Journal of Marine Systems xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Journal of Marine Systems

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

Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment

Jason K. Jolliff [a,*], John C. Kindle [b], Igor Shulman [b], Bradley Penta [b], Marjorie A.M. Friedrichs [c], Robert Helber [b], Robert A. Arnone [b]

[a] Building 1009, Naval Research Laboratory, Stennis Space Center, (NRL-Stennis) Mississippi 39529, USA
[b] NRL-Stennis, USA
[c] Virginia Institute of Marine Science, P.O. Box 1346, Gloucester Point, VA 23062-1346, USA

ARTICLE INFO

Article history:
Received 30 April 2007
Accepted 2 May 2008
Available online xxxx

Keywords:
Modeling
Marine ecosystem model
Statistical analysis
Remote sensing
Phytoplankton

ABSTRACT

The increasing complexity of coupled hydrodynamic-ecosystem models may require skill assessment methods that both quantify various aspects of model performance and visually summarize these aspects within compact diagrams. Hence summary diagrams, such as the Taylor diagram [Taylor, 2001, Journal of Geophysical Research, 106, D7, 7183–7192], may meet this requirement by exploiting mathematical relationships between widely known statistical quantities in order to succinctly display a suite of model skill metrics in a single plot. In this paper, sensitivity results from a coupled model are compared with Sea-viewing Wide Field-of-view Sensor (SeaWiFS) satellite ocean color data in order to assess the utility of the Taylor diagram and to develop a set of alternatives. Summary diagrams are only effective as skill assessment tools insofar as the statistical quantities they communicate adequately capture differentiable aspects of model performance. Here we demonstrate how the linear correlation coefficients and variance comparisons (pattern statistics) that constitute a Taylor diagram may fail to identify other potentially important aspects of coupled model performance, even if these quantities appear close to their ideal values. An additional skill assessment tool, the target diagram, is developed in order to provide summary information about how the pattern statistics and the bias (difference of mean values) each contribute to the magnitude of the total Root-Mean-Square Difference (RMSD). In addition, a potential inconsistency in the use of RMSD statistics as skill metrics for overall model and observation agreement is identified: underestimates of the observed field's variance are rewarded when the linear correlation scores are less than unity. An alternative skill score and skill score-based summary diagram is presented.

Published by Elsevier B.V.

1. Introduction

In general, mechanistic models that seek to simulate some natural phenomena must invariably be compared to observations in order to assess the model's skill. In accordance with this special volume on model skill assessment, we define *skill* as the model's fidelity to the truth. We further presume that since the truth cannot be known, assessment of model skill must begin with a quantification of the misfit between model results and imperfect observations. An overview of various model skill metrics, which may include known statistical quantities or novel functions and mathematical techniques, is given in Stow et al. (submitted for publication). In this paper, we present a pragmatic evaluation of some widely known statistical quantities for the purpose of model skill assessment as well as how relationships between these quantities may be exploited to make compact diagrams that summarize multiple aspects of model performance, i.e., summary diagrams. An important component of this analysis is the relationship

* Corresponding author. Tel.: +1 228 688 5308; fax: +1 228 688 4149.
E-mail addresses: jolliff@nrlssc.navy.mil (J.K. Jolliff), kindle@nrlssc.navy.mil (J.C. Kindle), igor.shulman@nrlssc.navy.mil (I. Shulman), penta@nrlssc.navy.mil (B. Penta), marjy@vims.edu (M.A.M. Friedrichs), helber@nrlssc.navy.mil (R. Helber), bob.arnone@nrlssc.navy.mil (R.A. Arnone).

0924-7963/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.jmarsys.2008.05.014

between various statistical quantities, which may be utilized to produce summary diagrams, but may also be deceptive if additional information is not presented. It is the general aim of this paper to demonstrate that a comprehensive and balanced approach to quantitative model skill assessment should include, at the very least, an acknowledgement of these relationships and an understanding of how they may influence the appearance of model skill.

