area

Our model domain covers the continental shelf and adjacent open ocean waters of the northwest North Atlantic Ocean, including the Newfoundland Shelf, the Grand Banks, the Gulf of St. Lawrence, the Scotian Shelf, and the Gulf of Maine (Fig. 1). The region is at the confluence of the two major North Atlantic current systems, the equatorward flowing Labrador Current and the northeast flowing Gulf Stream, and is influenced by their adjoining Shelf and Slope Water currents (Loder et al., 1998). This leads to complicated dynamics, including both cold and warm sub-regions originating from the North Atlantic subpolar and subtropical gyres (Townsend et al., 2004).

3. Model descriptions

3.1. Physical configuration

We use the 3D physical framework of the Regional Ocean Modeling System (ROMS, version 3.5, http://myroms.org, Haidvogel et al., 2008). The model is nested within the regional ocean-ice model of the northwest North Atlantic of Urrego-Blanco and Sheng (2012). The physical model implementation, and detailed sensitivity analyses and validation of the simulated physical variables of this model are

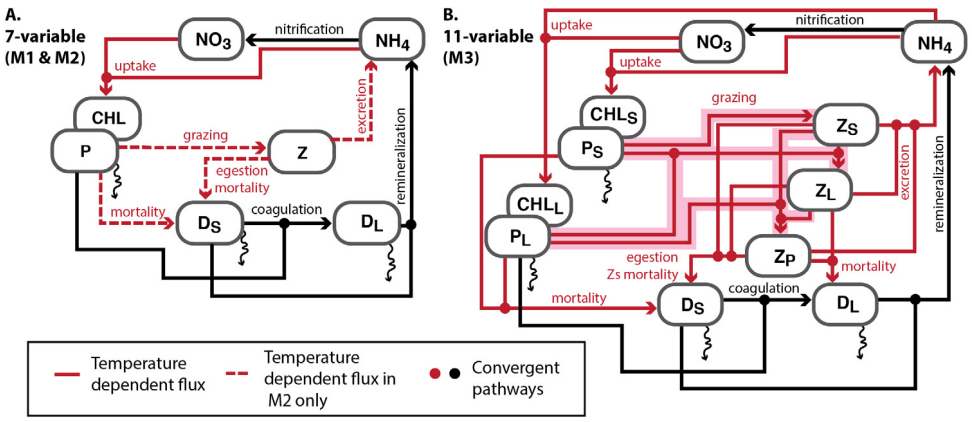

Fig. 2. Schematics of the ecosystem models used in the study. Fluxes that do not depend on temperature are represented by black lines, and temperature-dependent fluxes are represented by red lines A.) M1 and M2 both have the same 7 state variables: phytoplankton (P), chlorophyll (CHL), nitrate (NO_3), ammonium (NH_4), zooplankton (Z), small detritus (D_S) and large detritus (D_L). In M2 more fluxes are temperature-dependent, these are marked as dashed red lines. B.) M3 has 11 state variables: small phytoplankton (P_S), large phytoplankton (P_L), small phytoplankton chlorophyll (CHL_S), large phytoplankton chlorophyll (CHL_L), nitrate (NO_3), ammonium (NH_4), small zooplankton (Z_S), large zooplankton (Z_L), predatory zooplankton (Z_P), small detritus (D_S) and large detritus (D_L). Grazing fluxes are indicated by red lines with a pink shading. Fluxes that converge to feed the same variable are represented by solid circles (i.e., convergent pathways).

described in Brennan et al. (2016). The physical model has also been shown to realistically reproduce the distinct pathways of water mass movements in the domain (Rutherford and Fennel, 2018). Similar to the biogeochemical application by Bianucci et al. (2015), ocean temperature and salinity are weakly nudged (time scale of 140 days) to climatological fields from Geshelin et al. (1999).

3.2. Ecosystem models

We compare three nitrogen-based ecosystem model versions, which are shown schematically in Fig. 2 and referred to as M1, M2 and M3 in increasing order of complexity. M1 has previously been used in the Mid-Atlantic Bight, a region south of our model domain (Fennel et al., 2006). M3 is based on the North Pacific Ecosystem Model for Understanding Regional Oceanography (NEMURO) structure (Kishi et al., 2007). This model was chosen because an optimized version of NEMURO has been shown to outperform simple models in the California Current area (Mattern et al., 2017). M2 represents an intermediate step between M1 and M3.

M1 and M2 have 7 compartments tracking nitrate, ammonium, phytoplankton, chlorophyll, zooplankton, and two detritus size classes. In M1, only phytoplankton growth depends on temperature (Eppley, 1972). In M2, we introduced temperature dependency in other biological rates (i.e., phytoplankton mortality, zooplankton grazing, excretion and mortality). In M3, we further increased ecological complexity by adding plankton functional groups. M3 has 11 compartments that include 2 nutrient and 2 detritus pools similar to M1 and M2, 2 phytoplankton groups (representing small and large phytoplankton), and 3 zooplankton groups (small, large, and predatory zooplankton). While the trophic structure of M3 (i.e., the interactions among planktonic groups) is based on NEMURO, it utilizes the same functional forms as our M1 and M2 model versions (e.g., Holling III grazing, as in Fennel et al., 2006), instead of the Ivlev equation used in NEMURO) for sake of better comparability.

In summary, the three model versions we compare introduce additional ecological complexity in a step-wise fashion. In this way, we aim to tease apart the effects of increasing the dependency of the system on environmental factors, such as temperature, and increasing the trophic complexity itself. The equations for the three model versions are included in Appendix I.

Boundary and initial conditions for NO_3 are based on a monthly climatology constructed using *in situ* observations (see Section 4.1) and World Ocean Atlas monthly averages (Garcia et al., 2010). Initial and boundary conditions for all other biological variables are set to 0.1 mmol N m^{-3} as in Fennel et al. (2006, 2008). These variables adjust on short time scales (days); hence, the system has no memory of the initial values after a short spin-up phase. A phytoplankton-to-chlorophyll ratio of 0.76 mmol N (mg Chl)$^{-1}$ is assumed for the chlorophyll initial and boundary conditions (Bianucci et al., 2015).

3.3. Surrogate model

We apply a simple 1D framework to the 22 locations presented on Fig. 1, which are referred to as the "1D models" from now on. In general, the 1D models solve a vertical diffusion term $k_D \frac{\partial^2 C}{\partial z^2}$ using the Crank–Nicolson scheme, where k_D is the diffusivity, z is depth, and C refers to the biological state variables. The vertical resolution is 5 m, and the vertical grid is divided into two distinct layers with respect to mixing: a turbulent surface mixed layer (layer 1) and a quiescent layer below (layer 2). The interface between both layers is determined by the time-varying mixed layer depth, which is estimated using a criterion for the maximum density gradient. For this purpose, the density field is obtained from a base run of the 3D model. In 1D, a high diffusivity is assigned to all grid cells above the prescribed mixed layer depth ensuring complete mixing within the mixed layer on a time scale of 1 day with a minimum diffusivity of 100 m^2 d^{-1} imposed ($k_{D1} = \max[\mathrm{MLD}^2\ d^{-1}, 100] = 1.2 \times 10^{-3}\ m^2\ s^{-1}$). A lower diffusivity ($k_{D2} = k_{D1} \times 10^{-2}$) is assigned to all grid cells below the mixed layer depth. This 1D framework has been previously used (Lagman et al., 2014; Kuhn et al., 2015).

The 1D models also require shortwave radiation and temperature as inputs. The shortwave radiation is the same as in the 3D model and comes from the European Centre for Medium-range Weather Forecasts (ECMWF) global atmospheric reanalysis (ERA-Interim) (Dee et al., 2011). Temperature is taken from the 3D model base simulation. As 1D models do not include horizontal advection, NO_3 below the mixed layer depth is nudged to the 3D results from the 3D base simulation with a nudging time scale of 60 days. This nudging scheme avoids direct impacts on the photic surface layers, where it is assumed that biological activity has the strongest effect on nitrate. Since we focus on temporal changes in concentrations within the upper mixed layer we disregard the possible, but presumably small, effects of a few parameters on reshaping the vertical distribution of nitrate in the photic surface layers. Therefore, there is no conflict between the nudging treatment and the optimization (i.e., deep nitrate does not change with changes in the parameter values on the timescales considered here). Total depth in the 1D models is equal to the depth in the 3D model or truncated at 50 m below their maximum mixed layer depth, whichever is shallower. This

Table 1
List of 1D model locations used in the optimization exercises. Surface *in situ* NO_3, *in situ* Chl-a and satellite Chl-a are observed mean surface concentrations between 1999 and 2001. The 3D model depth differs from the 1D model depth when the bottom is deeper than the maximum mixed layer depth plus 50 m. Station numbers and transect acronyms are as in Fig. 1.

Station number	and transect acronym	Lat.	Lon.	3D model depth	1D model depth	Max MLD	Min MLD	*in situ* NO3(surface)	*in situ* Chl-a(surface)	satellite Chl-a
1	ANT	49.44	−64.61	315	92	42.4	2.5	5.59 ± 3.14	0.98 ± 1.57	1.02 ± 0.8
2	BBA	49.69	−59.25	214	117	68.0	9.5	5.27 ± 4.41	0.66 ± 1.17	0.38 ± 0.29
3	BON	48.89	−52.47	276	276	247.9	9.1	6.76 ± 6.45	1.49 ± 2.84	0.87 ± 1.00
4	BON	49.21	−51.48	295	171	121.8	10.0	5.45 ± 4.38	1.25 ± 2.34	0.78 ± 1.03
5	BON	49.52	−50.50	308	172	122.3	2.1	5.50 ± 4.43	1.17 ± 2.14	0.78 ± 0.81
6	CAB	47.27	−59.77	421	128	79.0	4.5	4.89 ± 3.88	1.36 ± 2.46	1.24 ± 2.24
7	FLC	47.00	−49.50	77	77	68.5	13.0	3.93 ± 3.49	1.75 ± 3.07	1.26 ± 2.93
8	FLC	47.00	−50.83	127	127	89.0	10.0	4.79 ± 3.68	1.52 ± 2.83	0.63 ± 0.76
9	FLC	47.00	−52.17	128	128	85.0	9.0	4.78 ± 4.13	1.47 ± 2.72	0.57 ± 0.32
10	HAL	42.83	−61.74	1180	116	66.9	9.0	4.79 ± 4.11	1.39 ± 2.64	0.67 ± 0.67
11	HAL	43.46	−62.43	83	83	71.0	10.0	4.05 ± 3.70	1.49 ± 2.68	0.72 ± 0.50
12	HAL	44.09	−63.13	162	103	53.5	9.0	4.48 ± 3.94	1.33 ± 2.51	0.92 ± 0.61
13	IMA	47.31	−62.63	63	63	61.0	8.5	3.26 ± 3.21	1.59 ± 2.73	0.68 ± 0.89
14	LOU	43.86	−57.89	2443	125	75.9	5.1	4.71 ± 4.10	1.29 ± 2.47	0.67 ± 0.49
15	LOU	44.65	−58.69	81	81	56.0	2.1	3.93 ± 3.59	1.50 ± 2.71	0.87 ± 0.63
16	LOU	45.44	−59.48	107	107	62.5	7.6	4.59 ± 3.96	1.42 ± 2.74	1.24 ± 1.00
17	SAB	42.09	−65.35	1110	118	69.0	3.0	4.62 ± 3.94	1.39 ± 2.69	0.71 ± 0.42
18	SAB	42.56	−65.41	105	105	82.9	2.0	4.62 ± 3.93	1.36 ± 2.63	0.84 ± 0.65
19	SAB	43.02	−65.47	104	104	64.0	3.0	4.65 ± 3.91	1.34 ± 2.58	0.97 ± 0.75
20	SEG	43.76	−50.63	62	62	59.5	5.2	3.04 ± 3.03	1.69 ± 2.84	0.65 ± 0.57
21	SEG	44.89	−51.55	62	62	61.5	2.0	2.94 ± 3.00	1.71 ± 2.83	1.09 ± 2.19
22	SEG	46.02	−52.47	83	83	82.0	9.5	3.63 ± 3.46	1.54 ± 2.69	0.79 ± 1.03

treatment further reduces the surrogate computational time, without affecting its performance. Conditions for year 1999 are repeated at the beginning of the 1D model run as a model spin-up. Acronyms, geographical coordinates, depths, mean temperature, chlorophyll-a, and nitrate values for each location are presented in Table 1.

Despite its simplicity, the mechanistic emulator replicates the results of the full 3D model well for all three biogeochemical models. In Fig. 3, we use 2D histograms to compare 5-day averages of surface chlorophyll simulated by the 3D (ROMS) and the 1D models configured with the initial parameter guess. The target model (ROMS) and surrogate surface chlorophyll results are significantly correlated ($p<0.01$), with correlation coefficients of 0.77, 0.84 and 0.60, for M1, M2, and M3, respectively. Nevertheless, the surrogate of M3 is challenged to replicate low chlorophyll values and tends to overestimate them. Differences in chlorophyll (1D minus 3D) before the optimization are shown for two locations, BON (49.21°N 51.48°W, location 4) and HAL (43.46°N 62.43°W, location 11) in Fig. 4. All model versions exhibit discrepancies at the beginning and end of the mixed layer-shoaling period. Biases in the position of the deep-chlorophyll maximum occur during summer stratification, with the 1D models predicting a shallower position than the 3D model.

