Journal of Marine Systems 165 (2017) 139–152

Contents lists available at ScienceDirect

Journal of Marine Systems

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

Simple parameter estimation for complex models — Testing evolutionary techniques on 3-dimensional biogeochemical ocean models

Jann Paul Mattern*, Christopher A. Edwards

Department of Ocean Sciences, University of California, Santa Cruz, CA, USA

ARTICLE INFO

Article history:
Received 11 February 2016
Received in revised form 26 October 2016
Accepted 27 October 2016
Available online 30 October 2016

Keywords:
Parameter estimation
Biogeochemical ocean model
Evolutionary algorithm

ABSTRACT

Parameter estimation is an important part of numerical modeling and often required when a coupled physical–biogeochemical ocean model is first deployed. However, 3-dimensional ocean model simulations are computationally expensive and models typically contain upwards of 10 parameters suitable for estimation. Hence, manual parameter tuning can be lengthy and cumbersome. Here, we present four easy to implement and flexible parameter estimation techniques and apply them to two 3-dimensional biogeochemical models of different complexities. Based on a Monte Carlo experiment, we first develop a cost function measuring the model-observation misfit based on multiple data types. The parameter estimation techniques are then applied and yield a substantial cost reduction over ~ 100 simulations. Based on the outcome of multiple replicate experiments, they perform on average better than random, uninformed parameter search but performance declines when more than 40 parameters are estimated together. Our results emphasize the complex cost function structure for biogeochemical parameters and highlight dependencies between different parameters as well as different cost function formulations.

1. Introduction

Parameter estimation is an important aspect of numerical modeling in general and biogeochemical (BGC) ocean modeling in particular. The process of estimating parameter values based on observations is especially relevant in the context of BGC models because they feature uncertain parameters that depend on, or represent biological properties that, unlike fundamental physical constants, may change in time and space (Mattern et al., 2012). BGC parameters are required to represent the integrated response of the entire plankton community in a given region and within a specific time frame. This makes these parameters less portable from one model configuration to another and typically requires parameter estimation using local observations for a new model setup.

Many BGC model formulations exist; in terms of complexity they differ mainly in the number of BGC variables that they simulate and the pathways that connect them. Commonly used BGC models of low complexity have as few as three or four variables, while commonly used high complexity models may have more than 20 (Friedrichs et al., 2007). The number of uncertain parameters typically increases with the number of variables (Denman, 2003), rendering parameter estimation for complex models a more challenging and computationally intensive task. In addition, more complex models require more observations to constrain their parameters. As oceanic observations of BGC properties and processes are sparse, previous studies have shown that relatively few BGC parameters can effectively be constrained with widely available observations, even for less complex BGC models (Matear, 1995; Ward et al., 2010; Fiechter et al., 2013).

There is a wide variety of parameter estimation techniques that differ in complexity both in terms of cost for implementation and computational expense. The former refers to the effort required to implement the technique which may include model-specific changes or additions to the model code, while the latter refers to computer time that is needed to run the technique which can typically be expressed in the number of required model evaluations.

Variational techniques (see e.g. Lawson et al., 1996; Ward et al., 2010; Bagniewski et al., 2011) often employ gradient descent methods to search for cost function minima. This approach is a computationally efficient way to find minima, yet it typically requires a linearized or adjoint version of the model's code to obtain the cost function gradient or an approximation of it. These model-specific requirements increase the effort to implement the techniques, especially for complex models. In addition, variational techniques can easily get trapped in local minima (Ward et al., 2010), increasing the computational cost by requiring multiple estimation runs with

* Corresponding author.
E-mail address: jmattern@ucsc.edu (J. Mattern).

http://dx.doi.org/10.1016/j.jmarsys.2016.10.012
0924-7963/

different starting conditions to find global minima (Schartau et al., 2001).

Parameter estimation techniques that do not rely on gradient information directly are often referred to as heuristic techniques. They are typically easier to implement but there is no guaranteed convergence to a minimum even when operating in close proximity to it. Some techniques, such as the Nelder–Mead method (Doerffer and Fischer, 1994) approximate the cost function gradient locally based on systematic sampling of cost function values. Many other techniques do not attempt to approximate gradient information at all, they contain stochastic elements to find lower cost function values, which are also designed to decrease the likelihood of getting trapped in local minima. Heuristic Parameter estimation often borrows ideas from real world concepts, for example simulated annealing (e.g. Matear, 1995; Hurtt and Armstrong, 1996) that is conceptually based on a physical cooling process converging towards a global energy minimum.

Here, we focus on evolutionary algorithms and similar heuristic techniques based on biological concepts, mainly that of natural evolution (the terms evolutionary algorithm and genetic algorithm are often used interchangeably, although genetic algorithms are technically a subclass of evolutionary algorithms (Bäck and Schwefel, 1993)). In the following we will summarily refer to this class of parameter estimation techniques as *evolutionary techniques*.