More specifically, however, summary diagrams may be particularly suited to the task of skill assessment for spatially complex models with multiple state variables, such as a marine ecosystem model coupled to a hydrodynamic model (coupled models — e.g., Franks and Chen, 2001; Gregg et al., 2003; Walsh et al., 2003; Holt et al., 2005; Kindle et al., 2005; Allen et al., 2007). Indeed, summary diagrams present a useful method to succinctly communicate various aspects of coupled model performance since extensive lists of metric values in tabular form may become tedious. In addition, the use of summary diagrams should also be encouraged in order to address several other practical and scientific concerns. First, many coupled model skill assessment exercises that have appeared in recent literature still rely principally upon graphics that emphasize the direct visual comparisons between model results and observations (Stow et al., submitted for publication), such as a time series plot or a side-by-side comparison of one to two-dimensional property fields (chlorophyll, nitrate, etc.). If the statistical and graphical techniques that are integral to the summary diagram approach become more widely accepted and presented, then this may encourage more quantitative statements about coupled model skill. Second, summary diagrams are particularly useful for quantitatively comparing the performance of an ensemble of different models or multiple permutations of a single model. Given that there remains continuing uncertainly in the structure and parameterization of ecosystem models (e.g., Friedrichs et al., 2007), summary and quantitative skill assessment techniques may become an efficient facilitator of improved prognostic performance.

Accordingly, one potential statistical and graphical skill assessment approach is to render a Taylor diagram (Taylor, 2001). Taylor diagrams exploit relationships between known statistical quantities in order to provide summary information about particular aspects of model performance and were developed to aid in the monitoring of complex ocean–atmosphere climate models. The Taylor diagram, as is the case for many potential model skill assessment tools, is not discipline specific, and several recent marine ecosystem modeling papers have presented them as part of a model skill assessment scheme (Gruber et al., 2006; Raick et al., 2007). Here we begin with an assessment of the Taylor diagram and the statistics it communicates for the specific purpose of coupled model skill assessment. Taylor diagrams are an appropriate place to begin our evaluation of summary diagrams given their increasing use in a wide range of modeling disciplines; however, summary diagrams are only as useful as the metrics they communicate, and so our analysis includes an exposition of how relationships between widely known statistical quantities may be further utilized to construct other types of summary diagrams that communicate additional aspects of model performance.

While the statistical methods and diagrams developed and discussed here may potentially be applied to many other types of model result to data comparisons, we nonetheless present results from a coupled hydrodynamic-ecosystem model and ocean color products derived from SeaWiFS satellite ocean color data in order to explicitly illustrate potential problems arising from this type of skill assessment. To that end, summary information about the modeling and satellite ocean color methods is given below (Section 2), whereas detailed description of statistical methods and display techniques are fully explicated in due course of the main analysis (Section 3). In Section 3.1, we examine the Taylor diagram and the univariate statistics it summarizes by presenting several example applications that demonstrate the strengths and weaknesses of this approach. In Section 3.2, we develop an alternative summary diagram, the target diagram, which provides information about additional aspects of model performance that may be of particular concern to the skill assessment of ecosystem models. In Section 3.3, we identify a potentially undesirable property of RMSD-based metrics, and present an alternative skill score and skill score-based summary diagram.

2. Methods

Results from an experimental ecosystem modeling environment, the Naval Research Laboratory Ecological-Photochemical-Bio-Optical-Numerical Experiment (which for brevity is referred to as Neptune), are presented here as a prototypical example of a complex modeling system. Detailed description of the Neptune modeling construct, including all state equations, parameter designations, and optical calculations, may be found in Jolliff and Kindle (2007). The modeling system is composed of four core elements: (1) the biogeochemical model that describes the flow and transformation of elemental reservoirs (carbon, nitrogen, and phosphorus) as a result of phytoplankton primary production and subsequent physiological processes and trophic interactions; (2) a visible optics module that relates the biogeochemical elemental reservoirs to spectrally explicit optical properties, describes the vertically resolved attenuation of incident, spectrally decomposed irradiance, and budgets photons absorbed by living phytoplankton to perform light-growth calculations; (3) an ultraviolet (UV) optics module that determines the attenuation of spectrally decomposed UV irradiance and the potential UV-stimulated photochemical degradation of colored dissolved organic matter (CDOM); and (4) a description of the spectrally decomposed UV and visible irradiance boundary conditions.

