Estimating biological parameters of a coupled physical–biological model of the Indian Ocean using polynomial chaos

Rashmi Sharma[1,*], Smitha Ratheesh[1] and Sujit Basu[1,2]

[1]Oceanic Sciences Division, Atmospheric and Oceanic Sciences Group, Space Applications Centre, Ahmedabad 380 015, India
[2]Present address: Department of Mathematics, Indus University, Ahmedabad 382 115, India

A statistical emulator technique, namely polynomial chaos, has been used to estimate two time-dependent biological parameters of a coupled physical–biological model of the Indian Ocean. This has been achieved by minimizing a distance function representing misfit between model simulated and satellite-derived surface chlorophyll. First, the parameters have been assumed to be constant in time and optimized values have been found by minimizing a time-averaged distance function. Since no significant improvement in model simulation has been found using a fixed set of optimum parameters, minimization has been carried out daily, assuming the parameters to be time-dependent. Emulation with this set of parameters has led to a significant improvement in the simulated surface chlorophyll. Smoothing of the parameters with singular spectrum analysis has caused less noisy simulations, at the cost of increasing the model data misfit. Time-varying parameters have been found to be more suitable for the hindcast of daily averaged chlorophyll both in the Arabian Sea and the Bay of Bengal.

Keywords: Coupled physical–biological model, distance function, polynomial chaos, surface chlorophyll.

NUMEROUS studies are devoted to data assimilation in biological or coupled physical–biological ocean models. In data assimilation, the model simulations are combined with observations in an optimum manner so that the model hindcasts are improved. There are two broad classes of data assimilation techniques, namely variational techniques[1,2], and Monte-Carlo-based techniques like ensemble Kalman filter[3,4] and particle filter[5]. In these types of data assimilation, the model states are altered, leaving the model parameters untouched. One can also use observations to optimize poorly known model parameters[6–9], so that the model simulations are improved. Hence this can be also considered as another type of data assimilation technique. The techniques employed for state estimation can also be used for parameter estimation. Thus, variational technique was used for parameter estimation in an equatorial Pacific Ocean model[6], and Monte-Carlo techniques were applied to ecosystem models[7,8] as well as a circulation model[9].

The aim of the present study is optimal estimation of two biological parameters of a coupled physical–biological model of the Indian Ocean, so that the model hindcast of the chlorophyll concentration is improved significantly. We have adopted the emulator approach to carry out this optimal estimation. Emulators require a set of model runs with different specific values assigned to the parameters, the optimal values of which are to be estimated[10]. Later, these simulations are approximated in different ways. Once this is done, one can easily approximate the true model output for any arbitrary values assigned to the concerned parameters. This makes the emulator approach faster and more cost-effective than other approaches.

As far as emulator technique is concerned, again there are a variety of approaches. We mention some of them in the oceanographic context, e.g. emulators based on artificial neural networks[11], Gaussian process model[12] and polynomial chaos[10]. We adopt the last approach in the present study. The polynomial chaos expansion was introduced by Weiner[13] and extended later[14,15]. In polynomial chaos, a set of orthogonal polynomials are used as basis functions for approximating the model results. It has been widely used in physical sciences[16], with only a few applications in oceanography[10,17,18]. Before proceeding to the detailed description of the technique involved, we describe briefly the model and the data used in the study.

The model is a coupled physical–biological model of the Indian Ocean[19] and has been validated in the Arabian Sea basin[20]. It has been also used to study the phytoplankton bloom in the Bay of Bengal[21]. Its physical component is a variable-density, 4.5 layer model, while its biological component consists of a set of advective–diffusive equations in each layer that determine nitrogen concentration in four compartments, namely nutrients, phytoplankton, zooplankton and detritus. The physical variables are the layer thickness, horizontal velocity, temperature and salinity. The deep ocean below the active layers is quiescent, where pressure gradients vanish (the 0.5 layer).

