View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by OceanRe

JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 112, C08001, doi:10.1029/2006JC003852, 2007

Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups

Marjorie A. M. Friedrichs,[1,2] Jeffrey A. Dusenberry,[3] Laurence A. Anderson,[3] Robert A. Armstrong,[4] Fei Chai,[5] James R. Christian,[6] Scott C. Doney,[3] John Dunne,[7] Masahiko Fujii,[5,8] Raleigh Hood,[9] Dennis J. McGillicuddy Jr.,[3] J. Keith Moore,[10] Markus Schartau,[4,11] Yvette H. Spitz,[12] and Jerry D. Wiggert[2]

Received 31 July 2006; revised 19 January 2007; accepted 2 May 2007; published 2 August 2007.

[1] Application of biogeochemical models to the study of marine ecosystems is pervasive, yet objective quantification of these models' performance is rare. Here, 12 lower trophic level models of varying complexity are objectively assessed in two distinct regions (equatorial Pacific and Arabian Sea). Each model was run within an identical one-dimensional physical framework. A consistent variational adjoint implementation assimilating chlorophyll-a, nitrate, export, and primary productivity was applied and the same metrics were used to assess model skill. Experiments were performed in which data were assimilated from each site individually and from both sites simultaneously. A cross-validation experiment was also conducted whereby data were assimilated from one site and the resulting optimal parameters were used to generate a simulation for the second site. When a single pelagic regime is considered, the simplest models fit the data as well as those with multiple phytoplankton functional groups. However, those with multiple phytoplankton functional groups produced lower misfits when the models are required to simulate both regimes using identical parameter values. The cross-validation experiments revealed that as long as only a few key biogeochemical parameters were optimized, the models with greater phytoplankton complexity were generally more portable. Furthermore, models with multiple zooplankton compartments did not necessarily outperform models with single zooplankton compartments, even when zooplankton biomass data are assimilated. Finally, even when different models produced similar least squares model-data misfits, they often did so via very different element flow pathways, highlighting the need for more comprehensive data sets that uniquely constrain these pathways.

Citation: Friedrichs, M. A. M., et al. (2007), Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups, *J. Geophys. Res.*, *112*, C08001, doi:10.1029/2006JC003852.

[1]Virginia Institute of Marine Science, College of William and Mary, Gloucester Point, Virginia, USA.
[2]Center for Coastal Physical Oceanography, Old Dominion University, Norfolk, Virginia, USA.
[3]Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, USA.
[4]School of Marine and Atmospheric Sciences, State University of New York at Stony Brook, Stony Brook, New York, USA.
[5]School of Marine Sciences, University of Maine, Orono, Maine, USA.
[6]Fisheries and Oceans Canada, Victoria, British Columbia, Canada.
[7]Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA.
[8]Creative Research Initiative, Hokkaido University, Hokkaido, Japan.
[9]Center for Environmental Science, University of Maryland, Cambridge, Maryland, USA.
[10]Earth System Science, University of California, Irvine, California, USA.
[11]Institute for Coastal Research, GKSS-Forschungszentrum, Geesthacht, Germany.
[12]College of Ocean and Atmospheric Sciences, Oregon State University, Corvallis, Oregon, USA.

Copyright 2007 by the American Geophysical Union.
0148-0227/07/2006JC003852

1. Introduction

[2] As knowledge regarding the complex components of marine ecosystems continues to grow, the models being developed to examine these systems are correspondingly becoming more complex as they include increasing numbers of organisms and biological processes. The range of models currently in use extends from the simplest three or four compartment nutrient/phytoplankton/zooplankton (NPZ) or nutrient/phytoplankton/zooplankton/detritus (NPZD) models [e.g., *Franks*, 2002; *Denman and Pena*, 2002; *Schartau and Oschlies*, 2003; *Kantha*, 2004] to complex models with 20 or more components including different types of plankton, multiple nutrients, and a microbial loop [*Bissett et al.*, 1999; *Moore et al.*, 2002; *Gregg et al.*, 2003; *Lancelot et al.*, 2005].

[3] The decision as to how much complexity to include in a marine ecosystem model boils down to a balancing act between unwanted detail and unjustified simplification [*Flynn*, 2005]. There are certain system feedbacks that are

linked to specific phytoplankton functional groups [*Anderson*, 2005; *Le Quere et al.*, 2005] that clearly can only be successfully simulated if multiple phytoplankton groups are included. However, before multiple phytoplankton groups can be modeled, the robustness of the parameterizations must be demonstrated, and data must be available to evaluate the individual phytoplankton functional types [*Hood et al.*, 2006]. The principle of Occam's razor, which states that when other means of comparison are eliminated, the simplest explanation (model) is the best one, is a cornerstone of much marine ecosystem modeling activity [*Flynn*, 2005]. That is, if two models demonstrate equal model skill, then the least complex is the preferable one. However, as *Flynn* [2005] warns: ''over application of Occam's razor results in you cutting your own throat,'' i.e., biological features should be omitted only after careful consideration.

[4] By including multiple phytoplankton functional groups (e.g., nanoplankton, diatoms, and diazotrophs) and multiple limiting nutrients, (e.g., nitrate, silica, iron), many models are achieving greater realism. However, because the number of parameters that must be specified increases by as much as the square of the number of state variables [*Denman*, 2003], completely constraining these complex models with existing oceanographic observations becomes increasingly problematic. As a result, there is a trade-off between the complexity and realism of a model and the degree to which it can be constrained given the available data. In the experiments presented here, model comparisons are carried out in an effort to understand how much complexity is warranted in these models.

[5] Most marine ecosystem models have been developed for limited geographic regions. Extrapolating these model structures to basin-scale or global applications can be quite speculative [*Evans*, 1999]. Currently, there is a critical need to identify ecosystem model structures and formulations that are geographically portable, i.e., are applicable over a number of diverse ecosystems. If such structures can be identified, and the reasons for their success understood, the oceanographic modeling community will be significantly closer to marine biogeochemical and ecological prediction. Thus a second goal of our model intercomparisons is to identify which ecosystem structures are able to perform well in diverse regions and physical settings.

[6] To address the issues posed above, model performance and portability must be objectively compared. In general, independent investigators apply distinct physical forcing fields to biogeochemical models; they then proceed to tune their models to varying degrees and assess them with different validation data. Thus it is extremely difficult to objectively compare the many models currently in use. In addition, quantitatively assessing model performance is not a straightforward task [*Evans*, 2003; *Arhonditsis and Brett*, 2004]. Certainly, a model fails when it cannot reproduce a data set that it was developed to describe, yet satisfying this condition is an insufficient test. When comparing a number of different ecosystem models, the one with the greatest number of tunable parameters might be expected to provide the best fit to a given data set, just as a higher-order polynomial model could be used to generate a better fit to a given data set than a lower-order polynomial. However, generating the lowest model-data misfits to a given data set does not imply that such a model will provide the best mathematical description of the ecosystem. On the contrary, models with large numbers of unconstrained parameters may be characterized by lower predictive ability if they have been overtuned and thus forced to fit noise in the data; the price associated with fitting noise in a given data set is a loss of predictive ability [*Friedrichs et al.*, 2006].

[7] To facilitate objective model comparisons, we have developed a set of regional ecosystem modeling test beds. The test beds consist of a one-dimensional (vertical) numerical framework which includes subroutines for diffusion, advection, light attenuation and sinking, physical forcing time series of temperature, irradiance, mixed layer depth, vertical velocity, and the horizontal advective divergence of nutrients, as well as biogeochemical data for either assimilation or evaluation. Finally, a data assimilation (parameter optimization) framework is included to ensure that the models are all tuned to about the same degree. Implementing various ecosystem models using the same physical forcing fields and reducing subjective tuning through the use of formal parameter optimization routines allows a quantitative comparison of different ecosystem models and modeling approaches.

[8] In addition to evaluating models according to how well they can be tuned to reproduce a given data set, it is critical to assess model performance based on how well models can reproduce data that were withheld from the tuning/optimization process. Such cross-validation experiments are frequently used in physical oceanography but have not yet been widely implemented in marine biogeochemical modeling applications, often due to the sparsity of the observations [*Friedrichs*, 2002]. However, the importance of these experiments is clear: models that are able to accurately simulate data from locations that were not included in the tuning process are likely to be more robust and thus more ''portable,'' i.e., better able to perform well in diverse regions and physical settings.

[9] This study compares the performance of 12 models with varying levels of ecosystem complexity and their ability to describe two environments characterized by distinct ecosystem dynamics: the equatorial Pacific and the Arabian Sea. In the following section the ecosystem models, the physical forcing fields, the data to be assimilated, and the assimilation experiments are described. Section 3 presents the results of this comparison effort. A synthesis of these results and their implications is discussed in section 4 and a summary is presented in section 5.

2. Methods

2.1. Ecosystem Models

[10] In this analysis, simulations from 12 different ecosystem models were compared. Almost all of these models are well documented in the literature, and therefore only their general characteristics and appropriate references are provided here. In instances where models are not published in the literature or where changes have been made to published models, the relevant parameterizations and parameter values are described in the auxiliary material[1].

[1] Auxiliary materials are available in the HTML. doi:10.1029/2006JC003852.

2.1.1. Model 1

[11] This four-component (phytoplankton, zooplankton, dissolved inorganic nitrogen (DIN), and detritus) ecosystem model represents a classic diatom-mesozooplankton system. Unlike most other of the models participating in this comparison, model 1 has been developed and calibrated specifically for the Arabian Sea [*McCreary et al.*, 2001; *Hood et al.*, 2003]. Parameterizations and parameter values used here have been taken directly from *McCreary et al.* [2001].

2.1.2. Model 2

[12] This five-component (phytoplankton, heterotrophs, DIN, dissolved organic nitrogen (DON), and detritus) ecosystem model was developed by *Hood et al.* [2001] for use at the Bermuda Atlantic Time Series station and subsequently applied over the tropical and subtropical Atlantic [*Hood et al.*, 2004; *Coles et al.*, 2004]. Although this model was originally developed with a sixth diazotrophic state variable, here the model is implemented without this component. The heterotroph compartment is considered to represent the sum of all heterotrophic processes that are facilitated by bacteria, microzooplankton, and mesozooplankton. This model emphasizes the microbial loop by having all organic matter cycle through the heterotroph compartment at relatively high rates. Parameterizations and parameter values used here have been taken directly from *Hood et al.* [2001].

2.1.3. Model 3

[13] In this five-component (phytoplankton, chlorophyll-a (chl), zooplankton, DIN, and detritus) model, phytoplankton are limited by nutrients and light and are grazed by one class of zooplankton with a quadratic dependence on prey concentration [*Denman and Pena*, 1999]. The single nutrient compartment implicitly includes nitrate, ammonium, and urea. The detritus compartment combines dissolved, suspended, and sinking organic matter, with a constant sinking rate. Growth and remineralization rates are temperature-dependent. Nongrazing mortality includes both linear and quadratic terms for both phytoplankton and zooplankton. Chl is a separate prognostic variable based on a light-dependent chl:N ratio. Parameterizations and parameter values used here are provided in the auxiliary material.

2.1.4. Model 4

[14] This six-component (nitrate, ammonium, phytoplankton, chlorophyll, zooplankton, detritus) ecosystem model is a slightly modified version of a model used in the Gulf of Maine (L. A. Anderson et al., Fitting a biological model to 2-D data: The seasonal cycle of phytoplankton in Wilkinson Basin, Gulf of Maine, manuscript in preparation, 2007) and very similar to that of *Besiktepe et al.* [2003]. The chl:N ratio adjusts toward a depth-dependent equilibrium chl:N ratio with a timescale of 6 days. Details of this model appear in the auxiliary material.

2.1.5. Model 5

[15] This six-component model is a simplified model version of *Schartau et al.* [2007]. The test bed version lacks the distinction within the pool of dissolved organic substances. Carbon and nitrogen fluxes are resolved between five compartments: nutrients, phytoplankton, detritus, dissolved organic matter (DOM), and heterotrophs. Total alkalinity is regarded as a separate, sixth state variable, in order to specify the carbonate system in conjunction with dissolved inorganic carbon. Parameterizations for phytoplankton growth are adopted from *Geider et al.* [1998]. Therefore the model accounts for phytoplankton acclimation to nutrient and light availability, as well as to temperature changes. The loss of phytoplankton biomass is assumed to be due to grazing, particle aggregation, exudation, and leakage. The closure for the mass flux is described by maintenance respiration and by the breakdown of particulate organic matter into DOM, which is eventually mineralized. Heterotrophic respiration is determined by a restoring term that restores zooplankton biomass toward a constant stoichiometric carbon-to-nitrogen ratio. As a consequence, heterotrophic respiration increases when phytoplankton with a high carbon-to-nitrogen ratio has been grazed. The model has not been tuned specifically to the equatorial Pacific station nor to the Arabian Sea site. Parameterizations and parameter values that differ from those described by *Schartau et al.* [2007] are provided in the auxiliary material.

2.1.6. Model 6

[16] This nine-component model contains two size classes of phytoplankton, zooplankton, and detritus, as well as ammonium, nitrate, and iron. This model was developed to simulate the high nutrient-low chlorophyll conditions observed in the equatorial Pacific [*Christian et al.*, 2002]. The implementation here is identical to that described by *Christian et al.* [2002], except that the a priori maximum grazing rate parameter for large phytoplankton was increased to 50 d^{-1}.

2.1.7. Model 7

[17] This nine-component model is a modified form of model 6. Several changes designed to promote more dynamic phytoplankton bloom behavior were instituted so that the observed interregional variation in bloom magnitude in the Indian Ocean would be more accurately represented [*Wiggert et al.*, 2006]. Structural modifications to the ecosystem include application of hyperbolic mortality to the large phytoplankton and zooplankton and allowing zooplankton to graze on detritus (coprophagy). The implementation here is identical to that described by *Wiggert et al.* [2006], except that the mortality of large phytoplankton has been increased to 1.05 d^{-1}.

2.1.8. Model 8

[18] This nine-compartment model contains two functional groups of phytoplankton (picoplankton and diatoms), zooplankton (microzooplankton and mesozooplankton), and detritus (particulate organic nitrogen and biogenic silica), as well as nitrate, ammonium, and silicate [e.g., *Chai et al.*, 2002; *Dugdale et al.*, 2002; *Jiang et al.*, 2003]. This model was developed to simulate the silicate limitation on the diatom growth in the equatorial Pacific upwelling region, which was not reproduced with nitrogen-based ecosystem models. Minor changes were made to this model [*Chai et al.*, 2002] and are documented in the auxiliary material.