The Neptune system is designed for integration with any hydrodynamic model capable of describing the advection–diffusion of state variables. Here we examine the one-dimensional case by coupling the model to the Modular Ocean Data Assimilation System (MODAS). MODAS is described in Fox et al. (2002). Briefly, the system uses optimal interpolation (Bretherton et al., 1976) to render daily satellite estimates of sea surface temperature (SST) and sea surface height (SSH) onto a two-dimensional grid. A subsurface temperature profile is then retrieved from the U.S. Navy's Master Oceanographic Observational Data Set. Deviation from subsurface climatology is then estimated based upon SST and SSH deviation from surface climatology. The result is a synthetic three-dimensional temperature field.

The MODAS fields were averaged over 4 years (2001–2004) to approximate an average annual cycle of summer thermal stratification followed by winter overturn for a 1°×1° area in the western Gulf of Mexico (center position 24.0° N, 94.5° W). Vertical eddy diffusion coefficients were imputed from MODAS synthetic temperature fields using the Pacanowski and Philander (1981) vertical mixing scheme. Daily and vertically resolved (total depth (z) = 161 m; Δz = 1 m) eddy diffusion coefficients were used to solve for the vertical turbulent mixing of model state variables using a fully implicit method with a time step of 1800 s. The coupled model was initialized using temperature–nutrient relationships observed in the Gulf of Mexico (Jochens et al., 2002) and then run for ten simulation years to solve for the steady state solution for transformations of carbon, nitrogen, and phosphorus in the upper ocean. The system was forced to material conservation by implicit remineralization of all particulates that sank below the deepest grid cell (z = 161 m).

The coupled model results were compared to local area coverage SeaWiFS ocean color data that were received and archived at the Naval Research Laboratory (NRL), Stennis Space Center. The satellite data were processed and the intervening atmospheric signal removed using NRL's Automated Processing System (APS). The atmospheric correction procedures are compliant with National Aeronautics and Space Administration SeaWiFS data processing protocols. Three NRL APS products derived from SeaWiFS data were examined: (1) the surface chlorophyll-*a* concentration, which was determined from the OC4v4 band ratio algorithm (O'Reilly et al., 1998); (2) the surface phytoplankton absorption coefficient (443 nm); and (3) the surface colored detrital matter (CDM) absorption coefficient (412 nm). The latter two products were determined from the multiband quasi-analytic algorithm (Lee et al., 2002), which estimates total absorption coefficients over SeaWiFS visible bands and then further decomposes them into phytoplankton and detrital contributions. Each daily spatial mean of SeaWiFS data through 4 years (2001–2004) from the 1° western Gulf of Mexico grid was used to construct a satellite ocean color time series wherein missing days due to clouds were accounted for via linear interpolation. The time series was lowpass filtered to remove variability from frequencies higher than 10 days; the averages were then computed to construct the annual climatology.

3. Results

The model results are compared with the daily climatology calculated from 4 years of SeaWiFS data (Fig. 1) for three surface bio-optical fields: the surface chlorophyll-*a* concentration, the surface phytoplankton absorption coefficient (443 nm), and the surface CDM absorption coefficient (412 nm). The satellite estimate of these surface quantities will be herein referred to as the reference field and the model's simulated surface bio-optical quantities will be referred to as simply the model field.

The Neptune model's three size-based phytoplankton functional groups are presently parameterized so that picophytoplankton have a higher absorption efficiency (per unit chlorophyll-*a*) than larger phytoplankton, as has been observed in the laboratory and in the field (e.g., Bricaud et al., 2004; Millan-Nunez et al., 2004). Thus the model phytoplankton absorption and total chlorophyll fields may vary with respect to one another due to differences in the relative dominance of simulated phytoplankton size fractions. In the example given in the following section, the satellite estimates of phytoplankton absorption and chlorophyll are thus used as a potential observational constraint on the simulated competition between phytoplankton size fractions.