4. Optimization procedure and sensitivity analysis

The optimization is implemented using an evolutionary algorithm and applied for 3 years (January 1999–December 2001). The evolutionary algorithm simulates a process of natural selection by imposing a survival of the fittest strategy (Houck et al., 1995) on a population composed of different parameter sets, which represent individuals within the population. Evolution takes place through random recombinations and mutations of the parameter sets. A priori estimates of the minimum and maximum values have to be specified for each parameter to avoid obtaining unrealistic values (Table 2). Details of the evolutionary algorithm used here are described in Kuhn et al. (2015). We optimized subsets of the complete parameter sets required by each model version (Table 3), which were selected based on the sensitivity analysis described in Section 4.2.

4.1. Observational datasets for calibration

Satellite and *in situ* observations were used to calibrate the models. Surface chlorophyll satellite observations come from the Sea-viewing Wide Field-of-view Sensor (SeaWiFs) 8-day averages with 9-km resolution. *In situ* observations were obtained from the Atlantic Zone Monitoring Program (AZMP, http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/azmp-pmza/index-eng.html), which performs biannual monitoring along the 13 transect lines shown in Fig. 1. The AZMP dataset includes quality controlled CTD measurements of temperature and salinity (Mitchell et al., 2002) from which density was calculated using the Gibbs SeaWater TEOS-10 oceanographic toolbox. Bottle measurements used in this study include *in situ* chlorophyll-a and nitrate, as these are variables with direct counterparts in the model. The standardized chlorophyll-a analysis method is Turner fluorometry and the nitrate analysis is colorimetric on a Technicon AutoAnalyzer II (AA II) segmented flow analyzer (Mitchell et al., 2002).

Satellite observations were first validated against the *in situ* chlorophyll-a observations from the top 3 m by identifying all matching records between 1997 and 2010 (i.e., the duration of the SeaWiFs record). This matchup analysis was implemented by first searching all *in situ* observations available within every 8-day window of the satellite record. Then, matching satellite records were averaged within a 0.1-degree radius of their corresponding *in situ* measurement. Using vertical averages over the top 3 meters increased the number of match-ups and did not significantly affect the regression, compared to using only the top 1 meter (Table i, Supplement I.). Additionally, the same matchup analysis was performed using GlobColour (http://hermes.acri.fr/; a combined MODIS and SeaWiFS product), and non-standard AZMP measurements of *in situ* HPLC (High Performance Liquid Chromatography) chlorophyll-a. The results of these analyses reveal the same patterns of satellite performance as the SeaWiFs vs. standard measurements (Supplement I.). The comparison shows systematic biases at certain locations, with the most pronounced bias at locations inside the Gulf of St. Lawrence. Bias correction is an essential step when merging *in situ* and satellite data sets (e.g., Smith et al., 2008). Therefore, based on the relationship between *in situ* chlorophyll and satellite chlorophyll (see Supplement I.), we defined the bias between satellite chlorophyll and *in situ* chlorophyll at locations inside the Gulf of St. Lawrence as a function of satellite chlorophyll concentration (Fig. 5A). To correct this bias, satellite time series inside the Gulf of St. Lawrence were debiased by subtracting $e_{GoSL} = 0.01 + 0.19x^{1.42}$, where x is the log-transform SeaWiFs satellite observation and e_{GoSL} is the bias (Fig. 5A).

In order to provide observed counterparts to both small and large phytoplankton groups in M3, we estimated the chlorophyll-a fractions

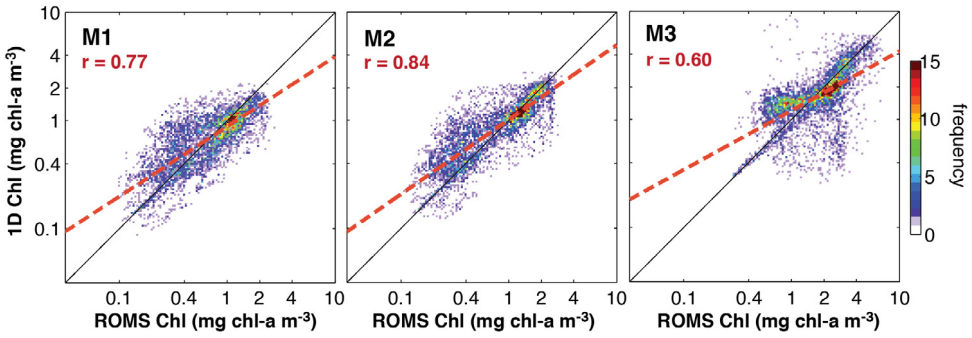

Fig. 3. Performance of the model surrogates with respect to surface chlorophyll for the three ecosystem model versions. Red dashed line indicates the regression line corresponding to correlation coefficients values shown in the upper left corner of each subplot.

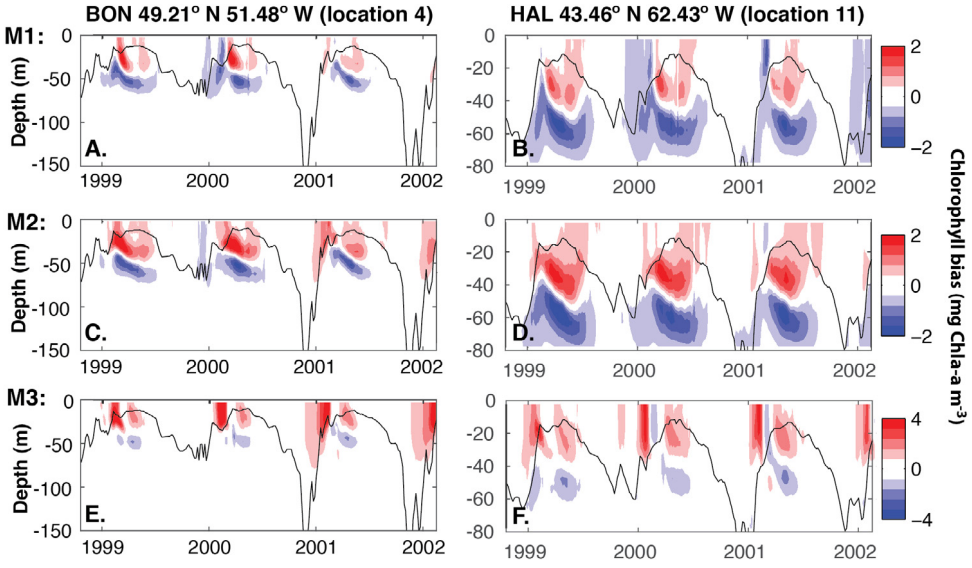

Fig. 4. Surrogate-target chlorophyll biases (1D minus 3D results) at two locations, using the three model versions. The black line represents the mixed layer depth. Note that the color scale and depth *y*-axis change in the subplots.

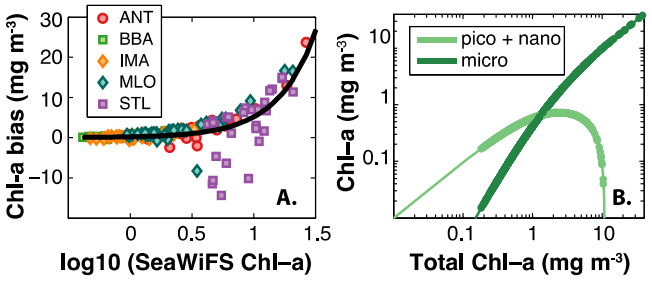

Fig. 5. Information used in the design of the cost function: A. Relationship between satellite chlorophyll and *in situ* versus satellite chlorophyll bias in locations of the Gulf of St. Lawrence. Symbols represent observations at different AZMP monitoring transects (see Table 1, Fig. 1). The black line shows the fitted de-biasing function $e_{GoSL} = 0.01 + 0.19x^{1.42}$, where x is the log-transform SeaWiFs satellite chlorophyll value and e_{GoSL} is the bias with respect to *in situ* observations. B. Chlorophyll concentrations of small (pico- and nano-) and large (micro-) phytoplankton estimated from satellite chlorophyll, following Hirata et al. (2011). Lines correspond to the Hirata et al. (2011) fractionation functions, and dots plotted on the lines correspond to chlorophyll fraction values calculated for observations in our study area.

from small and large phytoplankton in the satellite observations using the empirical relationships of Hirata et al. (2011). This study provides a set of equations and coefficients to estimate the chlorophyll concentration of various phytoplankton size classes and functional groups based on a global classification of HPLC pigment data into phytoplankton size classes. Here, we specifically used their equation to estimate the fraction of chlorophyll corresponding to microphytoplankton (Ψ):

$$\Psi = \left[\psi_0 + \exp\left(\psi_1 x + \psi_2\right)\right]^{-1}, \quad (1)$$

where x is the log-transform SeaWiFs satellite observation, and the coefficient values are $\psi_0 = 0.9117$, $\psi_1 = -0.27330$, and $\psi_2 = 0.4003$. We regard this fraction of chlorophyll as an observational counterpart of our large phytoplankton chlorophyll component; the remaining fraction (nano- and picophytoplankton) is considered the counterpart of small phytoplankton chlorophyll (Fig. 5B). Since Hirata et al.'s (2011) relationship was designed with SeaWiFs observations, we cannot apply the same formula to *in situ* measurements with any confidence.

4.2. Sensitivity analysis

In order to identify the most sensitive parameters and reduce the parameter space to be searched during optimization, the 1D models were rerun after perturbing each parameter one at a time. A reduced parameter space is desirable because parameters that are insensitive to the observations used in the optimization cannot be estimated. The sensitivity of the models to each of their parameter values is estimated as:

$$Q(Y,p) = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left|Y_{test} - Y_{base}\right|}{Y_{base}}, \quad (2)$$

which is the sum of the normalized absolute differences in the model state variables (Y), between the results of a base simulation (Y_{base}) and a test simulation (Y_{test}), where n is the total number of values compared. Q is calculated by varying each parameter p in $m = 5$ different tests. The tests change each of the base simulation parameter values, presented in Table 2, to the minimum, 25%, 50%, 75% and maximum values of their corresponding ranges also shown in Table 2. These minimum and maximum parameter values are based on the literature and also imposed in the optimization algorithm as bounds to avoid unrealistic values (Kuhn et al., 2015; Ward et al., 2010). This sensitivity analysis is thus independent from the observations. We consider this is important because the observations do not provide vertical structure information with sufficient frequency. Most *in situ* profiles are around the spring bloom, missing important aspects of the variability that could affect the parameter sensitivity results. This sensitivity analysis also provides information about the model sensitivity across all variables, which is valuable for comparison against other studies. Nevertheless, the ranking of parameters used for selecting the parameters to be optimized considers only the variables that have an observational counterpart (i.e., only chlorophyll and nitrate).

Results of this sensitivity analysis are shown in Fig. 6. Each stacked bar shows the contribution of all model variables to the total sensitivity. In order to select the parameters most sensitive to the available observations, the parameters were ranked with respect to the chlorophyll and nitrate contributions to Q. In models M1 and M2 equivalent parameters have similar rankings: the 3 most sensitive parameters are the maximum phytoplankton growth (μ_0), the mortality (m_P, m_{P0}), and the coagulation rate (τ). The initial slope of the P–I curve of photosynthesis (α) and grazing rate (g_{max}, g_0) have different ranks in M1 and M2, but are among the six most sensitive parameters. In addition to being important for the estimation of chlorophyll, this subset of parameters also has a significant effect on zooplankton and detritus. M1 and M2 are also sensitive to the zooplankton base metabolic rate (l_{BM}, l_{BM0}) and the remineralization of small detritus (r_{SD}); however, these parameters dominantly affect zooplankton and detritus, which are not part of the observation data used in the optimization.

In M3, parameters related to small phytoplankton are more sensitive than those related to large phytoplankton, e.g., the most sensitive parameter is the reference maximum growth rate of small phytoplankton (μ_{0Ps}). There are some similarities in parameter ranking with the rankings of M1 and M2, e.g., the small phytoplankton mortality rate (μ_{0PS}), the grazing rate of small zooplankton on small phytoplankton (g_{0ZsPs}), and the coagulation rate (τ) are among the most sensitive. The most sensitive parameters of M3 are similar to those reported for NEMURO (Yoshie et al., 2007). In general, the most sensitive parameters appear similar among biogeochemical models. In a parameter sensitivity analysis of 12 different biogeochemical models, Friedrichs et al. (2007) that the maximum phytoplankton growth rate and the remineralization rate frequently appear among the most sensitive parameters.

The parameter rankings in Fig. 6 guide the selection of parameters to be optimized, as detailed in the next section and in Table 3.