Since in this study we use satellite chlorophyll observations for the estimation of two biological parameters of the model, we mention just the time-evolution equation for the phytoplankton concentration P (μmol N kg^{-1}) in the topmost layer, since satellite-chlorophyll is related to this variable

$$\frac{\partial P}{\partial t} + V \cdot \nabla P - \kappa \nabla^2 P + \kappa_4 \nabla^4 P = S_{\mathrm{P}} + W_{\mathrm{P}}. \quad (1)$$

The terms on the left-hand side, starting from the leftmost term, denote local rate of change, advection (V being the horizontal velocity vector, in cm s^{-1}), Laplacian and bihar-

*For correspondence. (e-mail: rashmi@sac.isro.gov.in)

monic mixing, $\kappa = 10^7$ cm^2 s^{-1} and $\kappa_4 = 2 \times 10^{21}$ cm^4 s^{-1}, being the corresponding mixing coefficients. On the right-hand side of the equation, S_P contains the source and sink terms for the phytoplankton, and W_P represents almost all the vertical fluxes of phytoplankton. We omit details for the sake of brevity. Suffice it to mention that S_P contains the two biological parameters, which are to be estimated. These are maximum phytoplankton growth rate (g) and maximum zooplankton grazing rate (g_r; in sec^{-1}).

For carrying out the intended experiments, the physical model is spun up from a state of rest for a period of 5 years using climatological forcing fields. Then, the coupled model is spun up for another 5 years using the same forcing. From this initial condition, the coupled model is integrated from 1 January 2003 to 1 January 2006 using NCEP reanalysis fields for starting the emulation experiments with daily images of surface chlorophyll concentrations derived by SeaWIFS satellite for the year 2006 (ftp://podaac-ftp.jpl.nasa.gov/allData/seawifs/L3/chlA/9km/daily). For this work, we have to find the difference between a model-simulated chlorophyll field and an observed image. We call this difference a distance denoted by $d(t, \theta_1, \theta_2)$, for $t = 1, 2, \ldots, n_{obs}$, which is the usual root mean square difference between chlorophyll image and model chlorophyll field interpolated to the observation grid. In the present case n_{obs} is 362, since three days of observations (Julian days 289, 293 and 336) are missing. θ_1, θ_2 are the two imprecisely known biological parameters, which are to be estimated using polynomial chaos approach. Since for distance computation, we need simulated chlorophyll (in mg Chl a m^{-3}) and not phytoplankton, we have to convert the latter to the former. The chosen conversion factor is 1 and has been adopted from Vinayachandran *et al.*[21]. Note that the relevant parameters have been denoted as θ_1, θ_2 and not as g and g_r mentioned earlier. This is because the polynomial chaos expansion used by us (as will be shown presently), requires Legendre polynomials, which are functions of a dimensionless variable in the range [–1, +1].

For the estimation, we treat the biological parameters to be stochastic and make the strong assumption that all the uncertainties in the model hindcast are due to uncertainties in these two parameter values. We also adopt the hypothesis that the parameter probability distribution is a uniform one. Once the minimum and maximum values of each parameter are known, a simple linear mapping is sufficient to transform from physical variables g and g_r to the dimensionless arguments θ_1, θ_2 of the distance function. Once these are estimated using polynomial chaos, it is trivial to revert to the actual dimensioned variables using inverse mapping.

We adopt a general notation, and denote by function f the property of interest. In general, the function depends on a space coordinate $\mathbf{x}$, a time coordinate t and the θ parameters. Although there are two independent parameters, we describe the methodology only for one parameter for notational simplicity. Assuming independence of the growth and decay parameters, the theory translates in a seamless manner to the two (or greater) parameter case. In the polynomial chaos, the function is expanded as

$$f(\mathbf{x}, t, \theta) \approx \sum_{k=0}^{K} a_k(\mathbf{x}, t)\, \beta_k(\theta), \tag{2}$$

where $a_k(\mathbf{x}, t)$ are θ-independent expansion coefficients, and the kth basis function $\beta_k(\theta)$ is a polynomial of order k in the parameter space defined by θ. The upper limit of summation, i.e. K determines the quality of the approximation. If $K = \infty$, the expansion is exact. In practice, K has to be a small number, because of computational constraints. Fortunately, the series converges rapidly and K can be chosen to be small. Also, the polynomials are to be orthogonal (with respect to a weight function) in their domain of definition, symbolically written as