2.1.9. Model 9

[19] This 10-component model is a prognostic implementation of the *Laws et al.* [2000] model, which includes state variables for two dissolved nutrient pools (dissolved inorganic nutrients and dissolved organic nutrients), two phytoplankton size classes (large and small), and five heterotrophic

groups (bacteria, flagellates, ciliates, filter feeders, and carnivores). The model also includes a detritus compartment. The *Laws et al.* [2000] model structure is composed of two quasi-independent food webs. One is based upon primary production of large phytoplankton, which ultimately produces detritus that sinks, and the other is based upon primary production by small phytoplankton, which ultimately produces dissolved organic matter and bacteria (i.e., a "microbial loop") that does not sink. The relative production of these two food webs is controlled by temperature, with smaller size classes being favored by increasing temperature. Equations for the forward implementation of this model are provided in the auxiliary material.

2.1.10. Model 10

[20] This 11-compartment model contains two phytoplankton functional groups (diatoms and nondiatom small phytoplankton including coccolithophorids), three zooplankton functional groups (copepods, krill, and microzooplankton including foraminifera), as well as nitrate, ammonium, silicate, particulate and dissolved organic nitrogen, and biogenic silica [e.g., *Fujii et al.*, 2002, 2007; *Yamanaka et al.*, 2004; *Kishi et al.*, 2007]. The model was originally developed to simulate a lower trophic level ecosystem and its linkage with a higher trophic level ecosystem (fish) in the subarctic western North Pacific. The model is capable of reproducing major spring and minor fall diatom blooms by incorporating silicate limitation of diatom growth and seasonal vertical migration of the copepods, both of which are significant biogeochemical characteristics of the western subarctic Pacific. Parameterizations and parameter values that differ from those given by *Fujii et al.* [2002] are provided in the auxiliary material.

2.1.11. Model 11

[21] This 11-component model based on the allometric formulation of *Dunne et al.* [2005a, 2005b] tracks N and Fe in small and large phytoplankton, sinking detrital organic matter, two kinds of dissolved organic matter, and nitrate, ammonia, and dissolved iron nutrients. The growth of phytoplankton is described through colimitation by N, Fe, and light with variable chl:N ratios wherein the Fe:N ratio is allowed to modulate the chl:N ratio via the Fe:N ratio of phytoplankton, consistent with observations. Grazing of small and diazotrophic phytoplankton is set proportional to their concentration to the second power, consistent with an instantaneous steady state with an implicit grazer population. Grazing of large phytoplankton is set proportional to their concentration to the 4/3rd power, consistent with a moderate imbalance with an implicit grazer population providing the potential for phytoplankton blooms. The grazing and food web processing formulations were calibrated to global field data [*Dunne et al.*, 2005a, 2005b]. For both small and large phytoplankton, zooplankton concentration does not enter into the grazing term. Rather than exerting active control on the phytoplankton loss function, zooplankton growth is set as a constant fraction of the small and large phytoplankton-driven grazing rate. Zooplankton loss undergoes first-order growth with an Eppley temperature dependent rate. Iron scavenging is assumed to be second-order with respect to dissolved iron concentration as a proxy for colloidal scavenging.

2.1.12. Model 12

[22] The Biogeochemical Elemental Cycling 24-component model was originally developed for use with the National Center for Atmospheric Research Community Climate System Model. The version used here is nearly identical to that used by *Moore et al.* [2004]. It includes multiple limiting nutrients (N, P, Si, and Fe), multiple phytoplankton functional groups (picoplankton/nanoplankton, diatoms, diazotrophs), a single size adaptable zooplankton pool, and two detritus pools. Variable elemental composition is included as well as photoacclimation of chlorophyll. For this analysis, air-sea exchange dynamics and inorganic carbon thermodynamics are not modeled. The iron-scavenging dynamics are modified to use a single ligand equilibrium formulation following *Doney et al.* [2006]. A priori parameter values are taken from *Moore et al.* [2004], with the exception of the initial slopes (alpha values) and P_{max} of the P versus I curves for the phytoplankton groups. The alpha and P_{max} values were doubled from the *Moore et al.* [2004] values to accommodate a change from daily averaged irradiance fields to a diel cycle.

2.2. Physical Forcing Fields

[23] Time series of photosynthetically active radiation (PAR), temperature, vertical diffusivity, vertical velocity, and mixed-layer depth (MLD) are required to run the ecosystem models in the one-dimensional test bed framework. For the Arabian Sea site, PAR and temperature time series were obtained from the Office of Naval Research central mooring [*Weller et al.*, 1998; *Kinkade et al.*, 1999] and for the Equatorial Pacific site these time series were obtained from the Tropical Atmosphere Ocean mooring array (www.pmel.noaa.gov/tao) [*McPhaden et al.*, 1998]. Velocity and temperature data (above 120 m) from these moorings were also used to compute vertical diffusivity [*Pacanowski and Philander*, 1981]. Below 120 m a background value of 10^{-4} m^2 s^{-1} was assumed. Although this background value is somewhat higher than most microstructure and tracer measurements can support, this value is appropriate for one-dimensional (1-D) models as a means to compensate for the lack of explicit representation of 3-D physical processes of nutrient supply such as internal wave activity [*Friedrichs and Hofmann*, 2001] and mesoscale eddies [*McGillicuddy and Robinson*, 1997].

[24] MLD and vertical velocity time series were obtained for both sites from an interannual run using a reduced-gravity, primitive equation ocean model [*Murtugudde et al.*, 1996; *Murtugudde and Busalacchi*, 1999] with a variable depth mixed layer overlying 19 sigma layers. In this model, mixed layer thickness is determined using a "hybrid" mixed layer model [*Chen et al.*, 1994] that considers both wind stirring and shear instability. The thickness of each of the remaining 19 layers is a constant fraction of the total vertical distance between the base of the mixed layer and bottom of the model domain. In the Pacific the model grid is stretched to give a greater latitudinal resolution (1/3°) near the equator, with a uniform longitudinal resolution of 1°; in the Indian Ocean the model resolution is a uniform 1/2° longitude and 1/3° latitude.

[25] The effect of the horizontal advective divergence/convergence of biological quantities was examined through a scaling analysis using output from coupled biological-

physical models. In the Arabian Sea, two three-dimensional coupled biological-physical models [*Hood et al.*, 2003; *Wiggert et al.*, 2006] showed the magnitude of horizontal advective divergence to be small in comparison to vertical advection and other biological source/sink terms [*Friedrichs et al.*, 2006], and thus horizontal advective divergence was neglected at this site. A scaling analysis in the central equatorial Pacific [*Friedrichs and Hofmann*, 2001] revealed that the horizontal advective divergence of nitrate is a key process that may have first-order effects on the biogeochemical models implemented in this region. Therefore a time series of the inverse of the characteristic timescale for horizontal advection of nitrate (H_N) was computed from a three-dimensional coupled biological-physical model [*Christian et al.*, 2002; *Murtugudde et al.*, 1996] over a one-degree length scale:

$$H_N(z,t) = \frac{u}{N}\frac{\partial N}{\partial x} + \frac{v}{N}\frac{\partial N}{\partial y}$$

(Note that in this equation N, u and v as obtained from the 3-D model are all functions of x, y, z, and t.) Thus for the equatorial Pacific each model includes the advective term $-N(z,t)\,H_N(z,t)$ as an additional sink/source in its equation for the time rate of change of nitrate. Analogous terms for the horizontal advective divergence of silicate, phosphate, and iron were also used in those models containing these additional nutrients. Because H_N is partially dependent on the particular biological model used in the work of *Christian et al.* [2002], this formulation will inevitably add some additional error to the cost function; however, given that the DIN component of the cost function is typically only a small fraction of the total cost magnitude, this will not significantly affect our comparison results.

2.3. Model Implementation

[26] The ecosystem model equations are solved using a second-order Runge-Kutta scheme. Vertical advection and detrital sinking are described with a third-order direct space-time upwind-biased scheme [*Hundsdorfer and Trompert*, 1994] and the Sweby flux limiter [*Sweby*, 1984] and were simplified to work for 1-D (vertical) advection only. Vertical diffusion is applied using a Crank-Nicholson vertically variable diffusion operation [*Press et al.*, 1986], with a closed upper boundary and an open bottom boundary. Constant nitrate concentrations of 25 mmol N m^{-3} and 16 mmol N m^{-3} at the bottom boundary (150 m) are assumed in the Arabian Sea and equatorial Pacific, respectively. Detrital sinking velocities are individually chosen for each ecosystem model. At each time step, all state variables are homogenized throughout the mixed layer.

[27] The subsurface light field was computed using a downwelling attenuation coefficient (K_d) of:

$$K_d = K_w + K_{chl}Chl(z)$$

where K_w = 0.05 m^{-1} represents the attenuation due to water and K_{chl} = 0.1 m^{-1} (mg chl m^{-3})$^{-1}$ represents that due to the chlorophyll. Models without explicit phytoplankton carbon used a Redfield C:N ratio to convert from phytoplankton nitrogen to carbon units; C:chl ratios were model-dependent. This attenuation scheme and these attenuation coefficients were chosen as they gave the best fit to chlorophyll and PAR observations at both sites.

[28] The models were run from 1 October 1994 through 1 January 1996 for the Arabian Sea site and 1 October 1991 through 1 January 1993 for the equatorial Pacific site, with a time step of 1 hour and a depth resolution of 10 m. Initial and bottom boundary conditions were identical for all 12 models. These values were taken from the biogeochemical data available at each site.

2.4. Biogeochemical Data

[29] Four distinct data types were assimilated into the ecosystem models. In situ cruise observations of phytoplankton chlorophyll-a (ChL) DIN, and primary production, as well as time series of export flux from sediment traps were utilized from both the U. S. Joint Global Ocean Flux Study (JGOFS) equatorial Pacific Process Study (four cruises between February and November 1992 [*Murray et al.*, 1995]) and Arabian Sea Process Study (six cruises between January and December 1995 [*Smith et al.*, 1998]).

[30] Data were downloaded from the U. S. JGOFS Web site http://usjgofs.whoi.edu/jg/dir/jgofs/. For each Arabian Sea (AS) cruise, only data from station S7 (16.0°N, 62.0°W) were utilized and for each equatorial Pacific (EP) cruise (140°W), only data within one degree of the equator were used. Chl and primary production data (24-hour in situ ^{14}C incubations) were posted on this Web site by J. Marra and R. Barber and are available for five of the six Arabian Sea cruises [*Barber et al.*, 2001] and all four equatorial Pacific [*Barber et al.*, 1996] cruises. Total dissolved inorganic nitrogen (DIN = $NO_3^- + NO_2^- + NH_4^+$) were computed from data posted by L. Codispoti for the AS cruises [*Morrison et al.*, 1998] and by C. Garside and P. Wheeler for the EP cruises [*Garside and Garside*, 1995]. When multiple nutrient profiles were available for the same day, these were averaged prior to assimilation. All data were interpolated to the model grid. This resulted in six DIN and five chlorophyll and productivity profiles in the Arabian Sea and 40 DIN and 27 chlorophyll and productivity profiles in the equatorial Pacific. In addition to cruise data, particulate nitrogen export flux measurements are available in the AS from the 800 m sediment trap located at 16.0°N, 61.5°W and in the EP from the 880 m sediment trap, both posted by S. Honjo and J. Dymond [*Honjo et al.*, 1995, 1999]. Although these depths are below the bottom of the model domain (located at 150 m), model equivalents of these export fluxes are computed by applying the flux attenuation formula of *Martin et al.* [1987] to extrapolate from the lowest model layer detrital concentrations.

[31] In initial experiments in which DIN data were assimilated over the entire model domain (0–150 m), the resulting optimized ecosystem parameter values were inconsistent with the range of observed estimates. This outcome was a direct result of the assimilation attempting to compensate for a too diffuse nutricline that is a result of the physical framework employed and not of ecosystem parameter choices [*Friedrichs et al.*, 2006]. Thus only mixed layer DIN concentrations were assimilated so that the biogeochemical assimilation scheme would not attempt to compensate for this shortcoming in the physics. Primary production and chlorophyll cruise data were

Table 1. Mean, Standard Deviation, Weight, and Inverse Weight for Each Data Type at Both the Equatorial Pacific and Arabian Sea Test Bed Locations

	Eq. Pac. Nitrate mmolN m^{-3}	Eq. Pac. Chlorophyll-a mg chl m^{-3}	Eq. Pac. Productivity mmolC m^{-3} d^{-1}	Eq. Pac. Export mmolC m^{-3} d^{-1}	Arab. Sea Nitrate mmolN m^{-3}	Arab. Sea Chlorophyll-a mg chl m^{-3}	Arab. Sea Productivity mmolC m^{-3} d^{-1}	Arab. Sea Export mmolC m^{-3} d^{-1}
Mean	5.62	0.26	10.2	0.77	3.67	0.43	19.6	1.84
St. Dev.	1.87	0.09	8.51	0.47	2.48	0.22	19.2	1.44
Weight (W)[a]	1.87	38.9	0.82	7.45	1.41	15.91	0.36	2.43
Inv. Wt. (W^{-1})	0.54	0.03	1.22	0.13	0.71	0.06	2.78	0.41

[a]Weights are determined using equation (2) (see text), where the value of C is equal to 3.5, except for productivity for which $C = 7$.

assimilated over the full depth to which data were available.

2.5. Parameter Optimization Scheme

2.5.1. Variational Adjoint Method

[32] The variational adjoint method of data assimilation [e.g., *Lawson et al.*, 1995] was used to objectively determine optimal parameter values, such that the differences between the model solution and the observations are minimized. This method consists of (1) a numerical model, (2) the cost function, which is a measure of the misfit between the predicted and observed variables, (3) an adjoint model, which is used to compute the gradient of the cost function with respect to the subset of model parameters which will be adjusted (control parameters), and (4) an optimization procedure that uses this information to determine the adjustments to the control parameters that will minimize the cost function.

[33] Starting with an initial guess for the model parameter set, the numerical model is run in order to obtain a value of the cost function. The adjoint of the model, obtained from the Tangent linear and Adjoint Model Compiler (TAMC) [*Giering and Kaminski*, 1998] is then run backward in time in order to compute the gradients of the cost function with respect to the model control parameters. Values of these gradients are passed to a limited memory quasi-Newton optimization procedure [*Gilbert and Lemaréchal*, 1989], which computes the optimal direction toward the minimum of the cost function and the optimal step size in that direction. New values of the control parameters are found, and the procedure is repeated in an iterative manner until a convergence criterion based on the norm of the gradient of the cost function has been satisfied. In order to test whether a robust minimum of the cost function has been identified, the initial estimates of the control parameters are adjusted by 10–50%, and the assimilation process is repeated. This procedure did not result in significantly different estimates of the optimized parameter values; i.e., local minima were not found to be a significant problem in these experiments. The lack of local minima and the corresponding insensitivity to initial parameter guesses are results of the method used to identify the most appropriate control parameters (see section 2.5.3.).