4.3. Optimization experiments

We performed five optimization experiments (E1 to E5) for all three models and two additional experiments only for M3 (E4b and E5b). Each optimization experiment (E) utilizes a different cost function (J_E), depending on its objective. The different objectives of the experiments consider the number of variables included in the optimization, the number of locations evaluated, and the number of parameters optimized (Table 3). We define R_v as the weighted root mean square difference between the simulated ($\hat{y}$) and observed (y) values:

$$R_v = \frac{w_v^2}{N} \sum_{n=1}^{N} \left(\hat{y}_{l,n} - y_{l,n}\right)^2, \tag{3}$$

where the subscript v stands for the three observational data types in this study: satellite chlorophyll (R_{chl1}), *in situ* chlorophyll (R_{chl2}) and *in*

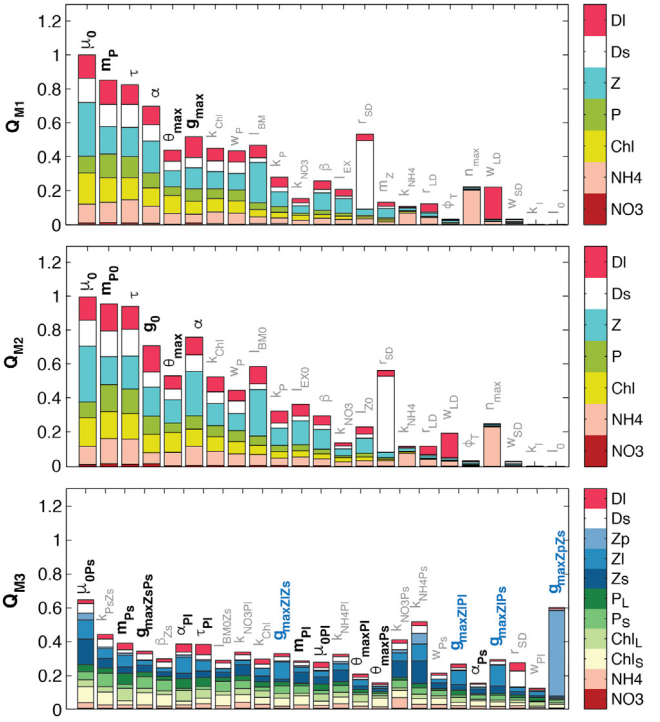

Fig. 6. Parameter sensitivities (Q) for M1, M2 and M3. Each stacked bar represents the total sensitivity of the model to a specific parameter, and is composed of the contribution of each model variables to the total. Different bar colors are assigned to each model state variable. Bars are ranked according to the parameter sensitivity to chlorophyll and nitrate variables. Parameters in black font are included in the main optimization exercise. In M3, parameters in blue font are grazing rates that were indirectly optimized (i.e., the ratio between all grazing rates was maintained constant). Parameters in gray font were not optimized in any of the model experiments, and were fixed to their a priori values (see Supplement II).

situ nitrate (R_{NO3}). N is the number of observed values, and the weight $w_v = \phi_v/\sigma_v$ of each variable at each location is inversly proportional to the standard deviation of its observations (Evans, 2003). A high weight was assigned to satellite chlorophyll by setting the weight coefficients $\phi_v = 3$ for satellite chlorophyll and $\phi_v = 1$ for the other two data types. The higher weight of satellite chlorophyll was determined during preliminary tests and was necessary due to the pronounced seasonal cycle in the region, resulting in a large standard deviation in surface chlorophyll values. If no additional weight was used, satellite chlorophyll became downweighted with respect to *in situ* chlorophyll profiles. The profiles could not fully capture chlorophyll temporal variability due to their significantly lower frequency. In the case of M3, R_{chl1} was modified to compare both the small and large phytoplankton chlorophyll fractions against corresponding satellite-derived fractions estimated using the Hirata et al. (2011) algorithm (see Section 4.1).

Experiments E1 and E2 are "joint" optimizations that assimilate chlorophyll data from all locations shown by the orange dots in Fig. 1. These experiments use cost functions J_1 and J_2, respectively:

$$J_1(\vec{p}) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{\omega_l} R_{chl1} \tag{4}$$

$$J_2(\vec{p}) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{\omega_l} \left(R_{chl1} + R_{chl2}\right) \tag{5}$$

where $\vec{p}$ refers to parameter vector, and $L = 22$ is the total number of 1D model locations (l). The location weighting factor $\omega_l = \frac{1}{L} \sum_{v=1}^{V} \frac{\overline{y}_v^2}{\sigma_v^2}$ uses the mean ($\overline{y}$) and standard deviation (σ) of the observed variables (v) to avoid biasing the cost towards locations with lower variability (Schartau and Oschlies, 2003a; Friedrichs et al., 2007).

Table 2
A priori model parameter estimates and ranges of the subset of parameters optimized in experiments E1 to E5. For M1, a priori parameter values were subjectively modified from Fennel et al. (2008) prior to the experiments in this study and used in Bianucci et al. (2015). For M2, we assume that the fixed biological rates are reference values at the average surface temperature in the domain (approximately $T = 10$ °C), and back-calculated the corresponding reference value at $T = 0$ °C. M3 uses the M2 parameters, except for zooplankton grazing rates. A full list of parameters can be found in the supplementary information.

Parameters		M1	M2	M3	Range	Units
Reference phytoplankton maximum growth rate at $T = 0$ °C (generic, small and large, respectively)	μ_0	0.28	0.28	–	0.1–3.5	d^{-1}
	μ_{0P_S}	–	–	0.28	0.1–3.5	
	μ_{0P_L}	–	–	0.28	0.1–3.5	
Initial slope of the P–I curve of photosynthesis for phytoplankton	α	0.025	0.025	–	0.007–0.13	mg C (mg Chl Watts m^{-2} day)
	α_{P_S}	–	–	0.025	0.007–0.13	
	α_{P_L}	–	–	0.025	0.007–0.13	
Phytoplankton mortality rate	m_P	0.03	–	–	0.01–0.25	d^{-1}
Reference phytoplankton mortality rate at $T = 0$ °C (generic, small and large, respectively)	m_{P0}	–	0.027	–	0.01–0.25	
	m_{PS0}	–	–	0.027	0.01–0.25	
	m_{PL0}	–	–	0.027	0.01–0.25	
Zooplankton maximum grazing rate	g_{max}	0.6	–	–	0.2–4	d^{-1}
Reference zooplankton maximum grazing rate at $T = 0$ °C (generic and small on small prey)	g_0	–	0.54	–	0.2–4	
	$g_{0Z_SP_S}$	–	–	0.54	0.2–4	
Phytoplankton and small detritus aggregation rate	τ	0.02	0.02	–	0.001–1	d^{-1}
Large phytoplankton and small detritus aggregation rate	τ_{P_L}	–	–	0.02	0.001–1	d^{-1}
Maximum chlorophyll to carbon ratio (generic, small and large)	θ_{max}	0.053	0.053	–	0.005–0.15	mg Chl (mg C)$^{-1}$
	θ_{maxP_S}	–	–	0.053	0.005–0.15	
	θ_{maxP_L}	–	–	0.053	0.005–0.15	

Table 3
Summary of optimization experiments, detailing observed variables included in the cost function, and the number of parameters optimized. The ranked sensitivity of parameter values here referred to is presented in Fig. 6.

			M1 & M2		M3	
Exp.	F(p)	Stations included	Observations in cost function	Optimized parameters	Observations in cost function	Optimized parameters
E1	J_1	All	R_{chl1}	6 most sensitive in M1 and their M2 equivalents: μ_0, m_P(orm_{P0}), τ, α, θ_{max}, g_{max}(or g_0)	Size fractionated R_{chl1}	M3 equivalents to 6 most sensitive in M1/M2: μ_{0P_S}, μ_{0P_L}, m_{PS0}, m_{PL0}, τ_{PL}, α_{PS}, α_{PL}, θ_{maxPS}, θ_{maxPL}, $g_{maxZsPs}$
E2	J_2	All	$R_{chl1} + R_{chl2}$	Like E1	Size fractionated $R_{chl1} + R_{chl2}$	Like E1
E3	J_3	Single	$R_{chl1} + R_{chl2} + R_{NO3}$	Like E1	Size fractionated $R_{chl1} + R_{chl2} + R_{NO3}$	Like E1
E4	J_4	All	Like E3	Like E1	Like E3	Like E1
E4b	J_4	All	(Performed for M3 only)	N/A	Like E3	6 most sensitive in M1/M2: μ_{0P_S}, k_{ZsPs}, m_{PS0}, $g_{maxZsPs}$, β_{Z_S}, α_{PL}
E5	J_4	All	Like E3	3 most sensitive in M1 and their M2 equivalents: μ_0, m_P(orm_{P0}), τ	Like E3	M3 equivalents to 3 most sensitive in M1/M2: μ_{0P_S}, μ_{0P_L}, m_{PS0}, m_{PL0}, τ_{PL}
E5b	J_4	All	(Performed for M3 only)	N/A	Like E3	3 most sensitive in M1/M2: μ_{0P_S}, k_{ZsPs}, m_{PS0}

Experiment E3 (Eq. (7)) corresponds to "single-site" optimizations (i.e., the optimization algorithm runs independently for each 1D model location). Note that the location-specific weight is not needed in this cost function:

$$J_3(\vec{p}, l) = \left(R_{chl1} + R_{chl2} + R_{NO3}\right) \tag{6}$$

The results of E3 were used to assess spatial patterns in the optimized parameters in a principal component analysis. This analysis assesses the similarities between parameter sets optimized for different locations, and identifies the parameters that are mainly driving such similarities. We also compared the portability (see Friedrichs et al., 2007) of our three model versions by optimizing the model at one single location, and then transferring these optimal parameters to the other stations.

Experiment E4 is a "joint" optimization where all 22 locations are included:

$$J_4\left(\vec{p}\right) = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{\omega_l}\left(R_{chl1} + R_{chl2} + R_{NO3}\right) \tag{7}$$

Comparison between the results of E3 and E4 allows us to evaluate the compromise required when fitting all observed variables at all locations using one common set of parameters (Section 5.2). Comparing results of E1, E2 and E4 aims to evaluate differences between single data type and multiple data types optimizations. All experiments from E1 to E4 aimed to optimize the 6 most sensitive parameters in M1 or their equivalents in M2 and M3. In the case of M1 and M2, the 6 most sensitive parameters are either the same or equivalent (Fig. 6). In the case of M3, equivalent parameters may be one or more. For example, the M3 equivalents of the phytoplankton reference growth rate μ_0, used in the single phytoplankton models (M1 and M2), are both the small and large phytoplankton reference growth rates μ_{0Ps} and μ_{0Pl}. Due to lack of observational constraints for zooplankton, only the most sensitive of the grazing rates is optimized ($g_{0Z_sP_s}$), keeping the ratio to the other six grazing rates constant. Notice that there are only three different estimates of the maximum grazing parameters that are assigned to seven parameters. Therefore, in essence, only three parameters to specify the grazing rates on M3.

Experiments E4b, E5 and E5b also use the cost function J_4. They evaluate optimizations for all compared variables, at all compared locations, modifying only the number and selection of parameters (Table 3).

In Experiment E5, we optimized only the top 3 most sensitive parameters of M1 and M2, or their equivalents in M3. Thus, the comparison between E4 and E5 evaluates how the number of optimized

parameters affects the results (Section 5.3). In preliminary tests, optimizing more than 6 parameters in M1 or M2 did not result in significant improvements.

By optimizing equivalent parameters, we aim to ensure an objective comparison between models. For instance, it has been theorized and shown that different models can produce similar fits to observations despite portraying different dynamics (Friedrichs et al., 2007; Quine, 1975). However, due to this parameter selection procedure, the number of optimized parameters in experiments E1 to E5 is higher for M3 than for M1 and M2. It could be argued that a better performance of M3 may be a consequence of more degrees of freedom. In order to address this issue, the additional experiments E4b and E5b, which were performed only for M3, replicate experiments E4 and E5, but using the same number of optimized parameters as for M1 and M2 (i.e. the 6 and 3 most sensitive parameters, respectively). In all cases, parameters not included in the optimization subsets are kept fixed at their initial guess value (Table 2). As the cost function is different for each experiment and thus non-comparable, we use J_4 as the function to evaluate differences between optimized models ($F(\vec{p}) = J_4$). This function compares all observational data types and all locations.

5. Results

We compare the results of the optimization experiments described in Table 3 using the cost function $J_4(\vec{p})$ (Eq. (7)) of each model version in 1D (Fig. 7). 3D simulations were performed only with parameters obtained in experiment E4 and their corresponding costs are also shown. Additionally, the optimized set of parameters obtained for E4 is presented in Table 4 as a reference.

Overall, model M3 presents lower costs than M1 and M2 in experiments E1 to E4, as well as in experiment E5b. However, in experiments E5 and E4b model M3 presents large model-data differences with respect to satellite chlorophyll. We describe and discuss these results in more detail in the following sections.

5.1. Single vs. multiple observed variables

Comparison of experiments E1, E2 and E4 illustrates the effects of optimizing the models against satellite chlorophyll alone (E1), versus including information about the vertical structure of chlorophyll (E2) and nitrate (E4). The inclusion of *in situ* chlorophyll profiles in the optimization (E2, Eq. (6)) degrades the performance of M1 with respect to satellite chlorophyll, but has no significant effect on the performance of M2 and M3. The inclusion of both chlorophyll and nitrate profiles (E4) results in lower model costs with respect to nitrate, but higher costs with respect to both satellite and *in situ* chlorophyll. Fig. 8A shows the optimized results of E4 in comparison to surface chlorophyll. *In situ* surface chlorophyll exhibits large ranges between October and January, whereas satellite chlorophyll appears less variable with relative constant low values. These months have the largest number of gaps in the satellite records (Fig. 8B).