$$\langle \beta_k, \beta_l \rangle_p = \delta_{kl} N_k. \tag{3}$$

Here the scalar product indicates integration in the parameter space, and the subscript p denotes that there is a weight function, the probability distribution $\boldsymbol{p}(\theta)$. δ_{kl} is the usual Kronecker delta function and N_k is a normalization factor, specific to the kth polynomial. In our notation

$$N_k = \langle \beta_k, \beta_k \rangle_p. \tag{4}$$

For the case considered by us (uniform distribution), the polynomials are the well-known Legendre polynomials. The expansion coefficients a_k are calculated as

$$a_k = (1/N_k) \langle f, \beta_k \rangle_p. \tag{5}$$

Since the property concerned (e.g. the distance of model chlorophyll from the observed chlorophyll) is not given as an analytical function, the integral appearing in eq. (5) has to be computed numerically. Usually, a Gaussian quadrature is employed to calculate the integral, and hence

$$a_k(\mathbf{x}, t) \approx (1/N_k) \sum_{i=0}^{K} f(\mathbf{x}, t, \theta^{(i)})\, \beta_k(\theta^{(i)}) \omega_i. \tag{6}$$

Here $\theta^{(i)}$ are quadrature points in parameter space and ω_i are the Gaussian quadrature weights.

Standard values (albeit, not optimized) of the parameters are available from the literature[20]. We choose minimum and maximum values of the distribution to be 0.5 and 2 times the standard values. The maximum order K was chosen to be 6. Thus, 49 model runs have to be performed. Once these runs are performed, it is trivial to calculate any function dependent on these stochastic parameters using eq. (2).

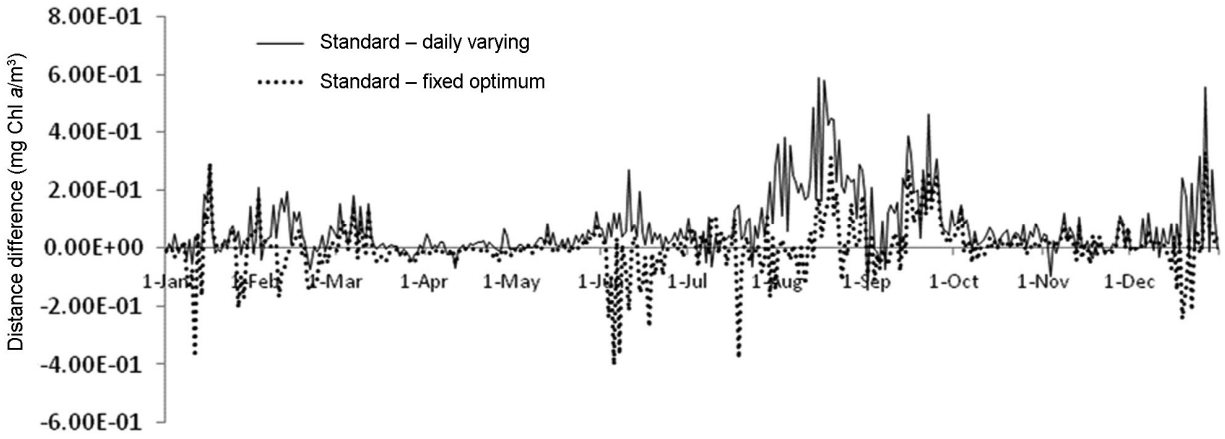

Figure 1. Time-series plot of the difference between daily distances obtained in the standard run and emulation with a fixed set of optimum parameters (dotted line). The same for emulation with daily varying parameters is also plotted (solid line).

Daily emulated distances using the parameters of the standard run were compared with corresponding distances obtained from the standard simulation, and a near-perfect match was obtained. Thus costly model simulation can indeed be replaced by a simple emulation. For the optimization, we calculated the distance function each day and for all the 49 runs. Then we eliminated the time dependence by time-averaging. Using them along with eq. (2), and a suitable minimization routine, we found the two optimized parameters. However, use of these time-independent parameters led to the counterintuitive result that on individual days performance of the run with optimized parameters was worse than that of standard run. There are twofold reasons for this result. The model is so well-tuned to observations in an average sense[20], that it is difficult to improve upon its result each day using a fixed set of parameters.