[34] Uncertainties in the estimated values of the model control parameters were computed from a finite difference approximation of the complete Hessian matrix (the matrix of the second derivatives of the cost function with respect to the control parameters). When computed at the minimum of the cost function, the inverse of the Hessian matrix can be used to estimate not only the errors for the optimal parameter estimates but also parameter correlations and the sensitivities of the cost function to each parameter [*Tziperman and Thacker*, 1989; *Matear*, 1995].

2.5.2. Cost Function

[35] The cost function, J, is a measure of the misfit between the predicted variables (a) and the observed variables ($\hat{a}$), and can be expressed as a weighted sum of squares:

$$J = \Psi \sum_{k=1}^{K} \frac{1}{\psi_k} \sum_{m=1}^{M} \frac{W_{km}^2}{N_{km}} \sum_{j=1}^{N_{km}} \left(a_{jkm} - \hat{a}_{jkm}\right)^2 \tag{1}$$

The sums are carried out over the number of test beds ($K = 2$ if data from both the AS and EP are assimilated; $K = 1$ if data from only one site are assimilated), the number of different data types ($M = 4$; chl, productivity, export, and nitrate), and the number of observations (N_{km}) for each data type. The weights (W_{km}) are inversely proportional to the standard deviations (σ_{km}) of the observations:

$$W_{km} = \frac{C_m}{\sigma_{km}} \tag{2}$$

In order to ensure that similar relative misfits at the different locations give similar cost function contributions and that the EP contributions are not weighted more strongly than the more variable (monsoonal) AS contributions [*Schartau and Oschlies*, 2003], ψ_k is defined as a function of the variance of the observations (σ_{km}^2):

$$\psi_k = \frac{1}{2} \sum_{m=1}^{M} \frac{\overline{\hat{a}_{km}}^2}{\sigma_{km}^2}$$

[36] The weights, W_{km}, represent not only the uncertainties associated with the accuracy of the observations but also our confidence in whether our modeled quantities truly represent the data. Therefore the factor C_m is included (equation (2)) to increase the weight of the primary production components of the cost function, which would otherwise be extremely low due to the high variance of the productivity data (Table 1). To ensure a fair comparison between models, identical values of C_m are used for all participating models. Thus the values of the inverse weights W_{km}^{-1} can be thought of as roughly representing a significance threshold for the magnitude of $a_{jkm} - \hat{a}_{jkm}$. In other words, if the model data difference $|a_{jkm} - \hat{a}_{jkm}|$ is less than or equal to the inverse weights

(W_{km}^{-1}) then this difference is assumed to be insignificant. In addition, the normalization factor Ψ

$$\Psi = \frac{1}{M}\left(\sum_{k=1}^{K}\frac{1}{\psi_k}\right)^{-1} \quad (3)$$

is included such that two values of J are not significantly different if $|a_{jkm} - \hat{a}_{jkm}| \leq W_{km}^{-1}$ and

$$J \leq \Psi\sum_{k=1}^{K}\frac{M}{\psi_k} = K$$

i.e., if two cost functions (J) differ by less than the number of sites (K), their difference is not considered to be significant.

2.5.3. Selection of Control Parameters

[37] A critical component of any application of a parameter optimization method to a marine ecosystem model is the selection of the control parameters. It is only possible to optimize parameters to which the cost function is highly sensitive: a parameter to which the model-data misfit is insensitive cannot be estimated with any degree of certainty. In addition, two highly correlated parameters cannot be simultaneously estimated successfully, since a change to one of the parameters will be counteracted by a change in the other, with multiple pairs of parameter values producing indistinguishable results.

[38] In general, the greater the number of control parameters, the lower the cost function; thus a more complex model with more tunable parameters will produce a lower cost function. However, as the number of optimized parameters grows and the corresponding uncertainty in these estimates increases, models tend to be less able to reproduce unassimilated (independent) data collected from different times or locations, i.e., they have lower predictive ability [*Friedrichs et al.*, 2006]. More complex models with a greater number of tunable parameters are particularly prone to this behavior. Consequently, the selection of control parameters is a critically important component of this model intercomparison exercise.

[39] A variation of the method introduced by *Friedrichs et al.* [2006] was implemented to objectively and systematically choose the subset of parameters for optimization. Initially, data from both sites were assimilated, all parameters were optimized, and the sensitivities of the cost functions to each model parameter and the correlations between each pair of parameters were computed from the inverse of the Hessian matrix. The optimized parameter to which the cost function was least sensitive, i.e., having the greatest normalized uncertainty, was then fixed to its original value, and another assimilation simulation was conducted with one fewer control parameter. This process was continued until a subset of control parameters was identified, for which no control parameter had an uncertainty greater than 100%. Correlations between parameters within the subset were low, almost always less than 0.8 and usually less than 0.1.

[40] In an additional set of experiments (section 2.6) using this subset of control parameters, data from each site (EP and AS) were individually assimilated, and if any of the control parameters did not meet the above criterion (uncertainties greater than 100%), they were removed from the subsets. In this manner, a single control parameter subset was selected for each model, which satisfies the above uncertainty and correlation criteria when both data sets are assimilated simultaneously and when the data sets are assimilated individually. This same parameter subset was used for experiments 2–4 (section 2.6). This method of parameter identification is not perfect, as the inverse Hessian method of estimating parameter sensitivity is local in parameter space. Thus it is possible that a parameter judged to be insensitive when all parameters are optimized might become sensitive when other parameters are held fixed; however, in our experience this was not the case. A large number of additional experiments were performed with the simplest models, in which the number and choice of optimized parameters, as well as the values associated with the fixed parameters, were altered. These additional experiments demonstrated that these changes always increased the magnitude of the cost function. With the large number of parameters associated with the more complex models, it was not possible to completely investigate the relevant parameter space with these models.

[41] For each of the 12 models used in this analysis, this process resulted in the selection of two to four key biogeochemical parameters. These control parameter subsets almost always included a parameter relating to the maximum growth rate of phytoplankton and the rate of remineralization. The remaining control parameters varied among models. A detailed analysis of the differences among the estimated parameter values obtained for the different models is beyond the scope of this study.

2.5.4. Penalty Terms

[42] Experiments were also conducted to examine the effect of imposing upper and lower bounds on the range of allowed parameter values as penalty terms in the cost function equation. In these experiments, penalties resulted in the optimized parameter values being either unaffected by the restriction (if the weight given to that penalty was low) or driven to the maximum or minimum value of that range (if the weight given to that penalty was high). Penalty terms were therefore not included in the cost function. Parameter values inconsistent with the ranges of observed estimates were not obtained in any of our final model runs, which we attribute to reliable first guesses (initial parameter values) obtained from the literature (typically acquired from papers arising from U. S. JGOFS process studies in the equatorial Pacific and the Arabian Sea) and to the objective and systematic selection of the model control parameters. In essence, to get around the need to rather arbitrarily attribute upper and lower bounds to each parameter value identified in the literature, the method described above (section 2.5.3) is used to identify an uncorrelated subset of parameters to which the cost function is most sensitive. By reducing the size of the parameter set to be optimized, the need for upper and lower bounds and corresponding penalty terms is removed.

2.6. Four Comparison Experiments

[43] Four experiments were conducted with the 12 ecosystem models described above. Experiment 1 was conducted without any form of formal data assimilation or

parameter optimization. Rather, model parameters were specified a priori, primarily based on literature values and varying degrees of subjective optimization (see below). In the remaining three experiments, the variational adjoint method of data assimilation was applied to optimize formally and objectively each participating model.

[44] In experiment 2, the assimilation procedure was performed individually at each site (AS and EP). Two distinct sets of optimized parameter values were obtained and used to generate optimal simulations and cost functions for each site. The sum of the costs for these two simulations in experiment 2 is defined as the "individual cost." In experiment 3, both data sets (AS and EP) were assimilated simultaneously, and one optimal parameter set was obtained that provided the best fit for both locations and was then used to generate simulations for both sites. The sum of the two cost values in experiment 3 is referred to as the "simultaneous cost." By definition, the magnitude of the simultaneous cost is always greater than or equal to that of the individual cost. Experiment 4 is an extension of experiment 2 wherein the two sets of optimized parameter values obtained in experiment 2 were applied such that the parameter set generated via the assimilation of the EP data was used in the AS simulation, and the parameter set generated via the assimilation of the AS data was applied to the EP simulation. The resulting cost is referred to as the "cross-validation cost" as it is similar to a cross-validation experiment.

2.7. Mean Model

[45] Each of the four experiments described above was conducted for the 12 ecosystem models, as well as for a simple empirical approximation we refer to as the Mean Model. This model assumes a constant value corresponding to each data type such that the chlorophyll, nitrate, productivity, and export estimates from this model are equal to their respective observational means computed over all depth and time. Individual, simultaneous, and cross-validation costs were obtained for the Mean Model just as they were for the other 12 participating models. For the preassimilation (experiment 1) and simultaneous (experiment 3) cases, the mean of each data type was computed over depth, time, and location, and the associated costs were calculated by computing the differences between these mean values and the observations. For the Mean Model, equation (1) becomes:

$$J = \Psi \sum_{k=1}^{K} \frac{1}{\psi_k} \sum_{m=1}^{M} \frac{W_{km}^2}{N_{km}} \sum_{j=1}^{N_{km}} \left(\bar{\alpha}_m - \hat{a}_{jkm}\right)^2 \tag{4}$$

where $\bar{\alpha}_m$ represents the mean (over depth, time, and location) of all observations of type m:

$$\bar{\alpha}_m = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N_{km}} \sum_{j=1}^{N_{km}} \hat{a}_{jkm}$$

[46] For the individual optimization (experiment 2) and cross-validation (experiment 4) cases, each model was individually tuned to one site at a time (rather than simultaneously as in equation (4)) and the two resulting costs were summed, giving the cost function for the Mean Model as:

$$J = \Psi \sum_{k=1}^{K} \frac{1}{\psi_k} \sum_{m=1}^{M} \frac{W_{km}^2}{N_{km}} \sum_{j=1}^{N_{km}} \left(\bar{\alpha}_{km} - \hat{a}_{jkm}\right)^2 \tag{5}$$

where $\bar{\alpha}_{km}$ represents the mean (over depth and time, but not location) of all observations of type m at site k:

$$\bar{\alpha}_{km} = \frac{1}{N_{km}} \sum_{j=1}^{N_{km}} \hat{a}_{jkm}$$

In this way the Mean Model was applied in the same way as the other 12 ecosystem models, and its results are shown for comparison in the following section.

2.8. Portability Index

[47] Another characteristic upon which ecosystem model skill was assessed is portability. It is highly desirable for a model to be able to reproduce data in multiple oceanographic regimes without retuning the biogeochemical parameters for each new location. A model possessing this characteristic is defined here as being highly portable. As a measure of portability, a "Portability Index" (*PI*) is defined as a function of the simultaneous costs and cross-validation costs. A highly portable model would be one which produces simultaneous and cross-validation costs that are similar in magnitude, and thus we define:

$$PI = J_s / J_x$$

where J_s represents the simultaneous cost and J_x represents the cross-validation cost. Models are increasingly portable as *PI* approaches a value of 1.0.

3. Results

3.1. Experiment 1: No Optimization

[48] In this first experiment the 12 participating ecosystem models did not undergo any formal data assimilation or parameter optimization, but instead parameter values were set a priori primarily based on literature values and varying degrees of subjective optimization (see details in auxiliary material). This resulted in cost functions that generally decrease with increasing ecosystem complexity (Figure 1a). Specifically, models 1–5, each of which contain only one phytoplankton state variable (i.e., "single-P models"), had a mean value of $J = 92.4 \pm 15.9$, whereas the models containing more than one phytoplankton state variable ("multi-P models"; i.e., models 6–12) had a significantly lower mean normalized cost ($J = 51.5 \pm 4.9$). (Uncertainties on these numbers and others throughout this section are computed as one standard error.) Some of the models did well in one location, whereas others did better in the other location. For example, model 5 did well in the equatorial Pacific (EP) but not as well in the Arabian Sea (AS). On the contrary, model 2 did well in the AS but performed poorly in the EP. Model 12 did well in both locations and was also the only model that produced a cost function that was lower than that obtained by simply comparing the data to the

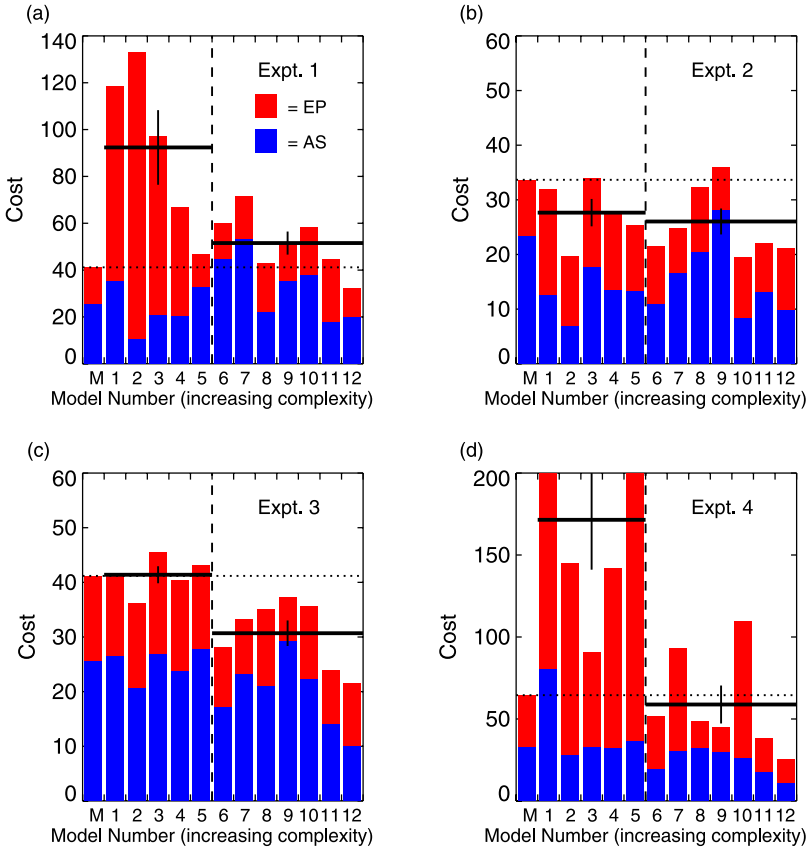

Figure 1. Cost function, J, as a function of model number. Vertical dashed line separates the single-P models (models 1–5) from the multi-P models (models 6–12). Red bars represent the equatorial Pacific (EP) component; blue bars represent the Arabian Sea (AS) component. Bars lower than the dotted horizontal line (cost of Mean Model "M") indicate that the model-data misfit is lower than that computed from the mean of the observations. Cost values are not significantly different if they vary by less than two. Two solid horizontal lines represent mean cost for the single-P and multi-P models, respectively; error bars illustrate one standard error. Note change in scales between panels: (a) experiment 1: no optimization, (b) experiment 2: individual optimization, (c) experiment 3: simultaneous optimization, and (d) experiment 4: cross-validation.

observational mean computed over all space and time (model "M" in Figure 1).