5.2. Single vs. multiple locations

In experiment E3, optimized parameters were found for each location individually using Eq. (7). Allowing different optimal parameters for the different locations makes it easier to fit the individual patterns of variability, and thus a lower total cost is achieved. The results of experiment E3 are used to analyze whether spatial patterns in the biological parameters emerge (Section 5.2.1) and to evaluate model portability from one location to another (Section 5.2.2).

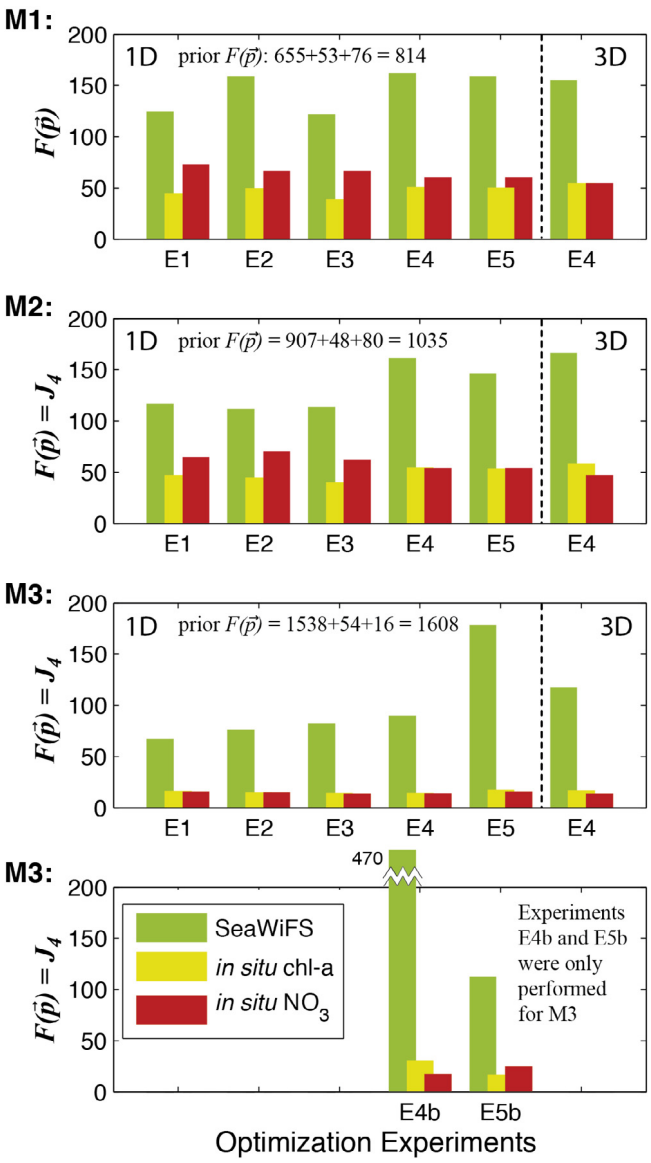

Fig. 7. Final cost metrics of the optimization experiments (see Table 3). However each optimization experiment uses a different optimization function (see Table 3), the optimization function J_4 is used to compare all experimental results. Optimized three-dimensional models were only run for the parameters obtained in E4.

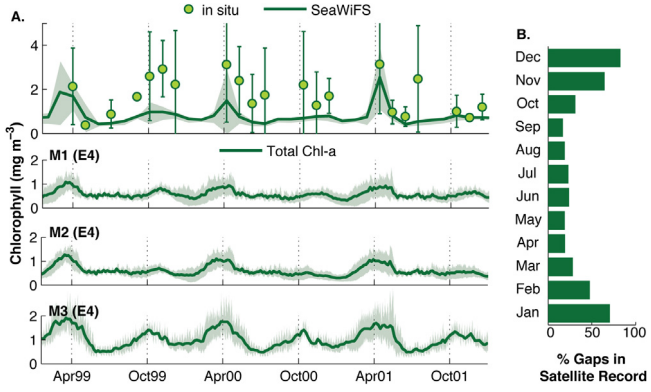

Fig. 8. A. Mean surface SeaWiFS satellite and *in situ* chlorophyll (top), compared to surface chlorophyll simulated by the three model versions (M1, M2, M3) with optimal parameters from experiment E4. Shading and errorbars represent the standard deviation between locations. B. Percentage of gaps in the satellite record per month.

Table 4
Model parameter estimates obtained from optimization experiment E4, which used a cost function including satellite and *in situ* Chl-a and *in situ* NO_3 for 22 locations in the study region.

Parameters	M1	M2	M3	Units
μ_0	1.11	1.14	–	d^{-1}
μ_{0P_S}	–	–	1.16	d^{-1}
μ_{0P_L}	–	–	1.12	d^{-1}
α	0.035	0.019	–	mg C (mg Chl Watts m^{-2} day)
α_{P_S}	–	–	0.041	mg C (mg Chl Watts m^{-2} day)
α_{P_L}	–	–	0.039	mg C (mg Chl Watts m^{-2} day)
m_P	0.13	–	–	d^{-1}
m_{P0}	–	0.063	–	d^{-1}
m_{PS0}	–	–	0.08	d^{-1}
m_{PL0}	–	–	0.04	d^{-1}
g_{max}	3.34	–	–	d^{-1}
g_0	–	1.62	–	d^{-1}
$g_{0Z_SP_S}$	–	–	2.32	d^{-1}
$g_{0Z_LP_S}$	–	–	1.16	d^{-1}
$g_{0Z_LP_L}$	–	–	0.39	d^{-1}
$g_{0Z_LZ_S}$	–	–	2.32	d^{-1}
$g_{0Z_PP_L}$	–	–	1.16	d^{-1}
$g_{0Z_PZ_S}$	–	–	1.16	d^{-1}
$g_{0Z_PZ_L}$	–	–	1.16	d^{-1}
τ	0.097	0.116	–	d^{-1}
τ_{P_L}	–	–	0.002	d^{-1}
θ_{max}	0.1	0.08	–	mg Chl (mg C)$^{-1}$
θ_{maxP_S}	–	–	0.03	mg Chl (mg C)$^{-1}$
θ_{maxP_L}	–	–	0.02	mg Chl (mg C)$^{-1}$

5.2.1. Spatial patterns in parameters

The analysis of the parameters optimized for individual locations may reveal spatial patterns with dominance of specific plankton groups in different areas. A principal component analysis was performed on the optimal parameter sets obtained for each model version (Fig. 9). In all model versions, the variability among locations is dominated by differences in the zooplankton grazing rates (PC1) and phytoplankton growth rates (PC2). Clearly defined clusters of locations are not identified by the analysis; however, some locations consistently arrange themselves along PC2 in all model versions. That is, some locations are consistently characterized by either high or low grazing rates. Spatial patterns in the grazing rates are, however, difficult to discern.

5.2.2. Model portability

Model portability experiments were performed in the 1D environment, by iteratively applying the optimized parameters from one location ("optimized model") to the rest of them ("test models"). Results are summarized in Fig. 10, where the cost of the test models is normalized by the corresponding optimized model cost. M2 has the largest percentage of test models with cost equal to or lower than the optimized model (M1: 23.3%, M2: 32.9%, M3: 17.4%). The highest percentage of tests with cost larger than the optimized run occur in M3 (M1: 76.7%, M2: 67.1%, M3: 82.6%); however, M1 presented the highest percentage of tests with cost larger than twice the optimized run (M1: 32.3%, M2: 25.5%, M3: 26.7%). According to these results, M2 can be considered the most portable of the three model versions, whereas M3 appears as the least portable. In particular for M3, it appears that model solutions optimized at open ocean sites perform poorly when transferred to shallower shelf-areas. Oppositely, when parameters are optimized at the shallow shelf-areas and then transferred to the deeper open ocean sites, the models do not perform as bad. Behavior is more symmetric for M2.

Fig. 11 shows an example of the portability experiments for two locations with contrasting oceanographic conditions: location 4 (BON 49.21°N – 51.48°W) in the Labrador Sea and location 11 (HAL 43.46°N – 62.43°W) in the Scotian Shelf. The satellite observations at the BON location have a distinct spring bloom peak, which is well replicated by M2 using either parameters optimized for this location (Fig. 11A) or for the HAL location (Fig. 11C). However, M3 can only replicate the annual peak when using the locally optimized set of parameters. The magnitude of the spring bloom at the HAL location is lower than at BON, the peak occurs earlier in the year, and other peaks of equal magnitude can occur at different times of the year. Due to this more irregular variability, both models are challenged to replicate the HAL location even when using locally optimized parameters. When locally optimized, both models appear calibrated to appropriately capturing the timing of maximum surface chlorophyll concentrations, such that large discrepancies with observations occur when the fall bloom is larger than the spring bloom, as in 1999. Locally optimized M2 favors maximum concentrations and produces lower than observed summer to fall concentrations. In contrast, M3 favors average concentrations, better capturing summer to fall concentrations but underestimating the spring bloom maxima. The HAL test run of M2 maintains the spring bloom peak timing, but overall increases concentrations with emphasis on the fall. M3 generates a well-defined spring bloom of shorter duration.

5.3. Number of optimized parameters

Experiments E4, E5, E4b and E5b aim to evaluate the effect of increasing or decreasing the number of optimized parameters on the optimization success. The cost metric results (Fig. 7) show that optimizing 3 versus 6 parameters does not significantly affect the cost function value of M1 and M2. This is consistent with the results from the parameter sensitivity analysis, where the top 3 most sensitive parameters in M1 and M2 present a dominant effect on the model results in comparison with the rest of parameters (Fig. 6A, B). In contrast, M3 was more evenly sensitive to all parameters (Fig. 6C), and thus the number and choice of parameters to include in the optimization significantly affects the model results. For example, in experiment E4 a total of 10 parameters were optimized for M3. Those included sensitive parameters for all of the phytoplankton and zooplankton groups, while the subset of 6 parameters optimized in E4b only included one of the large phytoplankton parameters (α_{P_L}). The optimization results of E4b successfully replicate the average small phytoplankton background concentrations. However, they fail to replicate the blooming of the large phytoplankton group,

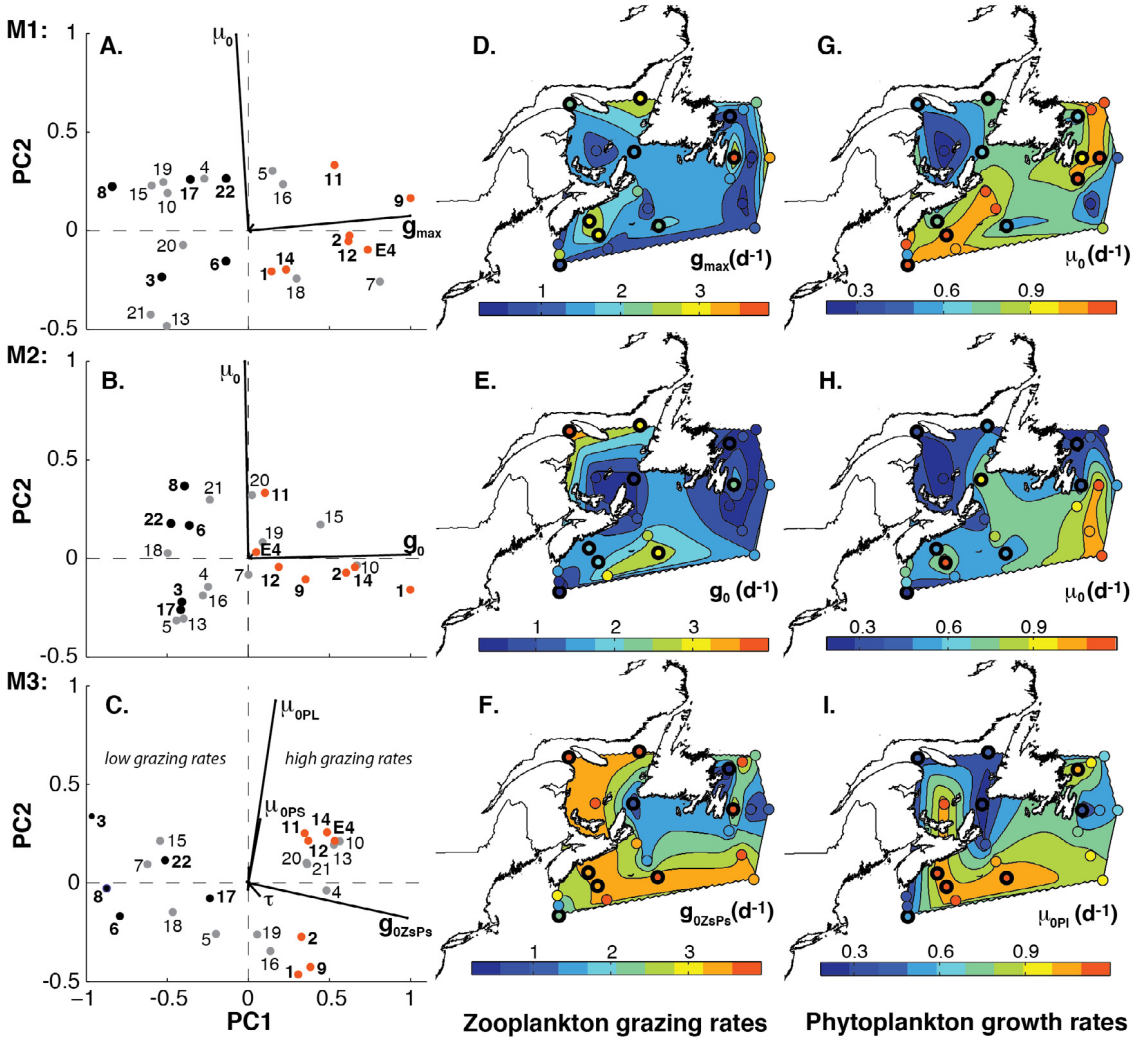

Fig. 9. Analysis of spatial patterns in optimal parameters for individual locations: Subplots A. to C. show the principal components analysis for the three model versions in the study. Numbers refer to the locations as depicted in Fig. 1. Some locations tended to consistently arrange themselves along PC1. Locations in blue are in the negative side for all three model versions, whereas locations in orange are always on the positive side. All other locations are shown in gray. Subplots D. to I. show the spatial distribution of the optimized parameters that dominated the variability on PC1 (reference zooplankton grazing rates), and PC2 (reference phytoplankton maximum growth rates). Circles of the locations with consistent behavior on the PC analysis have thick edges. For visualization purposes, the background contours show the linear interpolation of the corresponding parameter values.