Following Mattern *et al.*[10], we assumed that the two biological parameters vary with time and resorted to daily minimization of time-dependent distance function. After obtaining the optimized time-varying parameters, we again computed daily emulated distances. The daily differences from the standard run were computed and compared with those from emulation with a fixed set of optimum parameters found from the previous exercise. The result (Figure 1) clearly shows that the difference is always positive, indicating the benefit of using daily varying parameters. This difference set is also always more than the other difference set. In fact, the second distance even turned negative on individual days, indicating deterioration of simulation. The reason has been mentioned above. Thus, there is indeed merit in using daily varying optimized parameters.

The time evolution of individual distances appears to be noisy (corresponding figure not shown). This is due to outliers in the observation caused by a large number of missing values, which introduce noise in the optimized parameter set. It is thus quite logical to infer that reduction of noise in the parameters by some kind of smoothing may cause reduction of noise in the daily distances. The ultimate aim of any parameter optimization is, of course, the use of these parameters for carrying out model simulation or emulation. Smoothing causes the model evolutions to look less noisy and more realistic. Accordingly, we employed the singular spectrum analysis (SSA) technique of noise reduction[22]. Without going into the details of the technique, we just note that the method depends on an eigensystem analysis, and the degree of smoothing depends on the number of eigenvectors retained. The relation is, however, an inverse one. More eigenvectors retained means less smoothing and vice versa. Unlike the case of daily minimization, smoothing of a parameter time series requires a continuous set of data. In our dataset, there were no data for the 289th Julian day in 2006. Hence we used only the first 288 days of data for carrying out SSA smoothing. Figure 2 shows the daily varying maximum phytoplankton growth rate as well as two smoothed versions with different degrees of smoothing. The corresponding figure for the other parameter (maximum zooplankton grazing rate) is similar in nature and is not shown here. Although smoothing may make model simulations look less noisy, more smoothing means the resulting distances would deviate more from the optimized distances. We use the word 'nearness' (number of eigenvectors retained) to denote the fact that at more nearness, the parameters are closer to the optimized parameters (less smoothing) than at less nearness. The highest chosen nearness is 20 and the lowest is 2. For each nearness, one can find the corresponding daily varying parameters and then carry out model emulation. Then, one can find daily emulated distances and average them (Table 1). We can see from this Table 1 that with increase in the degree of nearness, the average distance value approaches the distance value for the unsmoothed

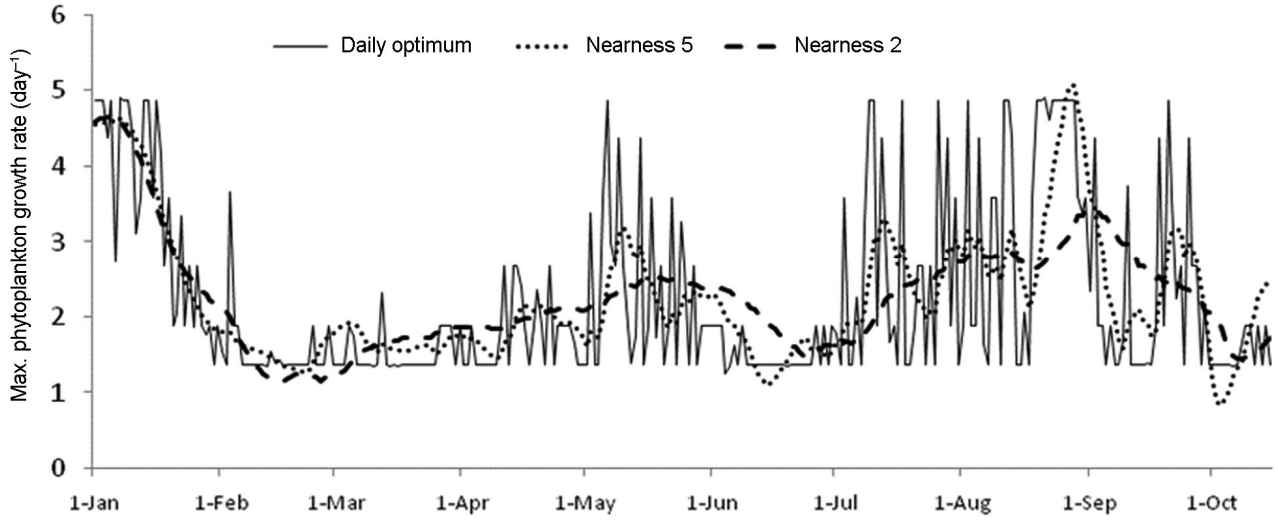

Figure 2. Time evolution of daily varying maximum phytoplankton growth rate for different degrees of nearness (explained in the text).