[49] The relative magnitudes of the four components of the cost function varied significantly between models in both the EP (Figure 2) and the AS (Figure 3). The magnitudes of these cost components were a function of the weights given to each data type, which were defined to be inversely proportional to the standard deviation of the data (equation (2)). To compare objectively the models' skill in reproducing the four data types, the ecosystem model cost components were compared to those produced by the Mean Model. In the equatorial Pacific without assimilation (Figure 2a), all but two ecosystem models produced model-data misfits that were smaller than those obtained with the Mean Model for both nitrate and export. The ecosystem models had much less success reproducing observed EP chlorophyll-a concentrations (chl). Without assimilation none of the models fit the chl data better than the Mean Model, with chl costs for certain models being more than 20 times greater than that obtained with the Mean Model. The models generally performed better for productivity, but still only five models produced model-data misfits that were lower than assuming a constant mean productivity value. In the Arabian Sea the situation was nearly opposite (Figure 3a): the ecosystem models produced nitrate and export costs that were generally higher than the Mean Model and chl and productivity costs that were typically lower than those obtained with the Mean Model.

[50] Although the implementation of each of the 12 models within an identical numerical and physical framework is a step toward an objective model comparison, the fact that some models were initially tuned to a greater degree than others must be taken into consideration. Without any data assimilation or parameter optimization, the 12 participating models clearly produced very different cost functions, with the more complex models generally yielding lower model-data misfits, particularly in the equatorial Pacific. However, certain models (e.g., models 11 and 12) were previously run in a global application and therefore already underwent a certain degree of prior tuning to ensure that they performed reasonably well in all ocean basins. As a result, these models perform particularly well in experiment 1. Similarly, model 1 was tuned for use in the Arabian Sea and therefore not surprisingly produced AS costs that

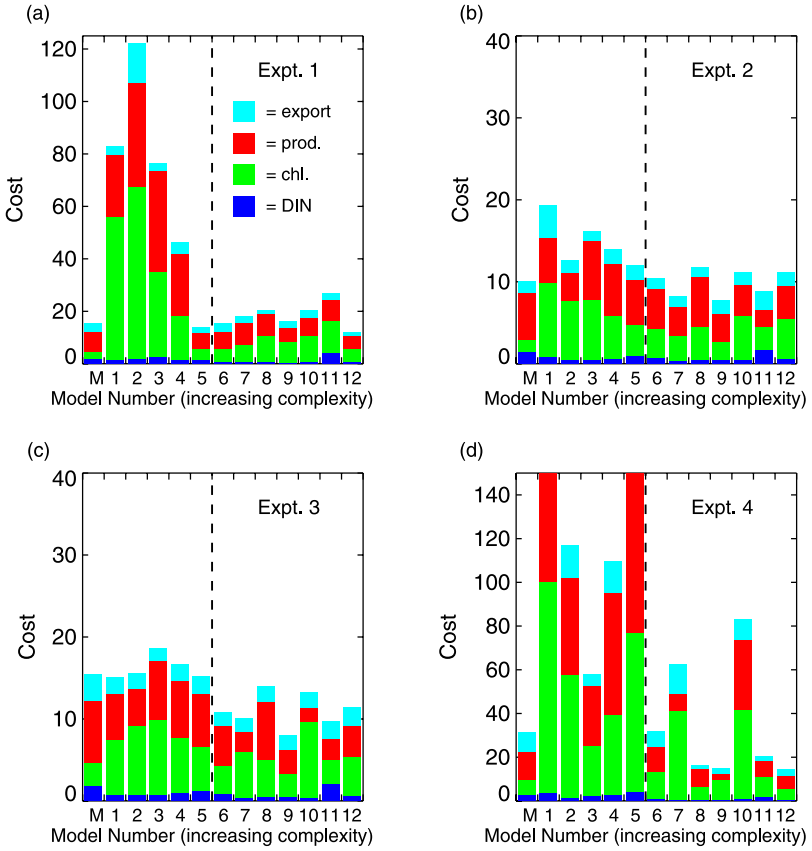

Figure 2. As in Figure 1, but for the equatorial Pacific (EP) component of the cost function, color-coded to show data type: cyan (export), red (productivity), green (chlorophyll), and blue (DIN). Total cost values are not significantly different if they vary by less than one. Note scales change between panels, but are consistent with those in Figure 3. (a) Experiment 1: no optimization, (b) experiment 2: individual optimization, (c) experiment 3: simultaneous optimization, and (d) experiment 4: cross-validation.

were less than EP costs, whereas model 6 was tuned for the EP, and produced EP costs that were less than AS costs.

[51] Thus from experiment 1, it is not clear whether certain models produce low model-data misfits because of their inherent model structure or simply because more effort was previously directed toward parameter tuning for application to a given region. To ensure that comparisons reflect real model differences and not simply differences in degrees of manual tuning, it is imperative that model comparisons such as experiments 2, 3, and 4 are performed in conjunction with a formal parameter optimization/data assimilation technique such as the variational adjoint method described above.

3.2. Experiment 2: Individual Optimization

[52] Site specific parameter optimizations using data from the EP and AS and the associated adjustment/optimization of two to four key biogeochemical parameters resulted in costs that were significantly (25%–85%) lower than those obtained prior to optimization (Figure 1b). The optimization significantly reduced the costs for all models. Whereas only one model produced costs lower than the Mean Model prior to the parameter optimization (Figure 1a), all but model 9 yielded combined costs lower than that of the Mean Model after assimilation (Figure 1b). Whereas the various model costs varied by as much as a factor of four prior to assimilation (Figure 1a, e.g., models 2 and 12), the results differed by less than a factor of two after assimilation (Figure 1b). Clearly, much of the performance variability between models with no parameter optimization (Figure 1a) was caused by differences in parameter tuning and not differences in model structure and complexity.

[53] Interestingly, the mean cost function of the single-P models ($J = 27.7 \pm 2.5$) was no longer significantly higher than that of the multi-P models ($J = 26.0 \pm 2.4$). In the EP, the multi-P models yielded slightly lower costs than the single-P models, but the opposite was true in the AS (Figure 1b). Again, the better performance of the more complex models in experiment 1 (Figure 1a) appeared to be a result of differential tuning and not due to any intrinsic advantage afforded by the more complex structure of these models.

[54] The implementation of the variational adjoint method resulted in a large reduction in the overall magnitude of both the EP and the AS costs. However, significant differences in model performance were observed in these two areas. In the EP, the Mean Model still produced costs that were lower than those obtained from most of the ecosystem models, whereas in the AS most models did as well or better than the Mean Model. The relative skill of the Mean Model in the EP results from the fact that in this region the data were more constant in time and depth (lower standard deviations; Table 1), compared to the AS where the amplitude of the

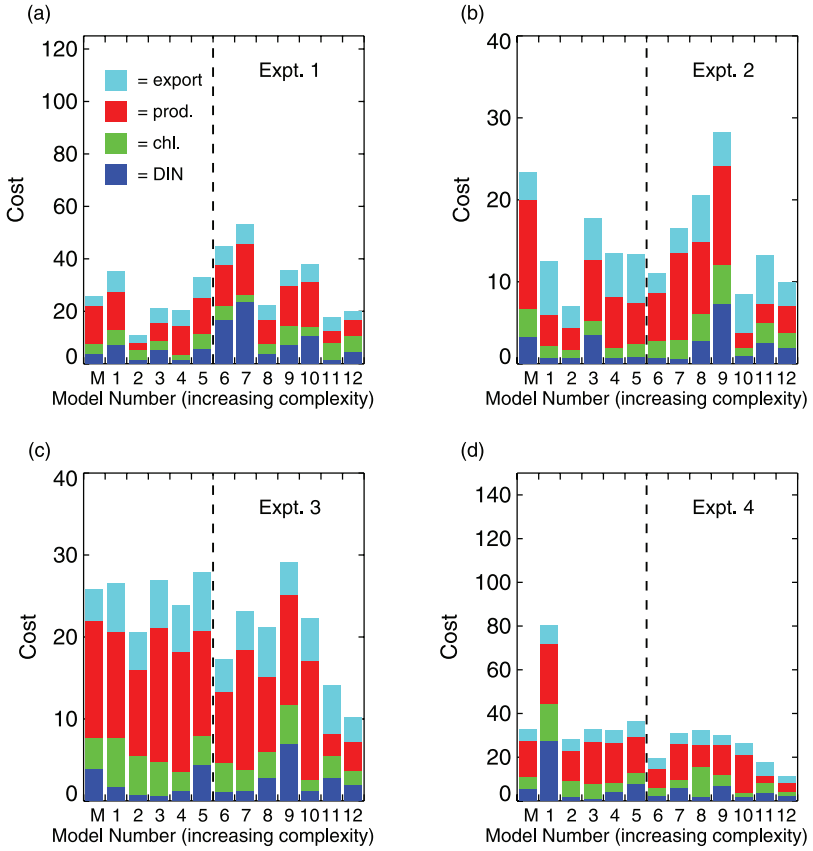

Figure 3. As in Figure 2, but for the Arabian Sea (AS) component of the cost function.

seasonal (monsoon) cycle is large and export showed episodic spikes associated with mesoscale upwelling.

[55] The parameter optimization dramatically reduced the cost for each of the four components of the cost function in the EP (Figures 2a and 2b; note change in scale). On average, the export, productivity, chl, and DIN costs were reduced by 35%, 49%, 55%, and 35%, respectively. The models with the highest preassimilation productivity and chl costs (models 1–4) underwent the most substantial cost reductions of 73–91% and 69–89% for productivity and chl costs, respectively (Figure 2b). After the optimization the EP export costs for all the models including the Mean Model were very similar. In terms of chl, the Mean Model produced lower costs than any of the participating models, whereas in terms of productivity, the participating models yielded costs that were generally at least as low as the Mean Model. The relative success of the Mean Model in terms of chl, rather than productivity, is largely a result of the small standard deviation (0.09 mg chl m^{-3}) of the chl measurements relative to their mean (0.26 mg chl m^{-3}). Similarly, the relatively high standard deviation of the productivity data (8.5 mmolC m^{-3} d^{-1}) as compared with their mean (10.2 mmolC m^{-3} d^{-1}) resulted in a poor productivity performance by the Mean Model (Table 1). Because of the strong variation in productivity both as a function of time and depth, the Mean Model generated a relatively large productivity cost.

[56] Within the AS test bed, the parameter optimization similarly reduced the export cost (on average by 18%), chl cost (50%), productivity cost (38%), and DIN cost (48%) for almost all models (Figures 3a and 3b). The one significant exception is model 11, for which the DIN and export costs increased by 63% and 12%, respectively (chl and productivity cost decreased 61% and 50%). In the AS (Figure 3b) all but one model (model 9) produced chl costs that were equal or lower than those generated by the Mean Model. The poorer performance of the Mean Model in terms of AS chl is due to its greater temporal/spatial variation of chl in this region compared to the EP.

[57] Although many of the models gave a similar low cost function value with a good fit to the data in experiment 2 (for example, models 2, 6, 10, 11, and 12 in Figure 1b), the concentrations of various model components varied significantly between models over the duration of the experiment. This is illustrated by examining time series of chl and DIN for four of these models, superimposed on the time series of the Ensemble Model (the average of all 12 participating models) ± one standard deviation (Figure 4). Model 12 produced large blooms of chl in the AS which were observed neither in the data set used herein, or in general in the AS where chl is maintained below 1.2 mg/m3 except in relatively isolated instances associated with mesoscale eddy activity [*Marra et al.*, 1998; *Dickey et al.*, 1998] which were not resolved in this study. Note, however, that Model 12 was also the only model that was able to produce the relatively low integrated chl observed around YD390 and YD580. Model 11 stands out from the other models by producing high nitrate concentrations in the intermonsoonal time periods when all other models produce much lower concentrations (Figure 4). In the EP after YD550, Model 11

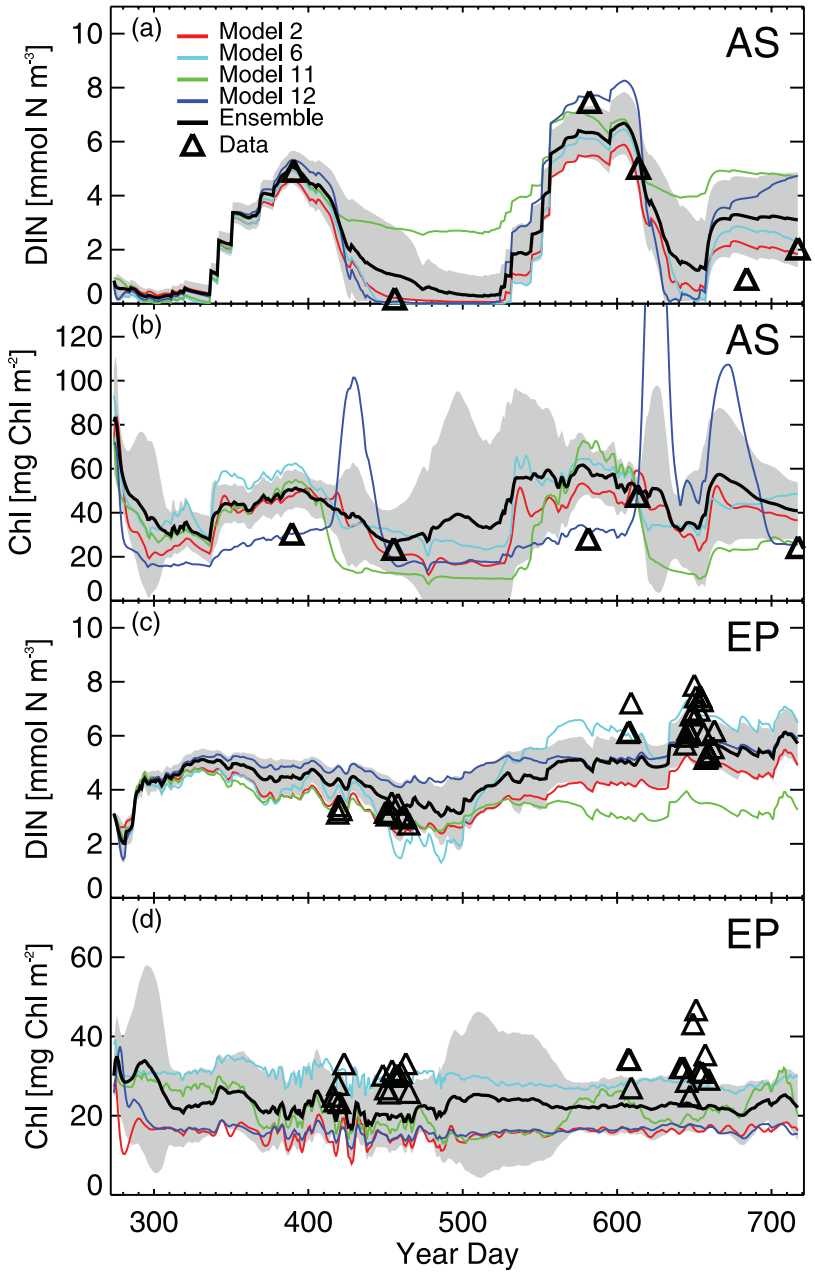

Figure 4. Simulations of (a,c) surface DIN and (b,d) integrated chlorophyll in the Arabian Sea (Figures 4a and 4b) and equatorial Pacific (Figures 4c and 4d) for models 2, 6, 11, and 12 plotted as a function of Year Day 1994 (AS) and Year Day 1991 (EP). Data (triangles) and the Ensemble Model time-series (mean of all 12 simulations) are shown for reference. Shaded area represents the Ensemble Model ± one standard deviation.

also produced much lower nitrate concentrations than any of the other models. Model 6 produced a much greater temporal variation of DIN in the EP, with relatively low concentrations (YD475) followed by quite high concentrations (after YD550). As expected, the models often converged at the points in time when data were available for assimilation (e.g., YD620 in the AS; Figure 4a) and diverged when no data were available (e.g., YD480-560; Figure 4b).