during which the biomass of large phytoplankton exceeds temporarily that of small phytoplankton (Fig. 12). Calibrating the initial slope of the P–I curve of photosynthesis of large phytoplankton was insufficient to obtain a realistically high proliferation of large phytoplankton in spring. This illustrates that in addition to an appropriate response to light, differentiation between the responses to nutrient availability is fundamental in multi-species models (e.g., gleaners vs. opportunists).

The cost of M3 in experiments E5 (5 optimized parameters) and E5b (3 optimized parameters) is within the range of those for M1 and M2, but the parameters obtained by these experiments generate unintended trophic dynamics where some functional groups become extinct in the model. The diagrams in Fig. 12B summarize these emergent structures. In E4b, predatory zooplankton (Z_P) disappear due to a combination of low prey biomass and low grazing rates. As large phytoplankton was not properly replicated, large zooplankton growth became inhibited by low prey densities, and both low large phytoplankton and low large zooplankton biomass affected predatory zooplankton. In E5, grazing rates were not part of the optimized parameters and did not scale with increasing phytoplankton growth rates. This resulted in the functional extinction of small zooplankton, while the optimization attempted to match zooplankton losses by increasing mortalities and coagulation rates. The negligible biomass of small zooplankton cascaded to the total extinction of large and predatory zooplankton. In experiment E5b, M3 essentially becomes a P–Z model similar to M1 and M2, due to the extinction of large and predatory zooplankton and the low concentrations of large phytoplankton.

5.4. Fluxes

The choice of parameters of M3 in experiments E1 to E5 was intended to optimize comparable fluxes among all model versions. Nonetheless, differences in the resulting gross fluxes between variables are present between the 7-variable models and the 11-variable model. Fig. 13 shows vertically integrated zooplankton grazing, phytoplankton growth (new and regenerated production), mortality and coagulation fluxes obtained for the models using parameters from E4. Differences in the new production fluxes are negligible between M1 and M2, but M1 presents a slightly higher annual peak in the regenerated production and grazing. The effect of temperature dependency on the phytoplankton mortality rates is noticeable during fall and winter, where M2 has lower rates than M1. In contrast to M1 and M2, M3 has more defined peaks in new production and more extended periods with high regenerated production. Grazing by small zooplankton is lowest in M3. Peaks in grazing by large and predatory zooplankton exceed

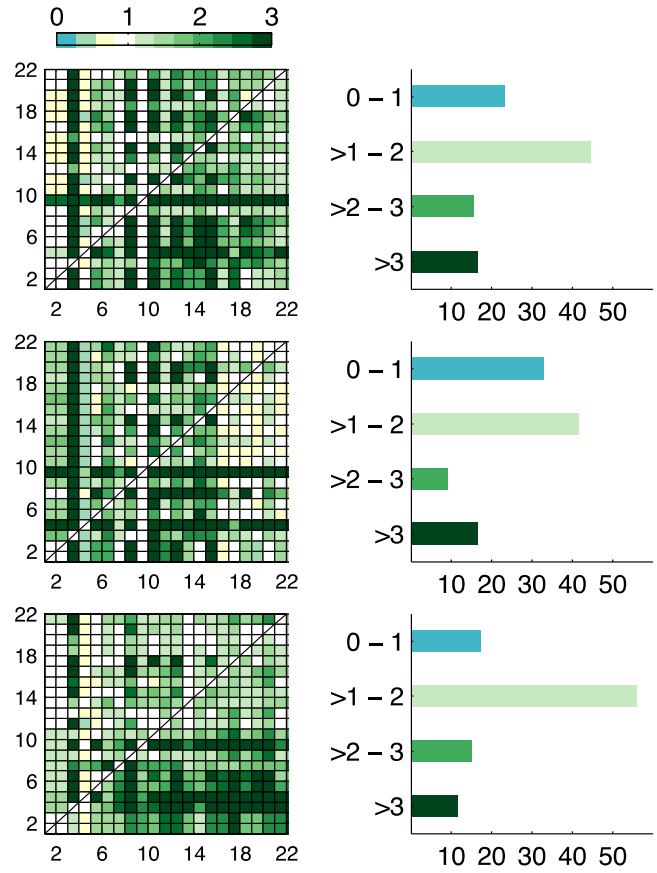

Fig. 10. Results of the portability experiments. On the left panels, the color scale represents the cost of running the 1D model at locations on the *y*-axis (test models), using parameters optimized for locations on the *x*-axis (optimized models). Cost (F) values have been normalized to the corresponding optimal cost, such that all optimized models have a cost equal to one (bins along the diagonal). On the right panels, bars summarize the results in four cost categories.

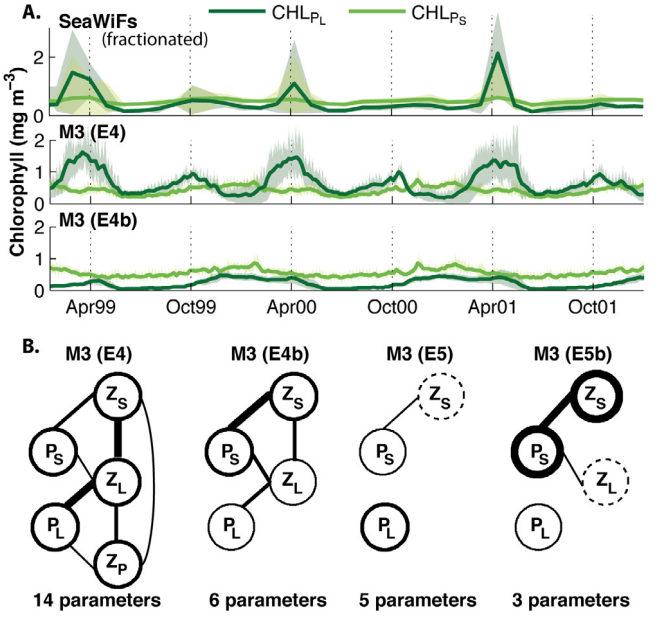

Fig. 12. A. Comparison of satellite-derived size-fractionated mean surface chlorophyll, and their model counterparts in M3 from optimization exercises E4 and E4b. B. Diagrams depicting the trophic model structures of M3 resulting from different optimization exercises. The line thickness of the circles' edges is proportional to the plankton group mean biomass, whereas the thickness of connecting lines is proportional to the fluxes between them. Dashed lines depict groups that have negligible biomass, but that are still part of the model dynamics by receiving a small but not negligible flux of nitrogen (i.e., functionally extinct plankton groups). Groups with negligible biomass and fluxes are removed from the diagram.

the grazing rates in M1 and M2 by approximately 2 mmol m^{-2} d^{-1} on average. Mortality of large phytoplankton is twice the phytoplankton mortality flux in M1 and M2 during spring and summer, but the same as M2 during winter. The coagulation flux of large phytoplankton is negligible.

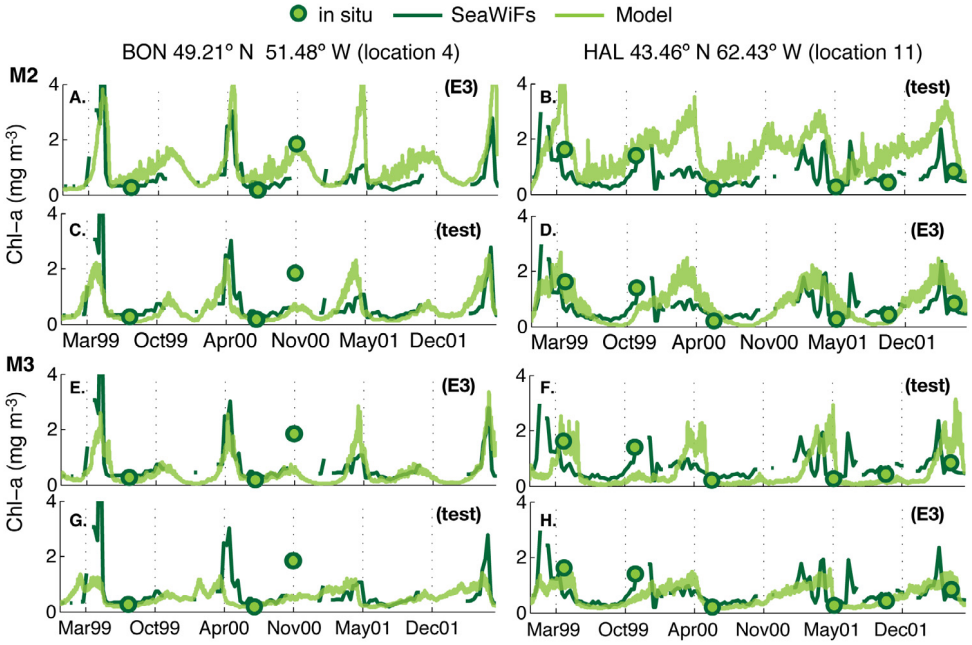

Fig. 11. Example of the portability experiments, showing satellite, *in situ*, and simulated surface chlorophyll at locations BON and HAL. Subplots A. to D. correspond to results of M2, whereas subplots E. to H. correspond to results of M3. Subplots A., D., E., and H. show results of models optimized for their corresponding location. Subplots B., C., F., and G. are model results using parameters optimized for a different location.

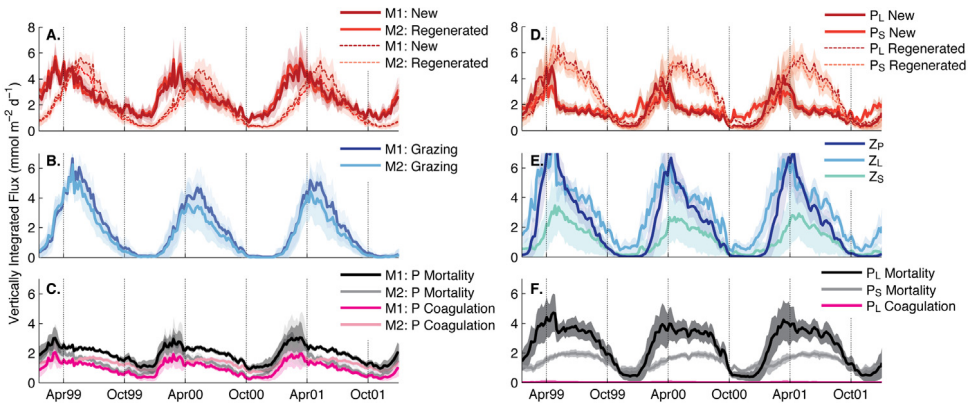

Fig. 13. Vertically integrated phytoplankton new and regenerated growth, zooplankton grazing, and other phytoplankton sinks (mortality and coagulation). Subplots A. to C. show the fluxes in the 7-compartment models (M1 and M2). Subplots D. to F. show the fluxes in M3 for all plankton groups.