Table 1. The degree of nearness of the smoothed parameters and the daily-averaged distances found from emulation with these parameters. Perfect nearness means all eigenvectors have been used in SSA smoothing and the smoothed parameters are identical with the unsmoothed ones

Nearness	Daily-averaged distance (mg Chl *a* m^{-3})
Perfect	0.5282
20	0.5375
15	0.5408
10	0.5423
5	0.5505
2	0.5542

parameters. All the distances are, however, greater than the corresponding distances for the unsmoothed parameters. Interestingly, however, the average distance values are less than that for the standard run, which is 0.6011 mg Chl *a* m^{-3} (for first 288 days of standard run). This shows that the emulated chlorophyll fields are better than the fields for the standard run with fixed parameters. The fact that the lowest distance is obtained when there is no smoothing indicates that at least part of the improvement is due to over-fitting the data. This is because for low smoothing the emulated values fit even the outlying values and noise well, completely disregarding the model dynamics.

It is known that the Arabian Sea basin of the Indian Ocean experiences seasonal extremes in forcing and biological response alternating from calm, stratified near-oligotrophic conditions during the intermonsoon periods to strongly forced euphotic conditions during the southwest and northeast monsoons[23]. During the southwest monsoon there are large gaps in satellite data due to cloudy weather. Hence we select bimonthly averaged data for February–March, in which the Noctiluca blooms occurring in the northeast Arabian Sea are clearly visible. It is interesting to study how this bloom is represented by the model emulation. Figure 3 shows a comparison of the averaged chlorophyll in observations, standard simulation and model emulation. The bloom is well captured in observations as well as in standard simulation and emulation. However, abnormally low chlorophyll occurring in the southeast Arabian Sea in the standard run is replaced by more realistic values in the emulation, leading to a better match with observations.

For further analysis we computed daily averaged surface chlorophyll in satellite observations, standard run and model emulation with daily varying optimized set, separately in the Arabian Sea and Bay of Bengal. The two basins were chosen because of their differences in the surface chlorophyll characteristics. Figures 4 and 5 show the results for each of the basins. These figures show that the emulated chlorophyll with daily varying parameters is closer to the observed chlorophyll than the chlorophyll simulated in the standard run.

In the present study we report the results of emulating chlorophyll fields obtained from a coupled physical–biological model of the Indian Ocean. For this purpose two of the biological parameters of the model were treated as stochastic and the model simulations were approximated by a low-dimensional emulator, using polynomial chaos expansion. By minimizing a distance function representing model-data misfit, the optimum parameter values were obtained. The parameters showed a clear time-dependence.

Once we allowed the parameters to vary in time, better fit to observations could be achieved. Thus polynomial chaos proved to be an efficient tool for analysing the results of the biological part of the model. Restriction on the number of parameters, considered to be stochastic, is