[58] All 12 participating models also differed significantly in terms of their depth-integrated productivity, depth-integrated chl, grazing rate, and export, averaged over the 1.25-year model run (Figure 5). For example, mean integrated productivity varied between 40 and 80 mmol C m^{-2} d^{-1} in the AS and 20 and 100 mmol C m^{-2} d^{-1} in the EP. Mean integrated chl ranged between 30 and 90 mg chl m^{-2} and 25 and 45 mg chl m^{-2} in the AS and EP models, respectively, with little correlation between the highest chl and productivity values. Grazing varied by a factor of four

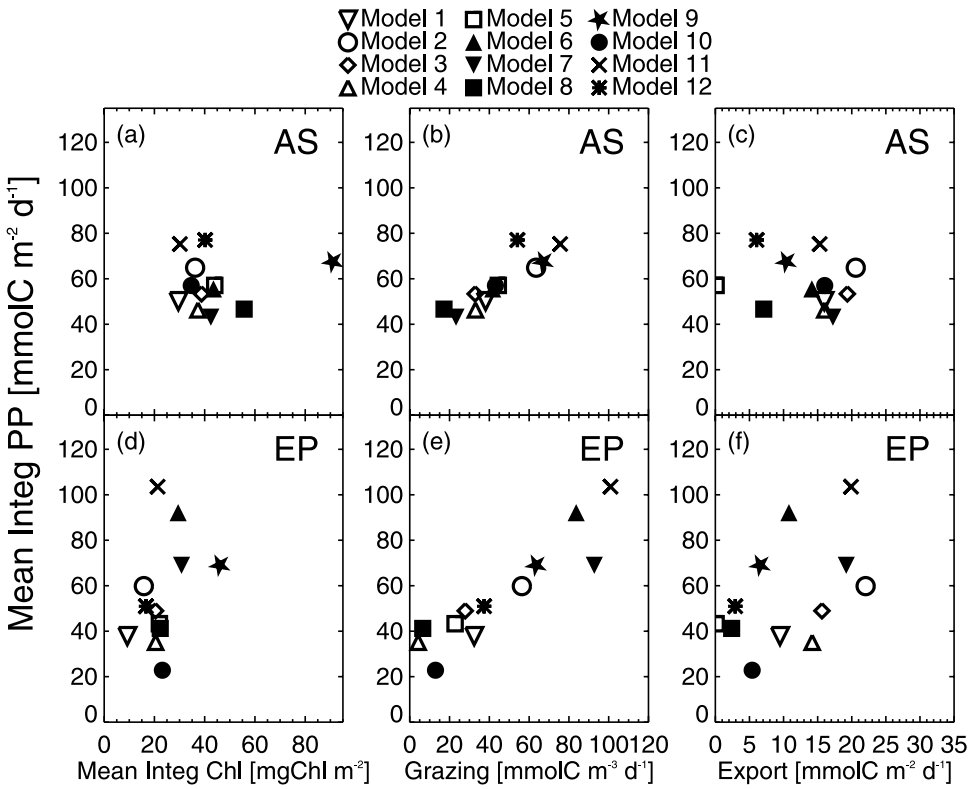

Figure 5. Mean integrated productivity (PP) computed over the 1.25 year run for (a–c) the Arabian Sea (AS) and (d–f) the equatorial Pacific (EP) as a function of mean integrated chlorophyll (Figures 5a and 5d), mean grazing (Figures 5b and 5e), and mean export at the bottom of the euphotic zone (Figures 5c and 5f). Experiment 2 results are shown for all 12 participating models; open symbols represent subset of single-P models.

among the AS models and by an order of magnitude among the EP models.

[59] In both test beds, models with the highest grazing rates were associated with the highest rates of productivity, which presumably indicates severe grazing pressure on phytoplankton at both sites. For these models (models 2, 9, 11, and 12 in the EP and models 2, 6, 7, 9, and 11 in the AS) the production/grazing loop is spinning faster and does not necessarily result in increased export (Figures 5c and f). Clearly, there are different ways in which the models are able to fit the data and thus more data, specifically data that can constrain ecosystem flows (e.g., community grazing rates), are required to determine which models are reproducing the data for the "right" reasons.

[60] Substantial differences between models exist even when mean values were computed only over the subset of model output that coincides in space and time with the available data. By comparing integrated productivity, integrated chl, export (at 800 or 880 m), and surface nitrate for the 12 models with the range (over different observation times) of analogous values computed directly from the data, it is possible to determine which models most closely reproduce the observations (Figure 6). Model 11 is the only model that produces a mean integrated productivity value for the AS that is within the range of values observed in this location (Figure 6a); however, this model produces AS export that is much lower than observed (Figure 6b). The remaining models all produce mean productivity values that are lower than any individual integrated productivity observation at this location. In the EP the range of observed integrated productivity is greater, but the models all still underestimate productivity, many by more than a factor of two.

[61] The fact that in both of these locations none of these models, with their different ecosystem structures and adjustable growth rates, are able to produce productivities as high as those observed suggests that these low productivities may result instead from a missing physical model component. For example, nutrient input due to mesoscale variability, which is prevalent in both of these regions, is undoubtedly underestimated here since these processes are not resolved in the three-dimensional circulation model used to generate mixed-layer depths and vertical velocities for the test bed framework. In addition, the one-dimensional test bed may not be resolving the increased productivity due to horizontal advection of highly productive water masses. Nonetheless, many of the models that underestimate productivity produce nitrate concentrations both at the surface (Figures 6c and 6f) and deeper in the mixed layer (not shown) that are in good agreement with the data. The cause of the systematic underestimation of productivity in these models, despite the rather high background value of K_v, is an active area of research. The solution may relate to the fact that data are from ^{14}C incubation experiments, whereas model productivities are typically computed from nitrate uptake rates using a constant C:N ratio. Investigation into

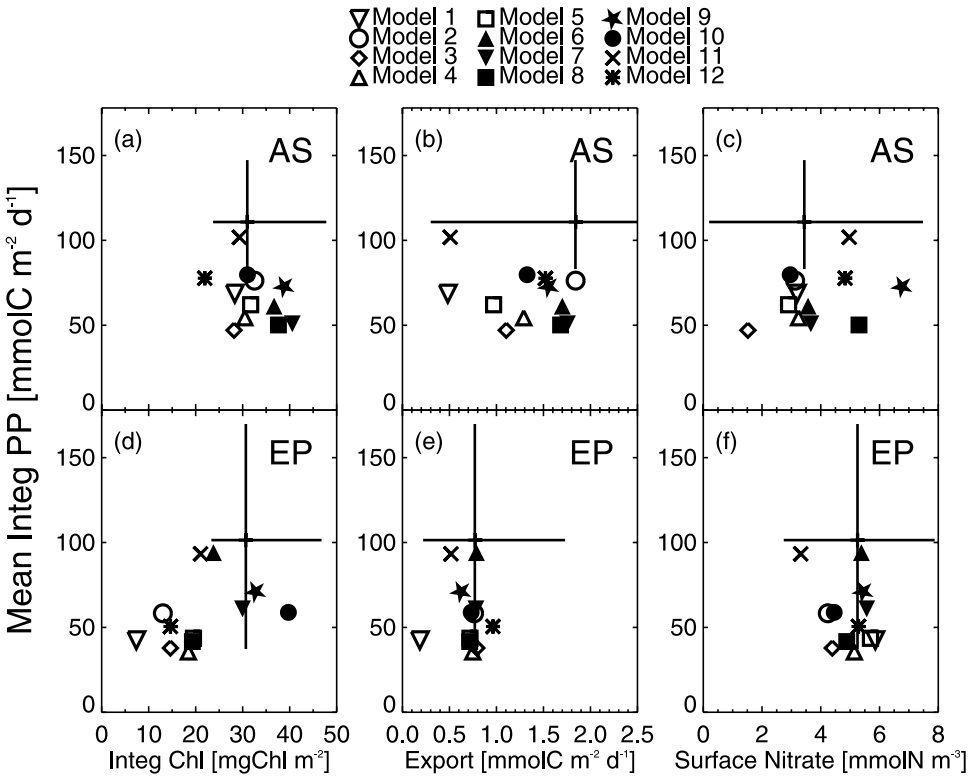

Figure 6. Mean integrated productivity (PP) for (a–c) the Arabian Sea (AS) and (d–f) the equatorial Pacific (EP), as a function of mean integrated chlorophyll (Figures 6a and 6d), mean export at 800 m (Figures 6c and 6f) and mean surface nitrate; all means are computed only over the model equivalents of the data. Solid black lines represent the ranges of the observations; observational means are illustrated as the intersection of each pair of solid black lines. Experiment 2 results are shown for all 12 participating models; open symbols represent subset of single-P models.

the explanation for the superior performance of model 11 in the AS and model 6 in the EP is also ongoing.

3.3. Experiment 3: Simultaneous Optimization

[62] Parameter optimization using data from both test bed sites tests the models' abilities to fit data from two very different environments with a single parameter set. If a model can successfully reproduce data collected in both the equatorial Pacific and the Arabian Sea without changing any biogeochemical parameters, it is more likely to have significant skill across the globe. Simultaneously fitting data from two such different sites using the same parameter set is a more challenging task for these models. Not surprisingly, the experiment 3 cost functions (Figure 1c) are significantly higher than those obtained from the individual optimization experiment (Figure 1b).

[63] With simultaneous optimization, the average cost function of the single-P models ($J = 41.4 \pm 1.5$) was no lower than that of the Mean Model ($J = 41.4$). In other words, for the single-P models, the sum of the squares of differences between each data point and its corresponding overall mean (including both locations over all depths and time) was typically not significantly greater than that computed between each data point and its model equivalent. The Model 2 AS simulation is a notable exception. Perhaps as a result of the high rates of organic matter cycling through its heterotrophic compartment, this model still produced relatively low model-data misfits in the AS.

[64] In this experiment the more complex (multi-P) models on average produce cost functions ($J = 30.7 \pm 2.3$) that were significantly lower than that of the single-P models ($J = 41.4 \pm 1.5$). By including more complex ecosystem dynamics with multiple pathways these multi-P models were better able to reproduce observations from two very different ecosystems.

[65] Dividing up the simultaneous cost in the Arabian Sea into its various components (Figure 3c) reveals that the magnitude of the chl and productivity costs for the simultaneous optimization experiments, relative to those of nitrate and export, are higher than they are for the corresponding costs for the individual optimization experiments (Figure 3b). In the equatorial Pacific, on the other hand, the relative contributions of each data type to the total cost were similar in the simultaneous and individual experiments. In the EP, the chl and productivity costs contributed roughly equally to the total cost, whereas in the AS the productivity simultaneous cost almost always exceeded the other components. Although the single-P models (models 1–5) were able to reliably reproduce chl and productivity data when only the AS observations are assimilated (Figure 3b), these models were much less successful (higher chl and productivity AS costs; Figure 3c) if the same parameters were used in both ocean basins. Interestingly,

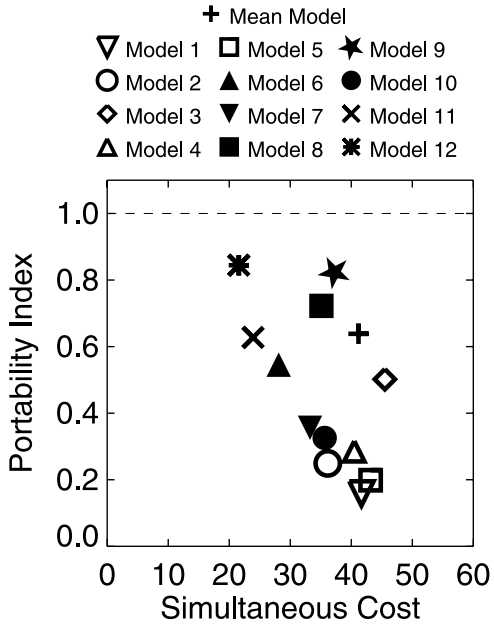

Figure 7. Portability Index (PI) as a function of simultaneous cost (J_S) for the 12 models participating in the intercomparison. The most portable models are closest to the PI = 1.0 (dashed) line. All the single-P models (open symbols) have either lower portability (low PI) or higher model-data misfit (high J_S) than all the multi-P models. The five most portable models all yield values of J_S that are lower than that of the Mean Model.

the models with three phytoplankton state variables (picoplankton, diatoms, and diazotrophs) and with only a single (or implicit) zooplankton compartment (models 11 and 12) were best able to reproduce productivity data simultaneously in the AS and the EP.

3.4. Experiment 4: Cross-Validation

[66] Although the cost function is a useful tool for examining model-data misfit for a particular data set, data assimilative models additionally need to be validated against independent, unassimilated data [*Friedrichs*, 2002; *Anderson*, 2005]. Thus in a final experiment the parameter values obtained via the assimilation of EP data in experiment 2 were used to generate an AS simulation, and the parameter values obtained via the assimilation of the AS data were used to generate an EP simulation. This is a very challenging test for any model, and one that can be used to assess the predictive ability of a set of models [*Friedrichs et al.*, 2006]. If a model tuned to one particular location cannot be used to reproduce data collected from another time or site, it is likely to be characterized by low predictive ability.