6. Discussion

6.1. Surrogates and surrogate-based optimization

Simplified models allow us to avoid the computational expense of 3D models when performing sensitivity tests and calibrations. Here, a simplified 1D physical framework was shown to replicate key aspects of the results of a 3D regional application at selected locations, using three different ecosystem model versions (Figs. 3, 4). After optimization, the model-data misfit was reduced in both 1D and performed similarly well in 3D applications (Supplement II, Fig. 7). Similar types of site-based or test-bed calibrations of marine ecosystem models using 1D models have previously been shown to improve the predictive skill of 3D models (Kane et al., 2011; McDonald et al., 2012; Oschlies and Schartau, 2005). In many cases, the 1D models are built to represent averaged conditions at a climatological scale or over a relatively large spatial area (Dadou et al., 2004; Losa et al., 2004; Matear, 1995; Schartau and Oschlies, 2003b). This is intended to reduce the effects of phase biases that result from noise in both the observations and models, and of the inability of models to replicate as much variability as is displayed in the observations (Hemmings et al., 2015; Leeds et al., 2012; Schartau and Oschlies, 2003a). We did not use climatological or spatial averaging; that is surrogate-target and model-data comparisons were done on a site-by-site and date-by-date basis. Although this can be considered more challenging, the surrogates were able to capture between 36% and 70% of the variance in surface chlorophyll estimates of the 3D model sample (Fig. 3).

The surrogate-based optimization was successful in improving the performance of the three ecosystem model versions (Supplement II, Fig. 7). However, there are some systematic differences between 1D and 3D models in terms of the position of the deep chlorophyll maxima in summer. Overall, the deep chlorophyll maxima are shallower in the 1D models than in the 3D model (Fig. 4), probably a consequence of our simplified two-layer vertical structure of turbulence in 1D. Similar discrepancies in the position and extend of the deep chlorophyll maximum have been previously noted in 1D models (Doney et al., 1996; Fasham et al., 1993), including 1D calibration studies (Schartau and Oschlies, 2003b). In the latter, the vertical diffusivities in the 1D model were directly taken from the target model. This suggests that biases in the deep chlorophyll maxima may be inherent to 1D models and are not entirely due to the specific oversimplification of the diffusive component applied here.

The mechanistic surrogate approach also allowed us to identify features in the variables of interest that are likely dominated by the biological module from those controlled by the physics. In our case, the timing of the peak of the spring bloom was overall well captured by the 1D models (Figs. 8, 11). This indicates that this phytoplankton phenological characteristic is well constrained by the observations used during the optimization, and sensitive to the choice of parameters optimized.

6.2. Deciding on a complexity level

A number of previous attempts to assess the most appropriate level of ecosystem complexity in models have been inconclusive (e.g., Matear, 1995; Dadou et al., 2004); while others argue that there is a humpback relationship between model complexity and performance, with intermediate complexity models presenting advantages over both simple and more complex models (Fulton et al., 2003; Raick et al., 2006; Xiao and Friedrichs, 2014a). When models do not have comparable equations and are not optimized, differences in their performance are likely related to the parameter selection and functional equations, rather than the model structure itself (e.g., Sailley et al., 2013). In a model assessment study by Kriest et al. (2010), it was demonstrated that increasing complexity of unoptimized models does not necessarily improve model performance. Another recent model skill assessment by Kwiatkowski et al. (2014) found no evidence that biological complexity could consistently improve all aspects of model performance in reproducing observed global-scale bulk properties of ocean biogeochemistry. The simple models performed better in terms of global spatial pattern correlations of pCO2, dissolved inorganic carbon and alkalinity, but complex models better captured the monthly and annual variance of DIC and correlation coefficients of chlorophyll and primary production (Kwiatkowski et al., 2014). Insufficient observational data may make it difficult to justify the use of more complex models over the commonly used nutrient–phytoplankton–zooplankton–detritus (NPZD) model (e.g., Matear, 1995; Bagniewski et al., 2011). For example, in Bagniewski et al. (2011) none of the model variants compared could be rejected based on their misfit against constraining observations; however, they generated significantly different estimates of the unconstrained export carbon fluxes. It has also been shown that systematically removing some of the unconstrained aspects of an ecosystem model does not significantly increase the minimum value of the cost metric (Ward et al., 2013).

In our results, the more complex (11-compartment) model M3 is able to generate the lowest model-data misfits in all optimization experiments where the intended model structure is preserved. Similarly, a previous comprehensive comparison of 12 individually optimized marine ecosystem models, by Friedrichs et al. (2007), showed that models with multiple phytoplankton groups outperformed the single phytoplankton group models. In our results, M3 particularly exhibits reduced differences against the observed chlorophyll and nitrate vertical distributions. The simpler (7-compartment) model structures have a higher cost, but are also able to capture the averaged seasonal variations in surface chlorophyll. Therefore, if the objective of a modeling study is to characterize an averaged seasonality in surface chlorophyll, a simple model may suffice. This is supported by previous optimization studies using NPZD models and observational climatologies from the

North Atlantic that have been able to significantly reduce model-data differences (Kuhn et al., 2015; Schartau and Oschlies, 2003a).

Even simpler biological models that do not include explicit representation of phytoplankton and zooplankton have been parameterized and optimized to represent global biogeochemical properties as well as complex models (Galbraith et al., 2015; Kriest, 2017). In these examples, the most complex models analyzed were successfully downscaled by carefully parameterizing all processes omitted in the simplified reduced model (Galbraith et al., 2015; Kriest et al., 2017).

Complex models have an obvious utility in the study of specific plankton traits (e.g., Kuhn et al., 2018), trophic interactions (e.g., D'Alelio et al., 2016), species distribution and diversity (e.g., Barton et al., 2010), and other complex ecological processes. According to our results, complex models may also be better able to capture vertical distributions. However, when applied to a different location than the one it was calibrated for, our model with multiple phytoplankton groups tended to maintain chlorophyll magnitude characteristics from its original location.

One aspect affecting the portability of the multiple phytoplankton model was the use of satellite-derived fractionated surface chlorophyll to compare against the simulated chlorophyll of small and large phytoplankton groups. The particular satellite fractionation method we used is mainly based on chlorophyll concentrations, such that low chlorophyll is interpreted as a dominance of small phytoplankton and high chlorophyll is interpreted a predominant bloom of large phytoplankton. If the model was calibrated for a location with low chlorophyll, the parameters selected for large phytoplankton may not be adequate for locations with high chlorophyll. Non-overlapping ranges of small and large phytoplankton growth parameters could be configured in the optimization algorithm to correct this problem. These results are consistent with geographical portability experiments performed using 1D models for 4 locations in the Mid-Atlantic Bight, south of our study area (Xiao and Friedrichs, 2014b). Xiao and Friedrichs (2014b), used a cost function that included satellite-derived size-fractionated chlorophyll and satellite-derived POC. In almost all cases, when parameters fitted to one location were tested in the other locations, the cost increased significantly (Xiao and Friedrichs, 2014b). Nevertheless, the use of size-fractionated chlorophyll was beneficial and reduced cost when optimizing all locations simultaneously (Xiao and Friedrichs, 2014b).

We also note that when optimizing models with multiple phytoplankton groups the value of a cost function based on total chlorophyll could be misleading. Some phytoplankton groups may become extinct during the optimization process, thus altering the intended model structure. Here we decided to use an estimate of size-fractionated surface chlorophyll. The independent constraining of small and large phytoplankton may affect the portability of the more complex model when calibrated for individual locations (Figs. 10, 11), as it tends to benefit one phytoplankton group over the other depending on the chlorophyll abundance patterns of the specific locations. Lower predictive ability in complex models has been posited to occur when the model becomes over-fitted to noise in the data (Friedrichs et al., 2006). We show that this indeed can happen when a model with multiple phytoplankton groups becomes too specific to its training data set. Therefore, to benefit from the higher performance of a site-based calibrated complex model, such calibration requires including samples from all biogeographical provinces to be simulated. A similar conclusion was reached by Xiao and Friedrichs (2014b). They found that successful optimization results for 1D models of the Mid-Atlantic Bight could be found using size-fractionated chlorophyll and particulate organic carbon, as long as data from multiple sites was assimilated. Even so, intermediate complexity models performed the best both against assimilated and unassimilated data (Xiao and Friedrichs, 2014a). In their analysis, both the simplest and most complex models typically obtained optimized parameters that were good approximations to the observations at optimized locations, were unrealistic parameters and non-portable to other locations. Similarly, M2 which had 7 compartments and temperature-dependent biological rates, showed the highest number of locations that could be replicated well with parameters optimized for a different location (Fig. 10). This suggests that temperature dependency in the biological rates plays a role in model portability. Hence improving mechanistic interactions, rather than introducing unconstrained diversity, should be preferred to improve the global applicability of an ecosystem model. Temperature dependent grazing and mortality rates have been previously noted to improve the performance of global models (Behrenfeld et al., 2013), and this clearly also applies to dynamically complicated regions like the northwest North Atlantic.

Finally, it is important to note that the performance of M3 could significantly degrade depending on the choice of parameters to be optimized, as biological parameter values have the ability to effectively modify the intended trophic interactions in a model (Cropp and Norbury, 2009; Sailley et al., 2013). In our optimization experiments, these unintended behaviors included the functional extinctions (i.e., biomass too low to affect model results) and total extinctions of plankton groups. Similar extinctions occur when a local minimum solution is found (e.g., Schartau et al., 2001), or when no scaling weights are assigned to different locations in the cost function, as the optimization becomes biased towards locations with higher biomass, and generates extinct functional groups at the locations with lower biomass (e.g., Schartau and Oschlies, 2003a). To correct the extinctions, Schartau et al. (2001) introduced zooplankton observations to the cost function without success in finding an optimal parameter solution that replicated the observations well. A different case was presented by Kriest (2017), where using a wide range for zooplankton parameter values resulted in a solution where zooplankton was almost extinct. This undesired behavior was corrected by restricting the range of zooplankton parameters, also resulting in a better fit to nutrient and oxygen and more realistic concentrations and fluxes overall (Kriest, 2017). In the absence of parameter boundaries, Ward et al. (2010) obtained optimized parameters with unrealistic negative grazing rates, indicative of extinct zooplankton. Due to the non-linear nature of ecosystem models, the extinction of one component can have unforeseen, however logical, consequences. For example, Cropp and Norbury (2009) showed that removing the predator of a given prey does not only allow such prey population to increase but can also lead to the extinction of competing prey and their predators, and ultimately generate the simulated system's collapse.

6.3. Limitations and uncertainties

The main uncertainty in the use of mechanistic surrogate-based calibrations with 1D models is in neglecting horizontal advection fluxes. We neglected horizontal advection, as is typically done in 1D models, assuming that horizontal divergence terms are small relative to the biological sources and sinks. This allowed us to have an estimate of how much a 3D application of an oceanographically complex region, the northwest North Atlantic, can be improved with the use of reduced-order models. Several 1D models have been previously used to study locations within or close to our study area, being successful at replicating key aspects of biogeochemical variability (Tian et al., 2003, 2004; Ji et al., 2006; Song et al., 2010, 2011; Xiao and Friedrichs, 2014a,b). In particular, based on a cross-validation analysis of the results of optimized parameters for 1D models of the Mid-Atlantic Bights, Xiao and Friedrichs (2014a) were optimistic about the potential use of these parameters in a 3D application for the US eastern continental shelf. Nonetheless, neglecting horizontal advection may impact the surrogate performance. Hemmings et al. (2015) explicitly examined the effect of introducing the horizontal advective flux in a mechanistic emulator composed of 1D models representing the ocean conditions of twelve sites located every 5-degrees latitude along 20°W in the North Atlantic. Their results showed that the addition of horizontal fluxes improved the correlation coefficient between 1D and 3D surface chlorophyll. The addition of horizontal advective fluxes in the surrogates is only

recommended if the velocities of the target 3D model are accurate; otherwise, the optimized biological parameters may tend to compensate biases introduced by an erroneous physical forcing. In our study, some effects of neglecting advective fluxes are compensated by nudging deep nitrate in the surrogates. In other approaches, like in Hemmings et al. (2015), the uncertainties of the emulator are evaluated and then used for the cost function.

In addition to the uncertainties due to unresolved advection, there are two main issues with the use of optimized simulations for comparing ecosystem model with different complexities: (1) the cost function, and (2) the parameters to be optimized. The cost function is not an entirely objective measure. Its design can affect the outcome of the optimizations, as we discussed in Section 6.2. Similarly, Evans (2003) exemplified that weighting and variable scaling factors applied in the cost function can generate parameters sufficiently different to affect the estimates of biogeochemical fluxes. The design of the cost function can also be used to partially compensate the absence of horizontal transport through the addition of correction terms to biological variables (Losa et al., 2004; Hemmings and Challenor, 2012; Prießet al., 2013a). Correction terms can also account for other systematic or random errors in the surrogate; however, as suggested in Section 4.1, the more the 1D model is forced to behave like the 3D model, the less useful it becomes in identifying the sources of deficiencies in either the physical or the biological components of the model. A consideration in the use of correction terms is that the distribution of errors in the 1D and 3D models may vary during the optimization (Hemmings et al., 2015). If correction terms are used, a statistical error term may be more robust than a parameter-dependent error term (Hemmings et al., 2015).

Another issue of importance in the design of the optimization cost function is the selection of weights to balance the contributions of different variables and/or locations. In optimizations with multiple observational data types, optimal parameters become a compromise between different biogeochemical conditions and sources of data. Hence, the optimization results are quite sensitive to the scaling approach. The lack of any explicit treatment of biases, and the weighing scheme used in our cost function are consistent with previous studies (e.g., Friedrichs et al., 2007; Ward et al., 2010). However, we emphasize the importance of correcting biases between observational data sets of the same simulated variable, as we did in the case of surface chlorophyll.