3.1. Taylor diagrams and pattern statistics

For the one-dimensional case wherein the model's surface values are averaged over the upper 10 m each simulated day and are compared with a single daily reference value, the model and reference fields resemble sinusoidal functions of time, or waveforms (Fig. 1). Analogously, a measure of the potential phase shift between the two waveforms is also more generally a common measure of the agreement between two

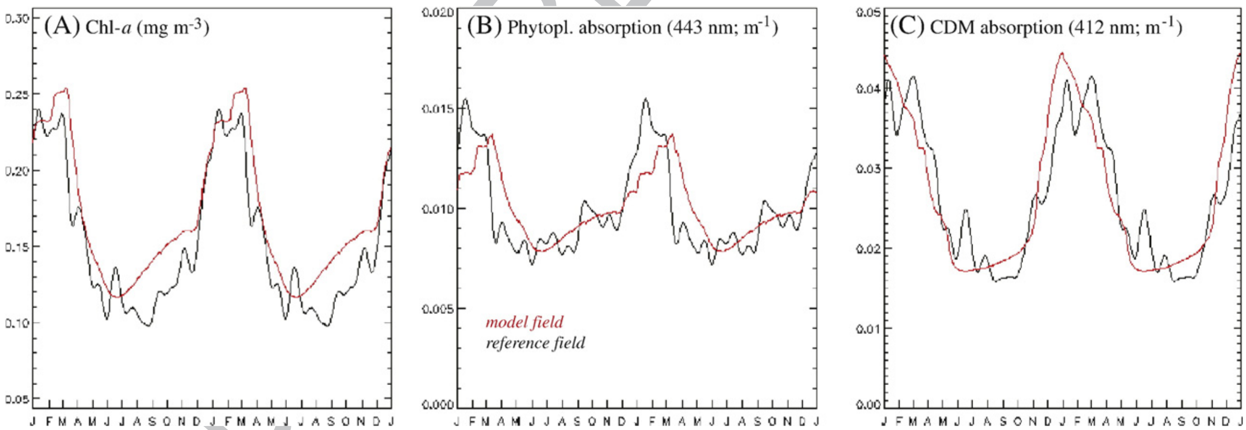

Fig. 1. Daily surface values for the (A) chlorophyll-*a* concentration (mg m^{-3}), (B) phytoplankton absorption coefficient (443 nm, m^{-1}), and (C) CDM absorption coefficient (412 nm, m^{-1}) are indicated for the final 2 years of the model's steady state solution (red line) and the SeaWiFS climatology (black line). Two years are shown in order to emphasize the winter peak and bring further emphasis to temporal misfits (i.e., phase misfits quantified by linear correlation coefficients). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: Jolliff, J.K., et al., Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, Journal of Marine Systems (2008), doi:10.1016/j.jmarsys.2008.05.014

fields: the linear correlation coefficient, R, which is defined by:

$$R = \frac{\frac{1}{N}\sum_{n=1}^{N}(m_n - \overline{m})(r_n - \overline{r})}{\sigma_m \sigma_r} \tag{1}$$

The letter m indicates the model field, r indicates the reference field, the overbar indicates the average, and σ is the standard deviation.

The correlation coefficient is bounded by the range $-1.0 \le R \le 1.0$. In general, as the phase between two temporal signals approaches agreement, R approaches 1.0. It is difficult, however, to discern information about the differences in amplitude between two signals from R alone. For this reason, another summary statistic, the normalized standard deviation, may be introduced:

$$\sigma* = \frac{\sigma_m}{\sigma_r} \tag{2}$$

The normalized standard deviation and the correlation coefficient from each of the three model to reference field comparisons may be displayed on a single Taylor diagram (Fig. 2). The Taylor diagram is a polar coordinate diagram that assigns the angular position to the inverse cosine of the correlation coefficient, R. A correlation coefficient of 0 is thus 90° away from a correlation coefficient of 1 (see scaling on Fig. 2). The radial (along-axis) distance from the origin is assigned to the normalized standard deviation, $\sigma*$. The reference field point, which is comprised of the statistics generated from a redundant reference to reference comparison, is indicated for the polar coordinates (1.0, 0.0). The model to reference comparison points may then be gauged by how close they fall to the reference point. This distance is proportional to the *unbiased* Root-Mean-Square Difference (RMSD′), as defined by:

$$\mathrm{RMSD}' = \left(\frac{1}{N}\sum_{n=1}^{N}\left[\left(m_n - \overline{m}\right) - \left(r_n - \overline{r}\right)\right]^2\right)^{0.5} \tag{3}$$

where the overbars indicate the mean. The term *unbiased* is used herein to emphasize that Eq. (3) removes any information about the potential bias (B), which is defined as the difference between the means of the two fields:

$$B = \overline{m} - \overline{r} \tag{4}$$

In other words, the unbiased RMSD (RMSD′) is equal to the total RMSD if there is no bias between the model and reference fields. This may be verified given the quadratic relationship between the unbiased RMSD, the bias, and the total RMSD:

$$\mathrm{RMSD}^2 = B^2 + \mathrm{RMSD}'^2 \tag{5}$$

where the total RMSD is a measure of the average magnitude of difference and is defined by:

$$\mathrm{RMSD} = \left[\frac{1}{N}\sum_{n=1}^{N}(m_n - r_n)^2\right]^{0.5} \tag{6}$$

In contrast, the unbiased RMSD may be conceptualized as an overall measure of the agreement between the amplitude (σ) and phase (R) of two temporal patterns. For this reason, the correlation coefficient (R), normalized standard deviation ($\sigma*$), and unbiased RMSD are collectively referred to herein as