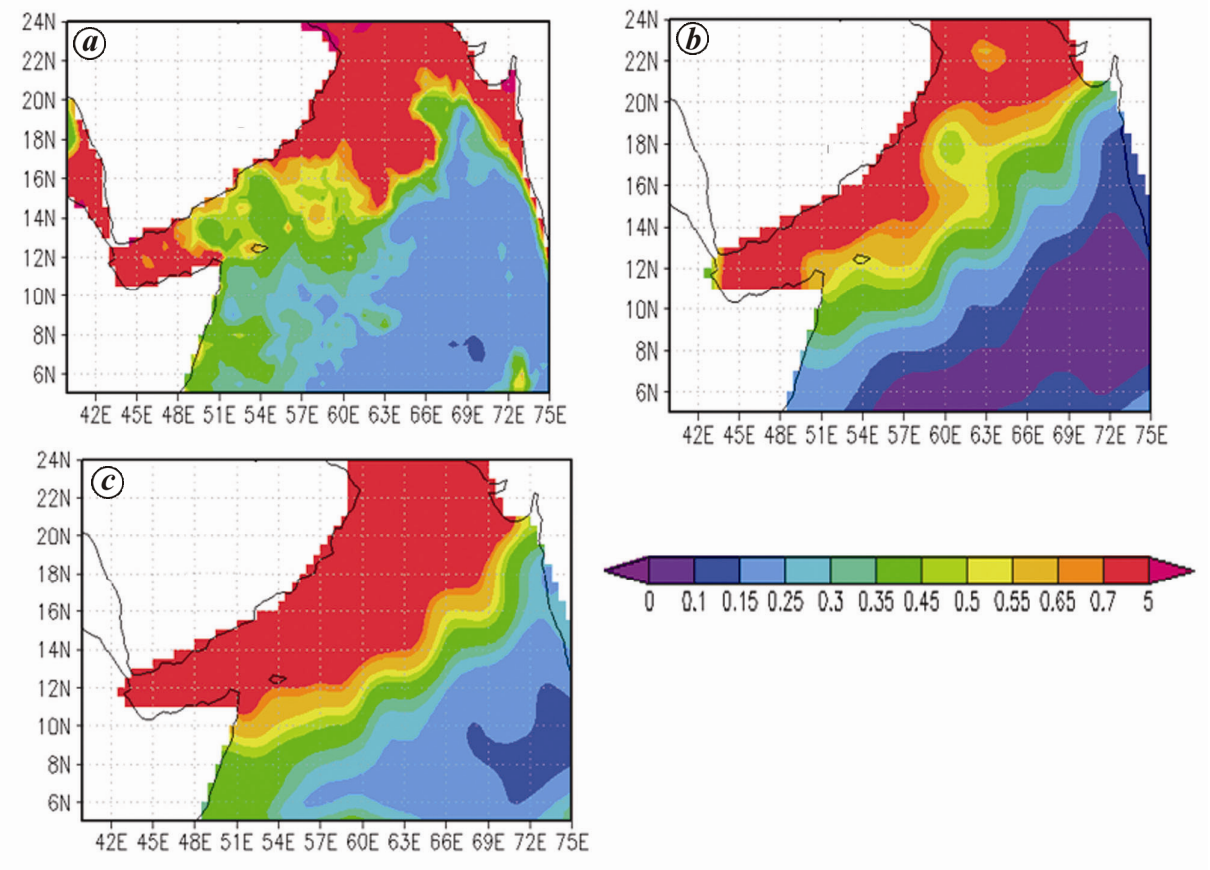

Figure 3. Bimonthy averaged chlorophyll concentrations in the Arabian Sea for February and March 2006. (***a***) SEAWIFS observations, (***b***) standard run and (***c***) model emulation with daily optimized parameters.

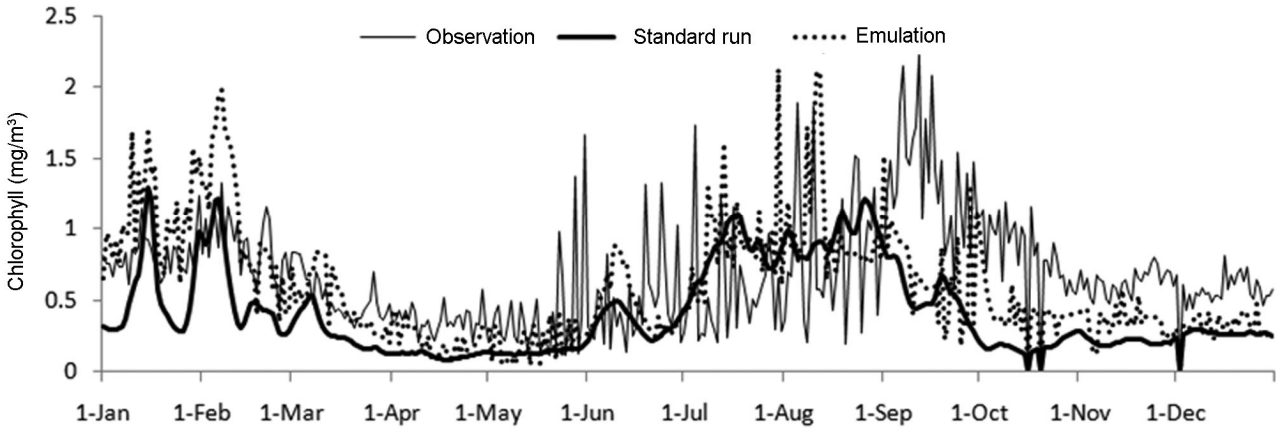

Figure 4. Comparison of chlorophyll estimates in the standard run (bold line) and emulation (dotted line) with time-varying parameters with observations (solid line) in the Arabian Sea.