[67] The cross-validation costs (Figure 1d) vary substantially for the single-P models (J = 171 ± 30) and were always significantly greater than those obtained from the Mean Model (J = 64). When tuned for use in a particular location, e.g., the equatorial Pacific, these simple models were unable to reproduce data collected from a different location, e.g., the Arabian Sea, implying that the predictive ability of these models, i.e., the ability of these models to reproduce data from a time/location other than that to which it was tuned, may be very low.

[68] On average, the more complex multi-P models were better able to reproduce the unassimilated data (J = 59 ± 12) than were the single-P models (J = 171 ± 30); however, not all the multi-P models outperformed the single-P models (Figure 1d). For example, model 3 yielded lower cross-validation costs than models 7 and 10. The reasons for the relative success of models 6, 8, 9, 11, and 12 are areas of active investigation. Results for the two sites under investigation in this analysis (the Arabian Sea and the equatorial Pacific) suggest that when multiple phytoplankton compartments are included, a model tuned to one specific location is better able to reproduce data collected from another site, yet this is not sufficient. Other model characteristics such as the inclusion of DOM, ammonium, silicate, iron, variable C:chl ratios, and differing grazing formulations are undoubtedly playing important roles here as well.

[69] As was the case in the simultaneous optimization experiment (experiment 3; Figures 2c and 3c), all models in the cross-validation experiment (experiment 4; Figures 2d and 3d) reproduce the observed nitrate and export costs similarly well, regardless of model complexity, while the multi-P models tended to produce lower chl and productivity cross-validation costs. Specifically, the single-P models tuned for the AS did particularly poorly in the EP (Figure 2d). Several of the multi-P models (models 6, 8, 9, 11, and 12) produced low costs at both locations. That is, parameters optimized for the AS data worked well in the EP and parameters optimized for the EP data worked well in the AS for these multi-P models.

3.5. Portability Index

[70] It is advantageous for a model to be able to reproduce data in multiple different oceanographic regimes without the need for retuning the parameters for each new location. A model possessing this characteristic is defined here as being highly portable, and will have a Portability Index (*PI*; see section 2.8) that approaches a value of 1.0. Of course in addition to examining the value of *PI*, it is also imperative to assess how well models reproduce data from multiple regimes without parameter tuning. That is, the magnitude of J_s is also important, since a very portable model ($PI \sim 1$) unable to reproduce data from multiple locations (high J_s) would not be desirable. Thus a particularly useful format for assessing model skill is a plot of *PI* versus J_s (Figure 7). Using the high portability/low simultaneous cost criteria described above, the model with greatest skill would be closest to the point (0,1).

[71] Of the 12 participating models, five yield $PI > 0.5$: models 6, 8, 9, 11, and 12 (Figure 7). These models all have either two (models 6, 8, 9) or three (models 11, 12) phytoplankton state variables. In contrast, these models differ considerably in terms of the number of zooplankton state variables: model 11 has only an implicit zooplankton compartment and model 12 has one zooplankton state variable, whereas models 6 and 8 have two and model 9 has four. These results strongly suggest that the portability of this set of models is more highly correlated with the number of phytoplankton functional groups than the degree of zooplankton complexity within these models. Of the five

most portable models, models 6, 11, and 12 produce the lowest simultaneous costs.

4. Discussion

4.1. Parameter Optimization

[72] Data assimilation has been demonstrated to be a highly valuable tool in marine biogeochemistry [*Hofmann and Friedrichs*, 2001; *Spitz et al.*, 2001; *Doney et al.*, 2001]. One relatively novel application of data assimilation is the objective assessment of model skill [*Matear*, 1995; *Friedrichs et al.*, 2006]. Without using a formal parameter optimization scheme, it is not possible to determine whether disparities among multiple model simulations result from differences in model structure or from differences in tuning of the various unconstrained biogeochemical parameters. In this study, for example, experiments without parameter optimization suggest that the more complex models with multiple phytoplankton functional groups and multiple limiting nutrients fit the available data better than the simpler models (Figure 1a), yet the data assimilative experiments reveal that if a diverse group of models are all objectively and equally optimized, the simpler models produce least squares fits to the data from individual locations just as well as the more complex models (Figure 1b). That is, the models with only one phytoplankton state variable and one limiting nutrient are able to reproduce the data at each site individually (with different parameter values used at each location) just as well as the more complex models that include picoplankton, diatoms, diazotrophs, and multiple limiting nutrients. These conclusions have important ramifications for the regional and global application of marine biogeochemical models, suggesting that regional conditions may be represented by either simple or complex ecological models so long as the model is well-tuned, while only the more complex models may be suitable for global applications. This result could not have been reached without the application of an objective method of parameter optimization.

[73] A critical component of the application of any optimization scheme is the choice of parameters to optimize. Typically, a more complex model with a greater number of tunable parameters can be tuned to fit a given data set better than a simple model with fewer tunable parameters. However, an improvement in model-data fit is not necessarily associated with higher predictive ability. In fact, increasing the number of optimized parameters (especially partially correlated parameters) typically reduces the ability of a model to reproduce an independent data set [*Friedrichs et al.*, 2006]. Here this trend was tested and observed for the most complex model (model 12). In an initial assimilative run where 14 parameters were optimized, model 12 produced an individual cost of $J = 18.8$ and a cross-validation cost of $J = 103.8$. In a final run in which only three parameters were optimized (Figures 1b and 1d), model 12 yielded individual and cross-validation costs of $J = 21.2$ and 25.6, respectively. Reducing the number of optimized parameters slightly degraded the model-data fit for individual locations (the individual cost increased by 13%) but greatly improved the model's ability to reproduce an unassimilated independent data set (cross-validation cost decreased by 75%).

[74] Although the optimization of too many unconstrained and/or partially correlated parameters results in decreased predictive ability, optimization of only a subset of the model parameters may also have potential drawbacks. In this case, parameters that cannot be constrained by the available data are held fixed, and it is assumed that reasonable a priori values for these parameters exist. If in fact this is not the case, one could argue that the validity of the model is questionable. By optimizing only the subset of parameters that can be constrained by the available data, it is not possible to refute the potential existence of an alternate parameter set that might yield an even lower cost function, even if this optimal parameter set cannot be found with existing data and fitting methods. However, given the insensitivity of the magnitude of the cost function to the "fixed" parameter values, this is not likely to be a significant problem for the model comparison results described herein.

4.2. Assessment of Model Skill

[75] Model performance can be evaluated using a variety of different metrics. When a number of models are being compared and limited observations are available for direct model-data comparison, performance can be assessed by the agreement between each model simulation and the Ensemble model (Figure 4), which is typically defined to be the mean of the model simulations [*Carr et al.*, 2006; *Mikaloff Fletcher et al.*, 2006]. This metric highlights outlying models, but outliers do not necessarily have less skill. In fact, it is possible that the outlying models are the only ones that are able to reproduce a validation data set. For example, in this study one of the models that most often falls outside the Ensemble ± one standard deviation envelope (model 11; Figures 4a and 4c) is also the model that most closely reproduces the mean integrated productivity in both test bed sites.

[76] If a set of models is being compared after a formal data assimilation (parameter optimization) technique has been imposed as has been done here, the most straightforward method for comparing the resultant simulations is to assess the magnitude of the minimized (a posteriori) cost function. In this study, if data are assimilated from individual test bed locations (Figure 1b), the simple models can produce costs as low as those produced by the most complex models. If models are required to use a single set of parameters simultaneously at two sites, however, the more complex models yield lower costs (Figure 1c). In other words, the simpler models can only fit the data well if key biogeochemical parameters such as growth and remineralization rates are varied from site to site. On the other hand, models with multiple phytoplankton functional groups inherently contain multiple growth parameterizations that might be appropriate for different sites and thus retuning is not as necessary for this class of models.

[77] The magnitude of the a posteriori cost function is one metric for assessing model performance, which, unlike the Ensemble metric, is based on model-data misfit; however a notable caveat to this approach is that this assessment yields no information regarding the predictive ability of these models. A high-order polynomial with enough free parameters could be tuned to fit a given data set and generate a low a posteriori cost, but such a model would have no predictive ability. Although assessing predictive ability is not a straightforward task, cross-validation experiments,

where a portion of a data set is retained for validation while the remainder is used for assimilation/optimization, can be performed to glean some information regarding the predictive ability of a set of models [*Friedrichs et al.*, 2006]. In this study, cross-validation experiments are conducted by applying parameters optimized for the equatorial Pacific data in an Arabian Sea simulation and vice versa. Some insight into the predictive ability of these models can thus be attained: if a model tuned to one location cannot successfully reproduce data collected from another region or another time, it is likely to have very limited predictive ability.

[78] The cross-validation costs for the 12 models participating in this comparison exercise vary significantly. In general, however, models with multiple phytoplankton functional groups produce lower cost functions than do single phytoplankton compartment models. The five most successful of the complex models differ in many ways: constant versus variable C:chl ratios, inclusion of multiple limiting nutrients (iron and/or silicate in addition to nitrogen), number of zooplankton state variables (zero to four), and inclusion of ammonium and dissolved organic matter. Yet each of these models includes either two or three phytoplankton state variables, again suggesting that this is a critical attribute for globally applicable biogeochemical models. Four of these five models also include limiting nutrients in addition to nitrogen (model 6 includes iron; model 8 includes silicate; models 11 and 12 include iron and silicate), suggesting that nutrient colimitation may also be a key property of global marine biogeochemical models. The success of model 9 in this experiment is intriguing; this may be due to the greater degree of freedom with respect to zooplankton pathways, leading to reasonable costs despite the lack of multiple limiting nutrients. The advantage of including other model features (e.g., variable C:chl ratio, increased number of zooplankton boxes, ammonium, DOM) is currently under investigation in a more systematic study involving a subset of the models participating in this comparison. In addition, inclusion of a model with limitation by multiple nutrients and a single phytoplankton compartment will enable a comparison of the relative importance of multiple phytoplankton functional groups versus multiple limiting nutrients.

[79] Another method for assessing model performance introduced in this study is the cost comparison with the null or Mean Model. One would hope that a mechanistic marine ecosystem model would be able to produce a cost function that is lower than that obtained by assuming constant mean (in depth and time) values for each data type at each location; however, rather alarmingly, without imposing a formal parameter optimization technique only one of the models participating in this intercomparison achieves a lower cost than assuming these constant mean values (Figure 1a). Even after parameter optimization, the simpler models do not necessarily produce lower least squares misfits for both regions simultaneously (experiment 3) than are obtained by this simple null model (Figure 1c), whereas the more complex multiple phytoplankton and nutrient models produce costs that as a whole are 10–50% lower than this simple observational mean calculation. In the cross-validation experiments, again none of the simplest models produce costs lower than that of this null model. In other words, the null model in the Arabian Sea gives a better approximation to the data in the equatorial Pacific than do any of these simpler ecosystem models when they are tuned to the Arabian Sea data. This is not necessarily the case for the more complex models: some of these significantly outperform the Mean Model, although others do worse.

[80] Of course it is important for models to not only generate low least squares misfits to the available data but also to reproduce the variability in the data. As described above, the Mean Model fits the data relatively well in a least squares misfit comparison. However, by definition the Mean Model is unable to reproduce any data variability, and the temporal correlation is zero. Contemporary metrics of model skill consider the mean and variance of the simulation relative to the observations, as well as the correlation of the modeled and observed fields [*Taylor*, 2001]. We did not attempt such comparisons because the data available (covering only limited, sporadic time periods and influenced by mesoscale variability not necessarily resolved by the circulation model used to generate the physical forcing) are not well suited to them. However, it is important to note that there are different metrics available that weight the mean (or root-mean-square) model-data misfit, the relative variance, and the correlation differently [*Taylor*, 2001], and straight least squares is at one extreme of this spectrum and strongly biased in favor of our simple null model.

[81] An important goal of marine biogeochemical modeling is to identify which model structures are most portable, i.e., which models can be successfully used in a wide variety of ecosystems with the same set of fixed biogeochemical parameter values. Thus the Portability Index (*PI*) defined here provides another useful metric of model performance, by specifically measuring the degree to which model structures can be transported among different environments without retuning. Plotting *PI* versus simultaneous cost clearly highlights models that are both portable and can reproduce data from different ecosystem regimes without changing parameter values. The five models with greatest portability (models 6, 8, 9, 11, 12) all contain multiple phytoplankton state variables but differ greatly in terms of zooplankton complexity.

[82] The results discussed above, i.e., those from the a posteriori cost, cross-validation, and portability comparisons, are by definition sensitive to the choices made in constructing the cost function. Specifically, choices as to which types of data to incorporate, how to weight the different types of data, as well as which sites to include and how to weight them individually will influence a comparison exercise based on cost magnitude. For example, in the simultaneous optimization experiment (experiment 3), model 11 performs especially well in the Arabian Sea (Figure 3c) because the fit to the productivity data is particularly good. If the cost function had been defined such that more weight was given to export and little or no weight was given to productivity, model 11 might not demonstrate an advantage over the other models. However, it is equally likely that without the constraint of matching the productivity data, the export cost could be substantially reduced for model 11, and this model might still perform better than any of the single P compartment models. Although the intercomparison results would indeed be

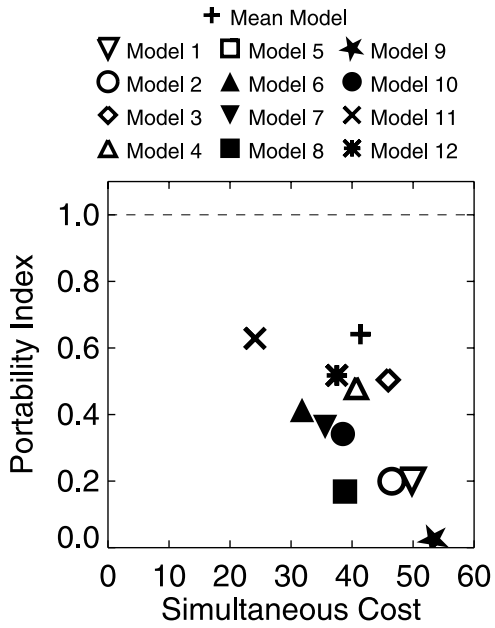

Figure 8. As in Figure 7, except for the case in which zooplankton data are included in the assimilation. In this case all the models except Model 11 have lower portability than the Mean Model.

different if a certain data type was omitted from or given a much lower weight in the assimilation process, there is no a priori reason why all data should not be utilized with the weights as defined here, being inversely proportional to the standard deviation of the data. The ramifications of including addition data in the cost function are further explored in the following section.

4.3. Additional Data for Assimilation

[83] Although in the individual optimization experiments many of the participating models yield cost functions with very similar magnitudes (Figure 1b), they are doing so using distinct model dynamics and are characterized by very different nutrient flow pathways. The simulated time series generated by these models tend to converge at the specific times when data are assimilated, but in between these times the models often produce very different results (Figure 4). This underscores the advantage of assimilating time series data with high temporal resolution (e.g., ocean color data) in addition to in situ cruise measurements [*Friedrichs*, 2001].