The use of fractionated chlorophyll to compare against the small and large phytoplankton groups in the design of the cost function influenced the results of our portability experiments. This approach was also used by Xiao and Friedrichs (2014a) for optimization experiments in the Mid-Atlantic Bight. We can expect that advancements in our understanding of how complex ecosystem models behave can be made with the use of other empirical sources of information for the optimization of unconstrained variables. For example, zooplankton abundances from Continuous Plankton Recorder measurements cannot be directly compared to model results, but could provide estimates of seasonal variability (Lewis et al., 2006) that can be scaled to the corresponding simulated zooplankton groups. This might be particularly useful, since phytoplankton losses are among the least constrained parameters, even for simple NPZD models (e.g., Fennel et al., 2001; Bagniewski et al., 2011).

The selection of parameters to optimize is, at some level, subjective as well, and can have a dramatic effect as we have shown. Here, we supported our decision of the target parameters with a sensitivity analysis where the model response to variations in the parameter values withing their corresponding range was tested systematically. Parameters were ranked according to how much they affected chlorophyll and nitrate (i.e., the same variables available in the observations). Poorly constrained parameters can be set to arbitrary values during the optimization without significantly affecting the model cost (Ward et al., 2010) or otherwise tend to hit their a priori distribution limits (e.g. Schartau and Oschlies, 2003a). When unconstrained parameters are fixed to their a priori estimates, the level of previous tuning to the original model domain can skew the results of geographical portability experiments (Ward et al., 2010). In our results, models benefit from optimizing a higher number of sensitive parameters (E4 vs. E5; Fig. 12). If the number of optimized parameters is further increased and the set contains unconstrained parameters, the ability of such optimized complex models to simulate unassimilated observations becomes reduced (Friedrichs et al., 2006, 2007). Our results thus support that the selection of parameters should be done with consideration of the optimal number of parameters that can be constrained by the observational data.

7. Conclusions

Parameter optimization methods offer a systematic approach to reduce subjective model tuning and quantitatively compare ecosystem models with different complexities; however, optimization is not an entirely objective methodology with a unique solution. We have illustrated that, in addition to the uncertainties of the physical environment, conclusions about the accuracy and portability of a model can differ depending on decisions about the design of the cost function, the selection of parameters to be optimized, and the level of preliminary calibration of each model. Due to the limitations of applying parameter optimization in 3D coupled physical–biogeochemical models, 1D surrogates represent an efficient alternative for the exploration of the parameter space and for geographical portability experiments. In an extensive application of this concept, we configured ensembles of 1D models to behave as their regional 3D model application counterparts and used them to compare the performance of three ecosystem model versions. Processes unresolved by the 1D physical models and the level of ecosystem model complexity did affect the accuracy of the surrogates; however, successful surrogate-based model calibrations were possible and generated similar model-data misfits when applied in the 1D and 3D environments.

When an appropriate set of parameters was optimized, the model with multiple phytoplankton and zooplankton groups was better able to replicate assimilated observations than the single phytoplankton and zooplankton models. Nonetheless, the simpler models were also able to replicate the observed averaged seasonal variations in surface chlorophyll well. These results are consistent with previous studies and suggest that more complex trophic structures in models can better capture the observed temporal variability and spatial distribution of biogeochemical variables at multiple locations. In an additional analysis, geographical portability experiments provided an indication of how each model structure behaves with respect to unassimilated information. In this case, the most complex model was found to be the least portable, as the parameters optimized at some locations tended to favor either small or large phytoplankton. This result is consistent with other studies and with early theoretical notions about the expected behavior of complex models. We note that conclusions drawn from portability experiments comparing optimized models with different complexities are strongly affected by the prior degree of calibration of models, the number of parameters optimized and the parameter boundaries in the optimization. Moreover, when we varied the selection of optimization parameters in the complex model, it was prone to unsatisfactory results and unintended model behaviors. Attempting to optimize an improper selection of parameters resulted in the extinction of certain plankton groups, thus modifying the intended structure of trophic relationships in the model. Hence, we highlight that a guided selection of the parameters to be optimized is necessary, especially when – as in our case – little or no prior model tuning has been performed.

We also highlight that in order to benefit from the improved ecosystem representation that a complex model provides, such model needs to be trained with observations from diverse geographical locations. Research is required on efficient sampling methodologies to calibrate global surrogates, allowing us to determine the number of locations that

would be sufficient, and ensuring that the most representative locations are being selected.

Finally, we also observed an improvement in our simplest model version when all biological fluxes were configured to depend on temperature. Therefore, we can conclude that improving the mechanistic relationships, rather than adding unconstrained diversity, can lead to more robust globally applicable models. Here we base this statement on the results of the model including temperature dependency, but the same argument may apply to the use of allometric or otherwise scaled and parameterized models, as well as to the combined use of temperature and parameterized dependencies. A subsequent study will analyze how these optimized model versions perform when applied to the 3D environment: Does complexity affect our conclusions about the drivers underlying phenology? How does complexity affect estimates of primary production? The answers to these questions are key when making decisions about which level of complexity should be used for our study region.

Acknowledgments

We thank C. Johnson and A. Cogswell at the Bedford Institute of Oceanography for providing observational data and guidance for the use of the AZMP database, J. Urrego Blanco for the physical boundary conditions used in the model, L. Bianucci, C. Brennan and R. Zhang for their contributions to the model development, as well as M. Dowd and M. Schartau for insightful discussions. This manuscript was greatly improved thanks to the constructive criticism and suggestions from three anonymous reviewers and our editor, Marjy Friedrichs. We gratefully acknowledge the financial support from Natural Sciences and Engineering Research Council of Canada (NSERC) and the Marine Environmental Observation Prediction and Response Network (MEOPAR).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ocemod.2019.101437.

References

Arhonditsis, G.B., Brett, M.T., 2004. Evaluation of the current state of mechanistic aquatic biogeochemical modelling. Mar. Ecol. Prog. Ser. 271, 13–26.

Bagniewski, W., Fennel, K., Perry, M.J., D'Asaro, E., 2011. Optimizing models of the North Atlantic spring bloom using physical, chemical and bio-optical observations from a Lagrangian float. Biogeosciences 8, 1291–1307. http://dx.doi.org/10.5194/bg-8-1291-2011.

Barton, A.D., Dutkiewicz, S., Flierl, G., Bragg, J., Follows, M.J., 2010. Patterns of diversity in marine phytoplankton. Science 327, 1509–1511. http://dx.doi.org/10.1126/science.1184961.

Behrenfeld, M.J., Doney, S.C., Lima, I., Boss, E., Siegel, D.A., 2013. Annual cyles of ecological disturbance and recovery underlying the subartic Atlantic spring plankton bloom. Glob. Biogeochem. Cycles 27, 526–540. http://dx.doi.org/10.1002/gbc.20050.

Bianucci, L., Fennel, K., Chabot, D., Shackell, N., Lavoie, N., 2015. Ocean biogechemical models as management tools: a case study for Atlantic wolfish and declining oxygen. ICES J. Mar. Sci. 73, 263–274. http://dx.doi.org/10.1093/icesjms/fsv220.

Brennan, C., Bianucci, L., Fennel, K., 2016. Sensitivity of Northwest North Atlantic shelf circulation to surface and boundary forcing: a regional model assessment. Atmos.-Ocean http://dx.doi.org/10.1080/07055900.2016.1147416.

Cropp, R., Norbury, J., 2009. Parameterizing plankton functional type models: insights from a dynamical systems perpective. J. Plankton Res. 31, 939–963.

Dadou, I., Evans, G., Garçon, V., 2004. Using JGOFS in situ and ocean color data to compare biogeochemical models and estimate their parameters in the subtropical North Atlantic Ocean. J. Mar. Res. 56, 5–594.

D'Alelio, D., Libralato, S., Wyatt, T., Ribera d'Alcalà, M., 2016. Ecological-network models link diversity, structure and function in the plankton food-web. Sci. Rep. 6, http://dx.doi.org/10.1038/srep21806.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., 2011. The ERA-interum reanalysis: Configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137, 553–597.

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

Doney, S., Glover, D., Najjar, R., 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 4, 591–624.

Dutkiewicz, S., Hickman, A.E., Jahn, O., Gregg, W.W., Mouw, C.B., Follows, M.J., 2015. Capturing optically important constituents and properties in a marine biogeochemical and ecosystem model. Biogeosciences 12, 4447–4481. http://dx.doi.org/10.5194/bg-12-4447-2015.

Eppley, R.W., 1972. Temperature and phytoplankton growth in the sea. Fish. Bull. 70, 1063–1085.

Evans, G.T., 2003. Defining misfit between biogeochemical models and data sets. J. Mar. Syst. 40–41, 49–54. http://dx.doi.org/10.1016/S0924-7963(03)00012-5, The Use of Data Assimilation in Coupled Hydrodynamic, Ecological and Bio-geochemical Models of the Ocean. Selected papers from the 33rd International Liege Colloquium on Ocean Dynamics, held in Liege, Belgium on May 7–11th, 2001.

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

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, 591–639.

Fasham, M.J.R., Sarmiento, J.L., Slater, R.D., Ducklow, H., Williams, R., 1993. Ecosystem behavior at Bermuda Station "S" and ocean weather station "India": A general circulation model and observational analysis. Glob. Biogeochem. Cycles 7, 379–415. http://dx.doi.org/10.1029/92GB02784.

Fennel, K., Losch, M., Schröter, J., Wenzel, M., 2001. Testing a marine ecosystem model: sensitivity analysis and parameter optimization. J. Mar. Syst. 28, 45–63.

Fennel, K., Wilkin, J., Levin, J., Moisan, J., O'Reilly, J.E., Haidvogel, D., 2006. Nitrogen cycling in the Middle Atlantic Bight: Results from a three-dimensional model and implications for the North Atlantic nitrogen budget. Glob. Biogeochem. Cycles 20 (14), http://dx.doi.org/10.1029/2005GB002456.

Fennel, K., Wilkin, J., Previdi, M., Najjar, R., 2008. Denitrification effects on air-sea CO2 flux in the coastal ocean: Simulations for the northwest North Atlantic. Geophys. Res. Lett. 35, http://dx.doi.org/10.1029/2008GL036147.

Friedrichs, M.A.M., Dusenberry, J.A., Anderson, L.A., Armstrong, R.A., Chai, F., Christian, J.R., Doney, S., Dunne, J.P., Fujii, M., Hood, R., McGillicuddy, D.J., Moore, K., Schartau, M., Spitz, Y.H., Wiggert, J., 2007. Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups. J. Geophys. Res. 112, 1–22.

Friedrichs, M.A.M., Hood, R., Wiggert, J., 2006. Ecosystem model complexity versus physical forcing: Quantification of their relative impact with assimilated Arabian Sea data. Deep-Sea Res. II 53, 576–600.

Fulton, E., Smith, A., Johnson, C., 2003. Effect of complexity on marine ecosystem models. Mar. Ecol. Prog. Ser. 253, 1–16. http://dx.doi.org/10.3354/meps253001.

Galbraith, E.D., Dunne, J.P., Gnanadesikan, A., Slater, R.D., Sarmiento, J.L., Dufour, C.O., de Souza, G.F., Bianchi, D., Claret, M., Rodgers, K.B., Marvasti, S.S., 2015. Complex functionality with minimal computation: Promise and pitfalls of reduced-tracer ocean biogeochemistry models. J. Adv. Model. Earth Syst. 7, 2012–2028. http://dx.doi.org/10.1002/2015MS000463.

Garcia, H.E., Locarnini, R.A., Boyer, T.P., Antonov, J.I., Zweng, M.M., Baranova, O.K., Johnson, D.R., 2010. Nutrients (phosphate, nitrate, and silicate). In: World Ocean Atlas 2009, NOAA Atlas NESDIS 71. U.S. Government Printing Office, Washington, D.C, p. 398.

Geshelin, Y., Sheng, J., Greatbatch, R., 1999. Monthly Mean Climatologies of Temperature and Salinity in the Western North Atlantic. Can Data Rep Hydrogr Ocean SciRapp Stat Can Hydrogr Sci Ocean.

Haidvogel, D., Arango, H.G., Budgell, W.P., Cornuelle, B.D., Curchistser, E., Di Lorenzo, E., Fennel, K., Geyer, W.R., Hermann, A.J., Lanerollee, L., Shchepetkin, A.F., Sherwood, C.R., Signell, R.P., Warner, J.C., Wilkin, J., 2008. Ocean forecasting in terrain-following coordinates: Formulation and skill assessment of the Regional Ocean Modeling System. J. Comput. Phys. 227, 3595–3624.

Hemmings, J.C.P., Challenor, P.G., 2012. Addressing the impact of environmental uncertainty in plankton model calibration with a dedicated sorftware system: the Marine Model Optimization Testbed (MarMOT 1.1 alpha). Geosci. Model Dev. 47, 1–498.