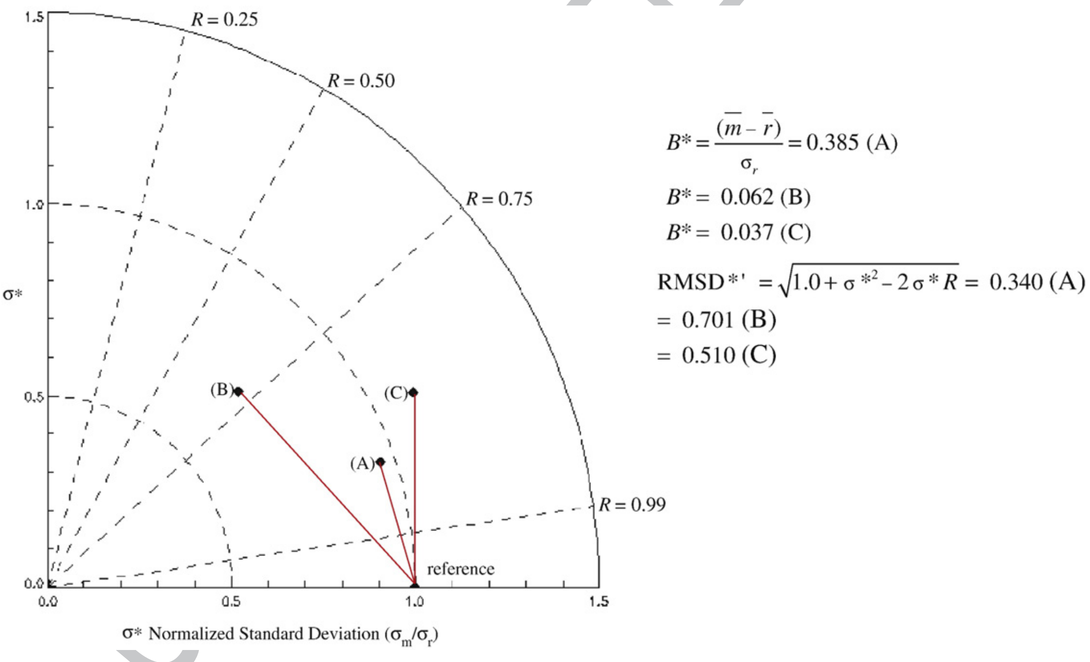

Fig. 2. Taylor diagram rendering of the model to reference field comparisons shown in Fig. 1: (A) chlorophyll-*a* concentration (mg m^{-3}), (B) phytoplankton absorption coefficient (443 nm, m^{-1}), and (C) CDM absorption coefficient (412 nm, m^{-1}). As explained in the text, the radial distance is proportional to the normalized standard deviation ($\sigma*$) and the angular position corresponds to the linear correlation coefficient (R values). In accordance with Eq. (7), the distances between the labeled points and the reference point are proportional to the unbiased RMSD, Eq. (3).