[84] Rates of productivity, grazing, and export also differ substantially between models, in certain instances by more than an order of magnitude (Figure 5), again illustrating that these models can produce similar least squares fits in terms of the aggregate cost function but achieve this in very different ways in terms of the individual cost components. Thus not only are more time series data required, but perhaps even more critical is the assimilation of in situ data from experiments that constrain these various model predictions of the ecosystem flows and dynamics. By working together in the experimental design process, modelers and observationalists can identify the observable quantities that best constrain the model dynamics and develop strategies for measuring them. However, the identity of these quantities may itself be model-dependent, and detailed analyses of the sensitivity of model skill to particular process parameterizations are needed to avoid experimental designs biased toward particular model structures.

[85] Within this study, one result that continually arises is that model performance appears to be more a function of the detail given to modeling phytoplankton complexity, rather than the detail given to modeling grazing, zooplankton biomass, and higher trophic level closure. For example, model 9, containing the greatest degree of zooplankton detail and the largest number (four) of zooplankton state variables, performs no better in the cross-validation experiment (Figure 1d) than model 11, which is located on the other end of the zooplankton detail spectrum in containing only an implicit zooplankton compartment. Model 6, with two size classes of zooplankton, also produces costs that are comparable with those of models 9 and 11. In the simultaneous cost experiment (experiment 3) the single-P models (models 1–5) produce costs ($J = 41.4 \pm 1.5$) that are significantly greater than those for the multi-P models ($J = 30.7 \pm 2.3$). On the contrary, the simplest models with regards to zooplankton complexity (models 1–5, 11, 12) produce costs ($J = 36.1 \pm 3.6$) that are not significantly different from the models with greater zooplankton complexity ($J = 33.9 \pm 1.6$).

[86] In order to test whether this result, insensitivity of model performance to zooplankton model complexity, is simply a result of the data types available for assimilation (e.g., DIN, chlorophyll-a, productivity, and export), the four experiments described above, including the parameter identification process, were repeated with 64–200 μm zooplankton data [*Roman et al.*, 2000] included in the cost function. (Note that the M^{-1} and N^{-1} normalization of the cost function (equations (1) and (3)) ensures that the magnitude of the costs with and without zooplankton data can be meaningfully compared.) An unavoidable caveat in this approach is the mismatch between what each model considers zooplankton (microzooplankton in many instances) and the data constraint. Models with a single zooplankton compartment are forced to match the size fractioned biomass (64–200 μm), which may be inconsistent with the model zooplankton definition. Models with multiple zooplankton compartments have somewhat more flexibility, as they are able to match a single zooplankton compartment to the 64–200 μm observations.

[87] In the simultaneous cost experiment where zooplankton data are assimilated, the single-P models again produce costs that are significantly greater than those for the multi-P models ($J = 50.6 \pm 5.1$ versus $J = 37.1 \pm 3.3$). As in the case in which zooplankton data are not assimilated, the simplest models with regards to zooplankton complexity still produce costs that are not significantly different from the models with greater zooplankton complexity ($J = 44.9 \pm 5.3$ vs. $J = 39.6 \pm 3.7$). In terms of portability, assimilating zooplankton biomass also did not improve and often degraded performance of the models with explicit zooplankton (Figure 8). Model 12, for example, produced $PI = 0.84$ when zooplankton data were not assimilated but only $PI = 0.52$ when zooplankton data were assimilated. As expected, the effect of assimilating zooplankton data was small for model 11, where zooplankton biomass is implicit and could

not feed back onto phytoplankton dynamics. The assimilation process merely scaled the loss rate constant determining the inferred biomass.

[88] It would be logical to assume that when zooplankton data are assimilated, the models containing the most realistic zooplankton representations would demonstrate the greatest improvement in model skill. However, in the simultaneous optimization experiments described above, a single model must reproduce data in two separate ocean basins so that increased portability only arises from increased model complexity if the model accurately represents the underlying mechanisms, and the mechanisms are in fact uniform across basins. The results from the experiments described above suggest that may not be the case in the current spectrum of models: assimilation of zooplankton data into models with greater zooplankton model complexity reduces the skill and portability of these models. One possibility is that the limited data constraints allow too much dynamical freedom within the models [*Armstrong*, 1999]. Alternatively, this could suggest that the underlying zooplankton behavior is misrepresented in these models. Clearly, more careful approaches are warranted in the assimilation of zooplankton data, and care must be taken to match the observations to the appropriate fraction of simulated zooplankton biomass.

[89] For example, in the equatorial Pacific microzooplankton occur prominently [*Calbet and Landry*, 2004], yet observations of this functional group are infrequent in part because there is no simple, widely used method of measuring microzooplankton biomass. Moreover, while observations of mesozooplankton (typically size partitioned biomass) are more prevalent, these species exhibit behavioral attributes [e.g., *Paffenhöfer*, 1998] that make them ill-suited for realistic inclusion within the continuum-based ecosystem models featured herein. Specifically, mesozooplankton are more difficult to model mechanistically than phytoplankton because of spatial-temporal scale mismatch, since in part their biomass fluctuations derive from vertical migration and nonlocal reproduction, which may depend on seasonal environmental cues [*Ashjian et al.*, 2002; *Idrisi et al.*, 2004]. Thus while more appropriate observational characterization of the grazer community's size-partitioned variability may be available for the Arabian Sea, the shortcomings in how mesozooplankton behavior is represented in our models are likely to overshadow this benefit.

[90] The observations assimilated here (wet weight of organisms in the 64–200 μm size class) represent a subset of the mesozooplankton that may not be a good match for any of the models, some of which aggregate all zooplankton into a single compartment, some of which distinguish between microzooplankton and mesozooplankton and some of which have multiple zooplankton compartments. Finally, the success of the models characterized by greater phytoplankton complexity may result from the simultaneous optimization of biomass (chlorophyll-a concentration) and rate (productivity) data. It is possible that if both zooplankton biomass and zooplankton grazing rate data were available and simultaneously assimilated, the models with greater zooplankton complexity would demonstrate greater model skill.

[91] Another question is, How would these results differ if additional nutrient data (iron, silicate, phosphate) were assimilated as well? Would these additional data improve the performance of the more complex models that contain multiple limiting nutrients, relative to the simpler models? Marine biogeochemical modeling remains challenged by limited data availability and in the long run the additional observations will improve model skill. However, in the short term, as increasing amounts of data become available, models will inevitably have more difficulty simultaneously producing low model-data misfits for all of the different data types.

[92] Whether or not the availability and assimilation of additional nutrient data would improve the performance of the more complex models relative to the simpler models depends on how performance is measured. Even if the cost function was normalized to the number of observations assimilated, model skill might not improve. As can be seen from the example of zooplankton observations, models can be forced to fit additional fields at the expense of deteriorating fidelity with some other field unless the additional data are entirely consistent with the underlying model dynamics. Gains in skill and portability from additional data therefore depend on understanding the underlying biological mechanisms and forcing the biogeochemical models with realistic physics. As the goal of this study was to be broadly comparative, assimilating data types common to all models participating in the intercomparison was chosen as the most broadly applicable approach.

5. Summary

[93] Twelve ecosystem models were run in an identical physical and numerical framework at two sites with distinct ecosystem dynamics: the Arabian Sea and the equatorial Pacific. Each model was optimized using the same variational adjoint technique. The resulting simulations were compared using a number of different metrics of performance. The simpler models with single plankton compartments were able to reproduce data within these specific ecosystem regimes (the equatorial Pacific and Arabian Sea) as well as the more complex models. However, models with greater phytoplankton complexity produce lower least squares misfits to data collected from both the equatorial Pacific and Arabian Sea, when parameter values are not allowed to vary between sites. Specifically, these models simultaneously produce better fits to the observed chlorophyll-a concentrations and rates of primary production at both locations. In addition, cross-validation experiments demonstrated that as long as only a few key biogeochemical parameters are optimized, models with greater phytoplankton complexity demonstrate enhanced portability, i.e., they are better able to reproduce data in multiple different oceanographic regimes without the need for retuning biogeochemical parameters for each new location. This result stems from the fact that these models inherently contain multiple growth parameterizations that are appropriate for different ecosystem states. The fact that models including more complex phytoplankton dynamics with multiple pathways are better able to reproduce observations from two very different ecosystems simultaneously may be intuitively expected, but this is the first time that this concept has been rigorously and objectively demonstrated. This study has also demonstrated that the optimization of too many uncon-

strained and/or partially correlated parameters may result in decreased predictive ability as measured by the success of the cross-validation experiments. Thus we see that there are both costs (possible degradation in predictive ability if too many parameters are optimized) and benefits (greater portability) to increasing ecosystem model complexity.

[94] The results of this model intercomparison effort demonstrate that multiple phytoplankton functional groups improve the ability of the models to simulate simultaneously the Arabian Sea and equatorial Pacific ecosystems, suggesting that this may be a critical characteristic of a globally applicable biogeochemical model. However, the additional dynamical freedom afforded to the models with greater zooplankton complexity does not in this case provide them with a tangible advantage. This is likely because the zooplankton data are an inappropriate constraint for the modeled zooplankton and/or because of an underlying misrepresentation of zooplankton dynamics in the models. In short, zooplankton in nature are particularly ill-suited for ecosystem models that implicitly assume that a continuum representation is adequate. For example, none of the models participating in this exercise include details of zooplankton life cycles, such as reproduction and diapause, or vertical migration.

[95] Although a number of the models participating in this comparison exercise were able to reproduce the data similarly well, they did so via very different element flow pathways. Results of this study specifically suggest that additional data such as zooplankton community grazing rates, phytoplankton speciation characterizations, and phytoplankton community nutrient uptake rates might be particularly helpful in further constraining these models. As ocean observing systems become more commonplace in the coming decades, the volume of marine biogeochemical data available for assimilation is expected to increase dramatically. However, these observations are likely to be of bulk concentrations (e.g., chlorophyll-a, nutrients) and will not be able to fully constrain the available models. This study truly highlights the need to conduct observational studies with experimental designs motivated by model deficiencies and differences, the need to identify the observable rates that can best constrain the models and differentiate among various optimal solutions, and the need to develop strategies for measuring these specific quantities. These goals can only be accomplished through the combined and concurrent efforts of modelers and observationalists; the importance of such a synergy between those running the models and those collecting and interpreting the data cannot be overstated.

[96] **Acknowledgments.** This research was supported by the U.S. National Science Foundation through the JGOFS Synthesis and Modeling Project (OCE-0097285) and the National Aeronautics and Space Agency (NAG5-11259 and NNG05GO04G), as well as numerous other grants to the various investigators who participated. Computer facilities and support were provided by the Commonwealth Center for Coastal Physical Oceanography at Old Dominion University. Our sincerest thanks go to Raghu Murtugudde for providing the 3-D model output fields used to force our 1-D models. We are especially grateful for the hard work of the many JGOFS scientists who helped acquire the excellent equatorial Pacific and Arabian Sea biogeochemical data sets used in our analysis. This is VIMS contribution 2795.

References

Anderson, T. R. (2005), Plankton functional type modeling: running before we can walk?, *J. Plankton Res.*, *27*(11), 1073–1081, doi:10.1093/plankt/fbi076.

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

Armstrong, R. A. (1999), Stable model structures for representing biogeochemical diversity and size spectra in plankton communities, *J. Plankton Res.*, *21*, 445–464.

Ashjian, C. J., S. L. Smith, C. N. Flagg, and N. Idrisi (2002), Distribution, annual cycle, and vertical migration of acoustically derived biomass in the Arabian Sea during 1994–1995, *Deep Sea Res. II*, *49*, 2377–2402.

Barber, R. T., S. T. Lindley, M. Sanderson, F. Chai, J. Newton, C. C. Trees, D. G. Foley, and F. Chavez (1996), Primary production in the equatorial Pacific during 1992, *Deep Sea Res. II*, *42*(4–6), 933–969.

Barber, R. T., J. Marra, R. C. Bidigare, L. A. Codispoti, D. Halpern, Z. Johnson, M. Latasa, R. Goericke, and S. L. Smith (2001), Primary productivity and its regulation in the Arabian Sea during 1995, *Deep Sea Res. II*, *48*, 1127–1172.

Besiktepe, S. T., P. F. J. Lermusiaux, and A. R. Robinson (2003), Coupled physical and biogeochemical data-driven simulations of Massachusetts Bay in late summer: real-time and postcruise data assimilation, *J. Mar. Syst.*, *40–41*, 171–212.

Bissett, W. P., J. J. Walsh, D. A. Dieterle, and K. L. Carder (1999), Carbon cycling in the upper waters of the Sargasso Sea: I. Numerical simulation of differential carbon and nitrogen fluxes, *Deep Sea Res. I*, *46*, 205–269.

Calbet, A., and M. R. Landry (2004), Phytoplankton growth, microzooplankton grazing, and carbon cycling in marine systems, *Limnol. Oceanogr.*, *49*(1), 51–57.

Carr, M.-E., et al. (2006), A comparison of global estimates of marine primary production from ocean color, *Deep Sea Res. II*, *53*, 741–770.

Chai, F., R. C. Dugdale, T.-H. Peng, F. P. Wilkerson, and R. T. Barber (2002), One-dimensional ecosystem model of the equatorial Pacific upwelling system. part I: Model development and silicon and nitrogen cycle, *Deep Sea Res. II*, *49*, 2713–2745.

Chen, D., A. Busalacchi, and L. Rothstein (1994), The roles of vertical mixing, solar radiation, and wind stress in a model simulation of the sea surface temperature seasonally cycle in the tropical Pacific Ocean, *J. Geophys. Res.*, *99*, 20,345–20,359.

Christian, J. R., M. A. Verschell, R. Murtugudde, A. J. Busalacchi, and C. R. McClain (2002), Biogeochemical modelling of the tropical Pacific Ocean, I: Seasonal and interannual variability, *Deep Sea Res. II*, *49*, 509–543.

Coles, V. J., R. R. Hood, M. Pascual, and D. G. Capone (2004), Modeling the effects of *Trichodesmium* and nitrogen fixation in the Atlantic Ocean, *J. Geophys. Res.*, *109*, C06007, doi:10.1029/2002JC001754.

Denman, K. L. (2003), Modelling planktonic ecosystems: Parameterizing complexity, *Prog. Oceanogr.*, *57*, 429–452.

Denman, K. L., and M. A. Pena (1999), A coupled 1-D biological/physical model of the northeast subarctic Pacific Ocean with iron limitation, *Deep Sea Res. II*, *46*, 2877–2908.

Denman, K. L., and M. A. Pena (2002), The response of two coupled one-dimensional mixed layer/planktonic ecosystem models to climate change in the NE subarctic Pacific Ocean, *Deep Sea Res. II*, *49*, 5739–5757.