Evolutionary techniques have been applied to low dimensional BGC models in earlier studies: Schartau and Oschlies (2003) use a micro-genetic algorithm and five types of BGC observations to estimate the values of 13 parameters across three 1-dimensional NPZD models at different locations, and Kettle (2009) uses a very similar setup to estimate 11 parameters using backscatter and chlorophyll-*a* observations. Ward et al. (2010) compare the use of a micro-genetic algorithm with that of a gradient descent-based variational adjoint technique and concludes that both perform similarly in terms of lowering the model-data misfit when estimating up to 10 parameters; the adjoint technique yields more precise estimates, yet is more prone to get trapped in local minima in higher dimensional cases. In Rückelt et al. (2010) a hybrid approach that combines the benefits of gradient descent with that of evolutionary techniques is employed to estimate 12 parameters. More recently, Kuhn et al. (2015) applied an evolutionary algorithm to estimate 13 BGC parameters at 6 locations in the subpolar North Atlantic.

All of the above studies use 1-dimensional models, that are configured to represent a vertical water column and rely on thousands of model evaluations to perform one estimation experiment — and many experiments with different starting values to ensure convergence to a global minimum or to obtain estimates of parameter uncertainty. Here, we apply several evolutionary techniques to estimate 9 and 43 parameters of two different 3-dimensional BGC models using data of chlorophyll-*a*, nitrate (NO_3) and, in the case of the more complex model, measured silicate concentrations. The computational cost of 3-dimensional models allows us to expend only a few experiments with ~100 model evaluations each. Hence, our goal is to test whether evolutionary techniques are usable and yield good results in terms of lowering the cost function, even for a low number of model evaluations and in high dimensional parameter spaces. While it is unlikely that many of the models' parameters can be fully constrained in this manner, we show that the choice of cost function and the types of observations included in the cost function help to constrain the estimated parameter choices and reduce problems of overfitting.

Prior to performing parameter estimation, we use Monte Carlo experiments in which parameter values are repeatedly drawn randomly in order to test several cost function formulations based on different types of observations and assess their ability to constrain parameter values and examine correlations between them. This helps us select a combined cost function for the parameter estimation that includes two surface chlorophyll-*a*-based contributions and a NO_3 contribution. The Monte Carlo experiments also help to evaluate the results of our parameter estimation experiments, by allowing us to compare the convergence rate of the parameter estimation techniques against results that can be expected from unguided, random sampling.

2. Methods

2.1. The NPZD and NEMURO models

The two BGC models we use in this study are embedded in a 3-dimensional physical model based on version 3.7 of the Regional Ocean Modelling System (ROMS; Haidvogel et al., 2008). The model domain covers the eastern Pacific coastline from 30°N to 48°N, extending westward from the coast to 134°W at a 0.1° horizontal resolution (Fig. 4a). Vertically, the model is split into 42 terrain-following layers. Forcing for the physical model (wind, solar radiation, air temperature, pressure and humidity data) as well as boundary conditions are provided by the Coupled Ocean Atmosphere Mesoscale Prediction System (COAMPS; Doyle et al., 2009). The setup of the physical model is nearly identical to that in Veneziani et al. (2009) where it is described in more detail. This physical model setup has also been combined with various BGC models (Goebel et al., 2010; Raghukumar et al., 2015; Song et al., 2016).

Our BGC model of low complexity resolves concentrations of NO_3, phytoplankton and zooplankton biomass as well as detritus (NPZD). It is based on the ROMS NPZD model of Fiechter et al. (2009) but with iron limitation excluded. This model uses nitrogen as its currency, measuring all variables in units of nitrogen concentration (mmol N m^{-3}). For the NPZD model, we selected 9 BGC parameters to be adjusted. This selection includes all the BGC parameters of the model, with the exception of parameters related to well known physical properties such as the fraction of photosynthetically active shortwave radiation or the seawater light attenuation. Note that our model configuration does not include direct remineralization of phytoplankton or zooplankton to NO_3, and zooplankton excretion to detritus. All parameters and their reference values are listed in Table A1. The reference values are based on literature sources and contain adaptations for the California Current System (Powell et al., 2006; Fiechter et al., 2009).

Our second model is a ROMS implementation of NEMURO (Kishi et al., 2007) that contains a total of 11 BGC variables, consisting of 3 nutrients (nitrate, ammonium, silicate), 2 phytoplankton groups (larger diatoms and smaller nanophytoplankton), 3 zooplankton groups (microzooplankton, mesozooplankton, predatory zooplankton), and 3 classes of detritus (particulate and dissolved organic nitrogen, opal). For NEMURO, we took the same approach as for NPZD and selected all relevant BGC parameters in our configuration as uncertain (our NEMURO configuration does not include inhibition of grazing by predatory zooplankton in the presence of other zooplankton). Our selection includes 48 parameters (listed in Table A2). Out of the 48 parameters, we allowed 43 to vary independently, while we identified the remaining 5 NEMURO parameters to be logically dependent on the values of other parameters. For example, some dependencies are linked to parameters serving multiple functions in the NEMURO model formulation (e.g. a phytoplankton growth parameter that also affects light attenuation), which are described in more detail in Appendix A. In all of our experiments, the values of these 5 parameters were determined based on the values of the other 43 parameters. The reference parameter values for NEMURO (Table A2) are based on those in Rose et al. (2015) and contain domain-specific modifications.

The 5 dependent parameters identified above leave 43 independent parameters that determine the dimension of the NEMURO parameter space in our experiments. It should be noted that there is