admittedly a constraint. This restriction was imposed purely for computational consideration. Nevertheless, the advantage of this emulation technique is immense, since any model result with a particular combination of the two stochastic parameters could be obtained immediately, without the necessity to actually carry out the run.

Comparison of bimonthly averaged chlorophyll (averaged over February–March) has shown the advantage of the emulator approach (with daily optimized parameters) over the standard run. Comparison of emulated estimates with those in the standard run has been also done separately in the Arabian Sea and the Bay of Bengal

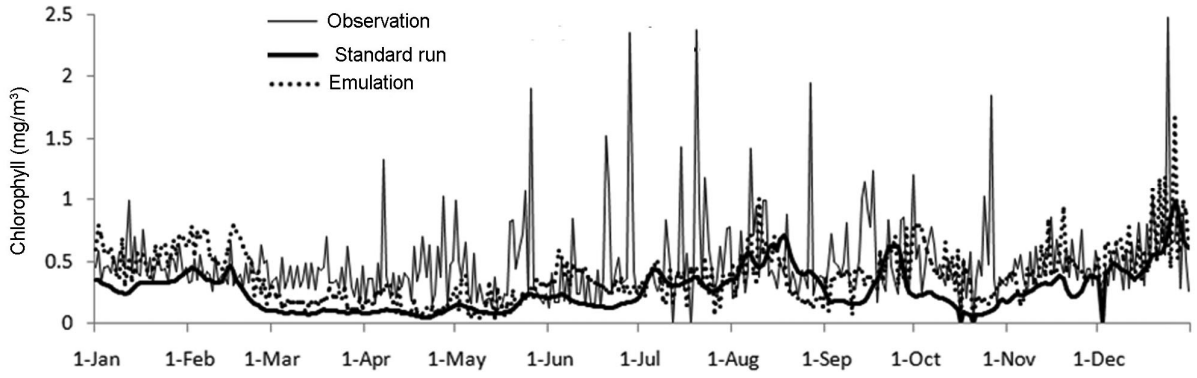

Figure 5. Same as in Figure 4, except for the Bay of Bengal.

basins of the north Indian Ocean. It has been found that, generally speaking, emulation with optimized time-varying parameters outperforms the standard run.