Dickey, T., J. Marra, D. E. Sigurdson, R. A. Weller, C. S. Kinkade, S. E. Zedler, J. D. Wiggert, and C. Langdon (1998), Seasonal variability of bio-optical and physical properties in the Arabian Sea: October 1994–October 1995, *Deep Sea Res. II*, *45*, 2001–2025.

Doney, S. C., I. Lima, K. Lindsay, J. K. Moore, S. Dutkiewicz, M. A. M. Friedrichs, and R. Matear (2001), Marine biogeochemical modeling, *Oceanography*, *14*, 93–107.

Doney, S. C., K. Lindsay, I. Fung, and J. John (2006), Natural variability in a stable 1000 year coupled climate-carbon cycle simulation, *J. Clim.*, *19*(13), 3033–3054.

Dugdale, R. C., R. T. Barber, F. Chai, T.-.H. Peng, and F. P. Wilkerson (2002), One-dimensional ecosystem model of the equatorial Pacific upwelling system. part II: sensitivity analysis and comparison with JGOFS EqPac data, *Deep Sea Res. II*, *49*, 2747–2768.

Dunne, J. P., R. A. Armstrong, A. Gnanadesikan, and J. L. Sarmiento (2005a), Marine ecology, biogeochemistry and atmospheric CO2 signature from a 43-year reanalysis in a global ice/ocean biogeochemical general circulation model, poster and extended abstract presented at the Seventh International Carbon Dioxide Conference, 25–30 September 2005, sponsor?, Boulder, Colo.

Dunne, J. P., R. A. Armstrong, A. Gnanadesikan, and J. L. Sarmiento (2005b), Empirical and mechanistic models for particle export ratio, *Global Biogeochem. Cycles*, *18*, GB4026, doi:10.1029/2004GB002390.

Evans, G. T. (1999), The role of local models and data sets in the Joint Global Ocean Flux study, *Deep Sea Res. I*, *46*, 1369–1389.

Evans, G. T. (2003), Defining misfit between biogeochemical models and data sets, *J. Mar. Syst.*, *40–41*, 49–54.

Flynn, K. J. (2005), Castles built on sand: dysfunctionality in plankton models and the inadequacy of dialogue between biologists and modelers, *J. Plankton Res.*, *27*(12), 1205–1210.

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

Friedrichs, M. A. M. (2001), A data-assimilative marine ecosystem model of the central equatorial Pacific: numerical twin experiments, *J. Mar. Res.*, *59*, 859–894.

Friedrichs, M. A. M. (2002), The assimilation of SeaWiFS and JGOFS EqPac data into a marine ecosystem model of the central equatorial Pacific, *Deep Sea Res. II*, *49*, 289–319.

Friedrichs, M. A. M., and E. E. Hofmann (2001), Physical control of biological processes in the central equatorial Pacific, *Deep Sea Res. I*, *48*, 1023–1069.

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

Fujii, M., Y. Nojiri, Y. Yamanaka, and M. J. Kishi (2002), A one-dimensional ecosystem model applied to time series station KNOT, *Deep Sea Res. II*, *49*, 5441–5461.

Fujii, M., Y. Yamanaka, Y. Nojiri, M. J. Kishi, and F. Chai (2007), Comparison of seasonal characteristics in biogeochemistry among the subarctic North Pacific stations described with a NEMURO-based marine ecosystem model, *Ecol. Modell.*, *202*, 52–67, doi:10.1016/j.ecolmodel.2006.02.046.

Garside, C., and J. C. Garside (1995), Euphotic-zone nutrient algorithms for the NABE and EqPac study sites, *Deep Sea Res. II*, *42*, 335–347.

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

Giering, R., and T. Kaminski (1998), Recipes for adjoint code construction, *ACM Trans. Math. Software*, *24*(4), 437–474.

Gilbert, J. C., and C. Lemarechal (1989), Some numerical experiments with variable-storage quasi-newton algorithms, *Math. Progr.*, *45*, 405–435.

Gregg, W. W., P. Ginoux, P. S. Schopf, and N. W. Casey (2003), Phytoplankton and iron: Validation of a global three-dimensional ocean biogeochemical model, *Deep Sea Res. II*, *50*, 3143–3169.

Hofmann, E. E., and M. A. M. Friedrichs (2001), Biogeochemical data assimilation, in *Encyclopedia of Ocean Sciences*, vol. 1, edited by J. H. Steele, S. A. Thorpe, and K. K. Turekian, pp. 302-308, Academic, London.

Honjo, S., J. Dymond, R. Collier, and S. J. Manganini (1995), Export production of particles to the interior of the equatorial Pacific Ocean during the 1992 EqPac experiment, *Deep Sea Res. II*, *42*, 831–870.

Honjo, S., J. Dymond, W. Prell, and V. Ittekkot (1999), Monsoon-controlled export fluxes to the interior of the Arabian Sea, *Deep Sea Res.*, *46*, 1859–1902.

Hood, R. R., N. R. Bates, D. G. Capone, and D. B. Olson (2001), Modeling the effect of nitrogen fixation on carbon and nitrogen fluxes at BATS, *Deep Sea Res. II*, *48*, 1609–1648.

Hood, R. R., K. E. Kohler, J. P. McCreary, and S. L. Smith (2003), A four-dimensional validation of a coupled physical-biological model of the Arabian Sea, *Deep Sea Res. II*, *50*, 2917–2945.

Hood, R. R., V. J. Coles, and D. G. Capone (2004), Modeling the distribution of Trichodesmium and nitrogen fixation in the Atlantic Ocean, *J. Geophys. Res.*, *109*, C06006, doi:10.1029/2002JC001753.

Hood, R., et al. (2006), Functional group modeling: progress, challenges and prospects, *Deep Sea Res. II*, *53*, 459–512.

Hundsdorfer, W., and R. A. Trompert (1994), Method of lines and direct discretization: a comparison for linear advection, *Appl. Numer. Math.*, *13*(6), 469–490.

Idrisi, N., M. J. Olascoaga, Z. Garraffo, D. B. Olson, and S. L. Smith (2004), Mechanisms for emergence from diapause of *Calanoides carinatus* in the Somali Current, *Limnol. Oceanogr.*, *49*, 1262–1268.

Jiang, M.-S., F. Chai, R. C. Dugdale, F. P. Wilkerson, T.-H. Peng, and R. T. Barber (2003), A nitrate and silicate budget in the equatorial Pacific Ocean: A coupled physical-biological model study, *Deep Sea Res. II*, *50*, 2971–2996.

Kantha, L. H. (2004), A general ecosystem model for applications to studies of carbon cycling and primary productivity in the global oceans, *Ocean Modell.*, *6*, 285–334.

Kinkade, C. S., J. Marra, T. D. Dickey, C. Langdon, D. E. Sigurdson, and R. Weller (1999), Diel bio-optical variability observed from moored sensors in the Arabian Sea, *Deep Sea Res. II*, *46*, 1813–1831.

Kishi, M. J., et al. (2007), NEMURO-Introduction to a lower trophic level model for the North Pacific marine ecosystem, *Ecol. Model.*, *202*, 12–25, doi:10.1016/j.ecolmodel.2006.08.021.

Lancelot, C., Y. H. Spitz, N. Gypens, K. Ruddick, S. Becquevort, V. Rousseau, and G. Billen (2005), Modelling diatom-Phaeocystis blooms and nutrient cycles in the Southern Bight of the North Sea: the MIRO model, *Mar. Ecol. Prog. Ser.*, *289*, 63–78.

Laws, E. A., P. G. Falkowski, W. O. Smith, H. Ducklow, and J. J. McCarthy (2000), Temperature effects on export production in the open ocean, *Global Biogeochem. Cycles*, *14*, 1231–1246.

Lawson, L. M., Y. H. Spitz, E. E. Hofmann, and R. B. Long (1995), A data assimilation technique applied to a predator-prey model, *Bull. Math. Biology*, *57*(4), 593–617.

Le Quere, C., et al. (2005), Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models, *Global Change Biol.*, *11*, 2016–2040.

Marra, J., T. D. Dickey, C. Ho, C. S. Kinkade, D. E. Sigurdson, R. A. Weller, and R. T. Barber (1998), Variability in primary production as observed from moored sensors in the central Arabian Sea in 1995, *Deep Sea Res. II*, *45*, 2253–2267.

Martin, J. H., G. A. Knauer, D. M. Karl, and W. W. Broenkow (1987), Vertex: carbon cycling in the northeast Pacific, *Deep Sea Res.*, *34*, 267–285.

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.

McCreary, J. P., K. E. Kohler, R. R. Hood, S. Smith, J. Kindle, A. S. Fischer, and R. A. Weller (2001), Influences of diurnal and intraseasonal forcing on mixed-layer and biological variability in the central Arabian Sea, *J. Geophys. Res.*, *106*(C4), 7139–7155.

McGillicuddy, D. J., Jr., and A. R. Robinson (1997), Eddy-induced nutrient supply and new production in the Sargasso Sea, *Deep Sea Res. I*, *44*(8), 1427–1450.

McPhaden, M. J., et al. (1998), The Tropical Ocean-Global Atmosphere observing system: A decade of progress, *J. Geophys. Res.*, *103*, 14,169–14,240.

Mikaloff Fletcher, S. E., et al. (2006), Inverse estimates of anthropogenic CO2 uptake, transport, and storage by the ocean, *Global Biogeochem. Cycles*, *20*, GB2002, doi:10.1029/2005GB002530.

Moore, J. K., S. C. Doney, D. M. Glover, and I. Y. Fung (2002), Iron cycling and nutrient-limitation patterns in surface waters of the World Ocean, *Deep Sea Res. II*, *49*(1–3), 463–507.

Moore, J. K., S. C. Doney, and K. Lindsay (2004), Upper ocean ecosystem dynamics and iron cycling in a global three-dimensional model, *Global Biogeochem. Cycles*, *18*, GB4028, doi:10.1029/2004GB002220.

Morrison, J. M., L. A. Codispoti, S. Gaurin, B. Jones, V. Manghnani, and Z. Zheng (1998), Seasonal variation of hydrographic and nutrient fields during the US JGOFS Arabian Sea Process Study, *Deep Sea Res. II*, *45*, 2053–2101.

Murray, J. W., E. Johnson, and C. Garside (1995), A US JGOFS process study in the equatorial Pacific (EqPac): Introduction, *Deep Sea Res II*, *42*(2–3), 275–293.

Murtugudde, R., and A. J. Busalacchi (1999), Interannual variability of the dynamics and thermodynamics of the tropical Indian Ocean, *J. Clim.*, *12*, 2300–2326.

Murtugudde, R., R. Seager, and A. Busalacchi (1996), Simulation of the tropical oceans with an ocean GCM coupled to an atmospheric mixed-layer model, *J. Clim.*, *9*, 1795–1815.

Pacanowski, R. C., and S. G. H. Philander (1981), Parameterization of vertical mixing in numerical models of tropical oceans, *J. Phys. Oceanogr.*, *11*, 1443–1451.

Paffenhöfer, G. A. (1998), On the relation of structure, perception and activity in marine planktonic copepods, *J. Mar. Sys.*, *15*, 457–473.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1986), *Numerical Recipes: The Art of Scientific Computing*, Cambridge Univ. Press, New York.

Roman, M., S. Smith, K. Wishner, X. S. Zhang, and M. Gowing (2000), Mesozooplankton production and grazing in the Arabian Sea, *Deep Sea Res. II*, *47*, 1423–1450.

Schartau, M., and A. Oschlies (2003), 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., A. Engel, J. Schröter, S. Thoms, C. Völker, and D. Wolf-Gladrow (2007), Modelling carbon overconsumption and the formation of extracellular particulate organic carbon, *Biogeosciences*, *4*, 433–454.

Smith, S. L., L. A. Codispoti, J. M. Morrison, and R. T. Barber (1998), The 1994–1996 Arabian Sea Expedition: An integrated, interdisciplinary investigation of the response of the northwestern Indian Ocean to monsoonal forcing, *Deep Sea Res. II*, *45*, 1905–1915.

Spitz, Y. H., J. R. Moisan, and M. R. Abbott (2001), Configuring an ecosystem model using data from the Bermuda-Atlantic Time Series (BATS), *Deep Sea Res. II*, *48*, 1733–1768.

Sweby, P. K. (1984), High resolution schemes using flux limiters for hyperbolic conservation laws, *SIAM J. Numer. Anal.*, *21*(5), 995–1011.

Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res.*, *106*(D7), 7183–7192.

Tziperman, E., and W. C. Thacker (1989), An optimal-control/adjoint-equations approach to studying the oceanic general circulation, *J. Phys. Oceanogr.*, *19*, 1471–1485.

Weller, R. A., M. F. Baumgartner, S. A. Josey, A. S. Fischer, and J. C. Kindle (1998), Atmospheric forcing in the Arabian Sea during 1994–1995: Observation and comparisons with climatology and models, *Deep Sea Res. II*, *45*, 1961–1999.

Wiggert, J. D., R. G. Murtugudde, and J. R. Christian (2006), Annual ecosystem variability in the tropical Indian Ocean: Results of a coupled bio-physical ocean general circulation model, *Deep Sea Res. II*, *53*, 644–676.

Yamanaka, Y., N. Yoshie, M. Fujii, M. N. Aita, and M. J. Kishi (2004), An ecosystem model coupled with Nitrogen-Silicon-Carbon cycles applied to Station A7 in the Northwestern Pacific, *J. Oceanogr.*, *60*, 227–241.

L. A. Anderson, S. C. Doney, J. A. Dusenberry, and D. J. McGillicuddy Jr., Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA.

R. A. Armstrong, School of Marine and Atmospheric Sciences, State University of New York at Stony Brook, Stony Brook, NY 11794, USA.

F. Chai and M. Fujii, School of Marine Sciences, University of Maine, Orono, ME 04469-5706, USA.

J. R. Christian, Fisheries and Oceans Canada, Victoria, BC, Canada V8L 4B2.

J. Dunne, Geophysical Fluid Dynamics Laboratory, Princeton, NJ 08540, USA.

M. A. M. Friedrichs, Virginia Institute of Marine Science, College of William and Mary, P. O. Box 1346, Gloucester Point, VA 23062, USA. (marjy@vims.edu)

R. Hood, Center for Environmental Science, University of Maryland, Cambridge, MD 21613, USA.

J. K. Moore, Earth System Science, University of California, Irvine, Irvine, CA 92697, USA.

M. Schartau, Institute for Coastal Research, GKSS-Forschungszentrum, Geesthacht GmbH, Max-Planck-Strasse 1, D-21502 Geesthacht, Germany.

Y. H. Spitz, College of Ocean and Atmospheric Sciences, Oregon State University, Corvallis, OR 97331, USA.

J. D. Wiggert, Center for Coastal Physical Oceanography, Old Dominion University, Norfolk, VA 23529, USA.