Hemmings, J.C.P., Challenor, P.G., Yool, A., 2015. Mechanistic site-bassed emulation of a global ocean biogeochemical model (MEDUSA 1.0) for parametric analysis and calibration: an application of the Marine Model Optimization Testbed (MarMOT 1.1). Geosci. Model Dev. 69, 7–731.

Hirata, T., Hardman-Mountford, N.J., Brewin, R.J., Aiken, J.W., Barlow, R., Suzuki, K., Isada, T., Howell, E., Hashioka, T., Noguchi-Aita, M., Yamanaka, Y., 2011. Synoptic relationships between surface chlorophyll-a and diagnostic pigments to phytoplankton functional types. Biogeosciences 8, 311–327.

Hooten, M., Leeds, W.B., Fiechter, J., Wikle, C.K., 2011. Assessing first-order emulator interference for physical parameters in nonlinear mechanistic models. J. Agric. Biol. Environ. Stat. 16, 475–494.

Houck, C.R., Joines, J.A., Kay, M.G., 1995. A Genetic Algorithm for Function Optimization: A Matlab Implementation (Technical Report No. NCSU-IE-TR-95-09). North Carolina State University, Raleigh, NC.

Ji, R., Chen, C., Franks, P.J.S., Townsend, D.W., Durbin, E.G., Beardsley, R.C., Gregory Lough, R., Houghton, R.W., 2006. Spring phytoplankton bloom and associated lower trophic level food web dynamics on Georges Bank: 1-D and 2-D model studies. Deep Sea Res. Part II 53, 2656–2683. http://dx.doi.org/10.1016/j.dsr2.2006.08.008, Dynamics of Plankton and Larval Fish Populations on Georges Bank, the North Atlantic U.S. GLOBEC Study Site.

Kane, A., Moulin, C., Thiria, S., Bopp, L., Berrada, M., Tagliablue, A., Crépon, M., Aumont, O., Badran, F., 2011. Improving the parameters of a global ocean biogeochemical model via variational assimilation of in situ data at five time series stations. J. Geophys. Res. Ocean. 116, http://dx.doi.org/10.1029/2009JC006005.

Kishi, M., Kashiwai, M., Ware, D.M., Megrey, B.A., Eslinger, D.L., Werner, F.E., Noguchi-Aita, M., Azumay, T., Fujii, M., Hashimoto, S., Huang, D., Iizumi, H., Ishida, Y., Kang, S., Kantakov, G.A., Kim, H., Komatsu, K., Navrotsky, V.V., Smith, S.L., Tadokoro, K., Tsuda, A., Yamamura, O., Yamanaka, Y., Yokouchi, K., Yoshie, N., Zhang, J., Zuenko, Y.I., Zvalinsky, V., 2007. NEMURO - a lower trophic level model for the North Pacific marine ecosystem. Ecol. Model. 202, 12–25.

Kriest, I., 2017. Calibration of a simple and a complex model of global marine biogeochemistry. Biogeosciences 14, 4965–4984. http://dx.doi.org/10.5194/bg-14-4965-2017.

Kriest, I., Khatiwala, S., Oschlies, A., 2010. Towards an assessment of simple global marine biogeochemical models of different complexity. Prog. Oceanogr. 86, 337–360.

Kriest, I., Sauerland, V., Khatiwala, S., Srivastav, A., Oschlies, A., 2017. Calibrating a global three-dimensional biogeochemical ocean model (MOPS-1.0). Geosci. Model Dev. 10, 127–154. http://dx.doi.org/10.5194/gmd-10-127-2017.

Kuhn, A.M., Fennel, K., Berman-Frank, I., 2018. Modelling the biogeochemical effects of heterotrophic and autotrophic N2 fixation in the Gulf of Aqaba (Israel). Red Sea. Biogeosciences 15, 7379–7401. http://dx.doi.org/10.5194/bg-15-7379-2018.

Kuhn, A.M., Fennel, K., Mattern, J.P., 2015. Model investigations of the North Atlantic spring bloom initiation. Prog. Oceanogr. 17, 6–193.

Kwiatkowski, L., Yool, A., Allen, J.I., Anderson, T.R., Barciela, R., Buitenhuis, E.T., Butenschön, M., Enright, C., Halloran, P.R., Le Quéré, C., de Mora, L., Racault, M.-F., Sinha, B., Totterdell, I.J., Cox, P.M., 2014. iMarNet: an ocean biogeochemistry model intercomparison project within a common physical ocean modelling framework. Biogeosciences 11, 7291–7304. http://dx.doi.org/10.5194/bg-11-7291-2014.

Kwon, E.Y., Primeau, F., 2006. Optimization and sensitivity study of a biogeochemistry ocean model using an implicit solver and in situ phosphate data. Global Biogeochem. Cycles 20 (4).

Kwon, E.Y., Primeau, F., 2008. Optimization and sensitivity of a global biogeochemistry ocean model using combined in situ DIC, alkalinity, and phosphate data. J. Geophys. Res.: Oceans 113 (C8).

Lagman, K., Fennel, K., Thompson, K., 2014. Assessing the utility of frequency dependent nudging for reducing biases in biogeochemical models. Ocean Model. 2, 5–35.

Le Quéré, C., Harrison, S.P., Prentice, I.C., Buitenhuis, E.T., Aumont, O., Bopp, L., Claustre, H., Cotrim da Cunha, L., Geider, R., Giraud, X., Klaas, C., Kohfeld, K.E., Legendre, L., Manizza, M., Platt, T., Rivkin, R.B., Sathyendranath, S., Uitz, J., Watson, A., Wolf-Gladrow, W., 2005. Ecosystem Dynamics Based on Plankton Functional Types for Global Ocean Biogeochemistry Models 11, 2016–2040. http://dx.doi.org/10.1111/j.1365-2486.2005.01004.x.

Leeds, W.B., Wikle, C.K., Fiechter, J., Brown, J., Milliff, R.F., 2012. Modeling 3-D spatio-temporal biogeochemical processes with a forest of 1-D statistical emulators. Environmetrics http://dx.doi.org/10.1002/env.2187.

Lewis, K., Allen, J.I., Richardson, A.J., Holt, J.T., 2006. Error quantification of a high resolution coupled hydrodynamic ecosystem coastal-ocean model: Part3, validation with Continuous Plankton Recorder data. J. Mar. Syst. 63, 209–224.

Loder, J.W., Petrie, B., Gawarkiewicz, G., 1998. The coastal ocean off northwestern North America: A large-scale view. In: The Sea. John Wiley and Sons, pp. 015–133.

Losa, S., Kivman, G.A., Ryanbchenko, 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. 1–20.

Matear, R.J., 1995. Parameter optimization and analysis of ecosystem models using simulated annealing: A case study at Station P. J. Mar. Res. 53, 571–607.

Mattern, J.P., Fennel, K., Dowd, M., 2012. Estimating Time-Dependent Parameters for a Biological Ocean Model using an Emulator Approach. pp. 32–47.

Mattern, J.P., Song, H., Edwards, C.A., Moore, A.M., Fiechter, J., 2017. Data assimilation of physical and chlorophyll a observations in the California Current System using two biogeochemical models. Ocean Model. 109, 55–71. http://dx.doi.org/10.1016/j.ocemod.2016.12.002.

McDonald, C.P., Bennington, V., Urban, N.R., McKinley, G.A., 2012. 1-D test-bed calibration of a 3-D Lake Superior biogeochemical model. Ecol. Model. 225, 115–126.

Mitchell, M.R., Harrison, G., Pauley, K., Gagné, A., Maillet, G., Strain, P., 2002. Atlantic Zonal Monitoring Program Sampling Protocol. Can. Tech. Rep. Hydrogr. Ocean. Sci..

Oschlies, A., Schartau, M., 2005. Basin-Scale Performance of a Locally Optimized Marine Ecosystem Model 63. pp. 335–358.

Prieß, M., Koziel, S., Slawig, T., 2013a. Marine ecosystem model calibration with real data using enhanced surrgoate-based optimization. J. Comput. Sci. 4, 423–437.

Prieß, M., Piwonski, J., Koziel, S., Oschlies, A., Slawig, T., 2013b. Accelerated parameter identification in a 3D marine biogeochemical model using surrogate-based optimization. Ocean Model. 68, 22–36.

Quine, W.V., 1975. On empirically equivalent systems of the world. Erkenntnis 9, 313–328.

Raick, C., Soetaert, K., Grégoire, M., 2006. Model complexity and performance: How far can we simplify? Prog. Oceanogr. 70, 27–57. http://dx.doi.org/10.1016/j.pocean.2006.03.001.

Riley, G.A., 1965. A mathematical model. Limnol. Oceanogr. 10, 202–215.

Rutherford, K., Fennel, K., 2018. Diagnosing transit times on the northwestern North Atlantic continental shelf. Ocean Sci. 14 (5), 1207–1221.

Sailley, S.F., Vogt, M., Doney, S., Aita, M.N., Bopp, L., Buitenhuis, E.T., Hashioka, T., Lima, I., Le Quéré, C., Yamanaka, Y., 2013. Comparing food web structures and dynamics across a suite of global marine ecosystem models. Ecol. Model. 261–262, 43–57.

Schartau, M., Oschlies, A., 2003a. Simultaneous data-based optimization of a 1D-ecosystem model at three locations in the North Atlantic: Part I - Method and parameter estimates. J. Mar. Res. 61, 765–793.

Schartau, M., Oschlies, A., 2003b. Simultaneous data-based optimization of a 1D-ecosystem model at three locations in the North Atlantic: Part II - Standing stocks and nitrogen fluxes. J. Mar. Res. 79, 5–821.

Schartau, M., Oschlies, A., Jürgen, W., 2001. Parameter estimates of a zero-dimensional ecosystem model applying the adjoint method. Deep-Sea Res. II 48, 1769–1800.

Schartau, M., Wallhead, P., Hemmings, J.C.P., Loptien, U., Kriest, I., Krishna, S., Ward, B.A., Slawig, T., Oschlies, A., 2017. Review and syntheses: parameter identification in marine planktonic ecosystem modelling. Biogeosciences 14, 1647–1701.

Smith, T.M., Reynolds, R.W., Peterson, T.C., Lawrimore, J., 2008. Improvements to NOAA's historical merged land–Ocean surface temperature analysis (1880–2006). J. Clim. 21, 2283–2296. http://dx.doi.org/10.1175/2007JCLI2100.1.

Song, H., Ji, R., Stock, C., Kearney, K., Wang, Z., 2011. Interannual variability in phytoplankton blooms and plankton productivity over the Nova Scotian Shelf and in the Gulf of Maine. Mar. Ecol. Prog. Ser. 426, 105–118. http://dx.doi.org/10.3354/meps09002.

Song, H., Ji, R., Stock, C., Wang, Z., 2010. Phenology of phytoplankton blooms in the Nova Scotian Shelf–Gulf of Maine region: remote sensing and modeling analysis. J. Plankton Res. 32, 1485–1499. http://dx.doi.org/10.1093/plankt/fbq086.

Tian, R.C., Deibel, D., Rivkin, R.B., Vézina, A.F., 2004. Biogenic carbon and nitrogen export in a deep-convection region: simulations in the Labrador Sea. Deep-Sea Res. I 51, 413–437. http://dx.doi.org/10.1016/j.dsr.2003.10.015.

Tian, R., Deibel, D., Thompson, R., Rivkin, R., 2003. Modeling of climate forcing on a cold-ocean ecosystem, Conception Bay, Newfoundland. Mar. Ecol. Prog. Ser. 262, 1–17. http://dx.doi.org/10.3354/meps262001.

Townsend, D.W., Thomas, A.C., Mayer, L.M., Thomas, M.A., Quinlan, J.A., 2004. Chapter 5: Oceanography of the Northwest Atlantic Continental Shelf (1, W). In: The Sea: The Global Coastal Ocean: Interdisciplinary Regional Studies and Syntheses. Harvard University Press.

Urrego-Blanco, J., Sheng, J., 2012. Interannual Variability of the Circulation Over the Eastern Canadian Shelf 50. pp. 277–300. http://dx.doi.org/10.1080/07055900.2012.680430.

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

Ward, B.A., Schartau, M., Oschlies, A., Martin, A., Follows, M., Anderson, T.R., 2013. When is a biogechemical model too complex? Objective model reduction and selection for North Atlantic time-series sites. Prog. Oceanogr. 4, 9–65.

Xiao, Y., Friedrichs, M.A.M., 2014a. The assimilation of satellite-derived data into a one-dimensional lower trophic level marine ecosystem model. J. Geophys. Res. Ocean. 119, 2691–2712. http://dx.doi.org/10.1002/2013JC009433.

Xiao, Y., Friedrichs, M.A.M., 2014b. Using biogeochemical data assimilation to assess the relative skill of multiple ecosystem models in the Mid-Atlantic Bight: effects of increasing the complexity of the planktonic food web. Biogeosciences 11, 3015–3030. http://dx.doi.org/10.5194/bg-11-3015-2014.

Yoshie, N., Yamanaka, Y., Rose, K.A., Eslinger, D.L., Ware, D.M., Kishi, M., 2007. Parameter sensitivity study of the NEMURO lower trophic level marine ecosystem model. Ecol. Model. 202, 26–37.