1. Lawson, L. M., Hofmann, E. E. and Spitz, Y. H., Time series sampling and data assimilation in a simple marine ecosystem model. *Deep Sea Res. II*, 1996, **43**, 625–651.
2. Powell, B. S., Arango, H. G., Moore, A. M., Di Lorenzo, E., Milliff, R. F. and Foley, D., 4DVAR data assimilation in the intra-Americas sea with the Regional Ocean Modeling System (ROMS). *Ocean Model*., 2008, **23**, 130–145.
3. Evensen, G., The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn*., 2003, **53**, 343–367.
4. Allen, J. I., Eknes, M. and Evensen, G., An ensemble Kalman filter with a complex marine ecosystem model: hindcasting phytoplankton in the Cretan Sea. *Ann. Geophys*., 2003, **21**, 399–411.
5. Mattern, J. P., Dowd, M. and Fennel, K., Particle filter-based data assimilation for a three-dimensional biological ocean model and satellite observations. *J. Geophys. Res*., 2013, **118**, 2746–2760, doi:10.1002/jgrc.20213.
6. Smedstad, O. and O'Brien, J. J., Variational data assimilation and parameter estimation in an equatorial Pacific Ocean model. *Prog. Oceanogr.*, 1991, **26**, 179–241; doi:10.1016/0079-6611(91)90002-4.
7. Losa, S. N., Kivman, G. A., Schröter, J. and Wenzel, M., Sequential weak constraint parameter estimation in an ecosystem model. *J. Mar. Syst*., 2003, **43**, 31–49.
8. Dowd, M., Estimating parameters for a stochastic dynamic marine ecological system. *Environmetrics*, 2001, **22**, 501–515.
9. Vossepoel, F. C. and van Leeuwen, J., Parameter estimation using a particle method: inferring mixing coefficients from sea-level observations. *Mon. Weather Rev*., 2007, **135**, 1006–1020.
10. Mattern, J. P., Fennel, K. and Dowd, M., Estimating time-dependent parameters for a biological ocean model using an emulator approach. *J. Mar. Syst*., 2012, **96–97**, 32–47.
11. Frolov, S., Baptista, A. M., Leen, T. K., Lu, Z. and van der Merwe, R., Fast data assimilation using a nonlinear Kalman filter and a model surrogate: an application to the Columbia River estuary. *Dyn. Atmos. Ocean*, 2009, **48**, 16–45.
12. Scott, V., Kettle, H. and Merchant, C. J., Sensitivity analysis of an ocean carbon cycle model in the North Atlantic: an investigation of parameters affecting the air–sea CO_2 flux, primary production and export of detritus. *Ocean Sci.*, 2011, **7**, 405–419.
13. Weiner, N., The homogeneous chaos. *Am. J. Math*., 1938, **60**, 897–936.
14. Askey, R. and Wilson, J. A., Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials. *Mem. Am. Math. Soc*., **2,** 1985, 319.
15. Wan, X. L. and Karniadakis, G. E., Beyond Wiener–Askey expansions: handling arbitrary PDF. *J. Sci. Comput*., 2006, **27**, 455–464; doi:10.1007/s10915-005-9038-8.
16. Xiu, D. B. and Karniadakis, G. E., The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM. *J. Sci. Comput*., 2002, **24**, 619–644; doi:10..137/S1064827501387826.
17. Lucas, D. D. and Prinn, R. G., Parametric sensitivity and uncertainty analysis of dimethylsulfide oxidation in the clear-remote marine boundary layer. *Atmos. Chem. Phys*., 2005, **5**, 1505–1525; doi:10.1029/2007JC004493.
18. Thacker, W. C., Srinivasan, A., Iskandarani, M., Knio, O. M. and LeHenaff, M., Propagating boundary uncertainties using polynomial expansions. *Ocean Model*., 2012, **43–44**, 52–63; doi:10.1016/j.ocemod.2011.11.011.
19. McCreary, J. P., Kohler, K. E., Hood, R. R., Smith, S., Kindle, J., Fischer, A. S. and Weller, R. A., Influences of diurnal and intraseasonal forcing on mixed-layer and biological variability in the Arabian Sea. *J. Geophys. Res*., 2001, **106**, 7139–7155.
20. Hood, R. R., Kohler, K. E., McCreary, J. P. and Smith, S. L., A four-dimensional validation of a coupled physical–biological model of the Arabian Sea. *Deep Sea Res. II*, 2003, **50**, 2917–2945.
21. Vinayachandran, P. N., McCreary, J. P., Hood, R. R. and Kohler, K. E., A numerical investigation of the phytoplankton bloom in the Bay of Bengal during northeast monsoon. *J. Geophys. Res*., 2005, **110**, doi.10.1029/2005JC002966.
22. Penland, C., Ghil, G. and Weickman, K. M., Adaptive filtering and maximum entropy spectrum with application to changes in atmospheric angular momentum. *J. Geophys. Res*., 1991, **96**, 22,659–22,671, doi:10.1029/91JD02107.
23. Smith, S. L., Codispoti, L. A., Morrison, J. M. and Barber, R. T., The 1994–1996 Arabian Sea Expedition: an integrated interdisciplinary investigation of the northwestern Indian Ocean to monsoonal forcing. *Deep Sea Res. II*, 1998, **45**, 1905–1915.

ACKNOWLEDGEMENTS. We thank the Director, Space Applications Centre (SAC), the Deputy Director, Earth, Ocean, Atmosphere, Planetary Sciences and Applications Area, and the Group Director, Atmospheric and Oceanic Sciences Group, SAC, Ahmedabad for their support and encouragement.

Received 3 March 2015; revised accepted 4 January 2016

doi: 10.18520/cs/v110/i8/1544-1549