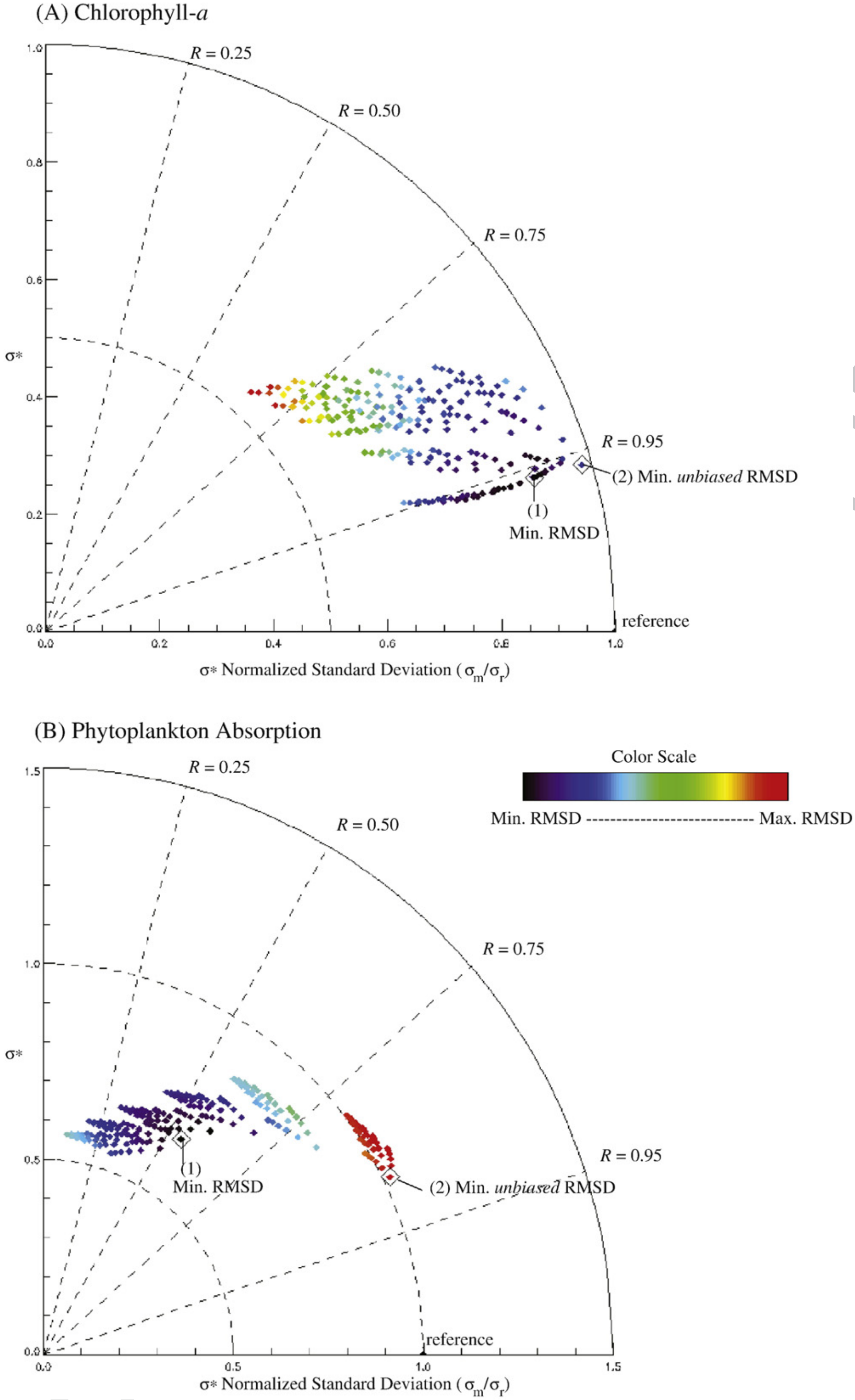

Fig. 3. Taylor diagrams for grazing sensitivity model executions showing model to reference statistics for the (A) surface chlorophyll-*a* field and (B) the surface phytoplankton absorption field. The minimum total RMSD (1) and the minimum unbiased RMSD (2) are indicated on each plot. The color scale is added to both Taylor diagrams and corresponds to the minimum total RMSD (black) to the maximum total RMSD (red) for each set of model to reference comparison statistics. The time series results corresponding to points (1) and (2) in (B) are shown in Fig. 4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pattern statistics. The three pattern statistics are related to one another by:

$$\mathrm{RMSD}'^2 = \sigma_r^2 + \sigma_m^2 - 2\sigma_r\sigma_m R \quad (7)$$

It is this relationship that makes the Taylor diagram useful: the individual contribution of misfits in amplitude may be compared to misfits in phase to discern how they contribute to the unbiased RMSD. Since the diagram is in standard deviation normalized space, the distance from the model points to the reference points is also proportional to Eq. (7), which recast in standard deviation normalized units (indicated by the asterisk) becomes:

$$\mathrm{RMSD}^{*\prime} = \sqrt{1.0 + \sigma*^2 - 2\sigma* R} \quad (8)$$

Note also that it can be shown that the minimum of this function occurs where $\sigma^* = R$. This is an important relationship that we will refer to at several points later in the text.

Fig. 2 shows that the chlorophyll model to reference field comparison point (A) appears closest to the reference point, whereas the phytoplankton absorption comparison point (B) appears farthest due to a poorer correlation as well as an underestimate of the standard deviation. Indeed, the chlorophyll comparison has the lowest normalized and unbiased RMSD. However, the normalized bias, defined as:

$$B* = \frac{(\overline{m} - \overline{r})}{\sigma_r} \quad (9)$$

is much larger for the model chlorophyll field, which consistently tends to overestimate the reference field (as shown in Fig. 1A). Thus caution must be applied when interpreting a Taylor diagram wherein no information about the bias is included.

The importance of adding information about the bias may also be further demonstrated using a large number of model executions, such as during a sensitivity analysis. The advantage of the Taylor diagram in such cases is that it allows one to discern how the phase and amplitude of a simulated field change as the model is modified. The disadvantage is that information about any potential model to reference field bias must be somehow added to the diagram.

For example, the mortality rate for phytoplankton (ε_r) in the Neptune ecological model is described using the Ivlev (1961) formulation:

$$\varepsilon_r = \varepsilon_m \left(1.0 - e^{-Iv(C)}\right) \quad (10)$$

where Iv is the Ivlev parameter that describes how the maximum potential mortality rate (ε_m) is attenuated with decreasing phytoplankton biomass (C). With three phytoplankton functional groups and an estimated Iv parameter space incremented for 6 values, there are 216 potential grazing permutations.

The results of 216 separate model executions are shown on two Taylor diagrams (Fig. 3). For brevity, only the first two field comparisons, phytoplankton chlorophyll and phytoplankton absorption, are shown since the CDM absorption field is somewhat less sensitive to the grazing parameter selections. It is important to note that the model and reference fields were not log-transformed. In this case, it would not make a considerable difference; however, if there were large outliers in either field then log-transformation may significantly impact the value of statistical quantities. Some investigators may choose to log-transform the fields first, particularly if the bio-optical fields range over several orders of magnitude. If the fields are log-transformed then the investigator should be cognizant that statistical quantities generated from non log-transformed values may be different.

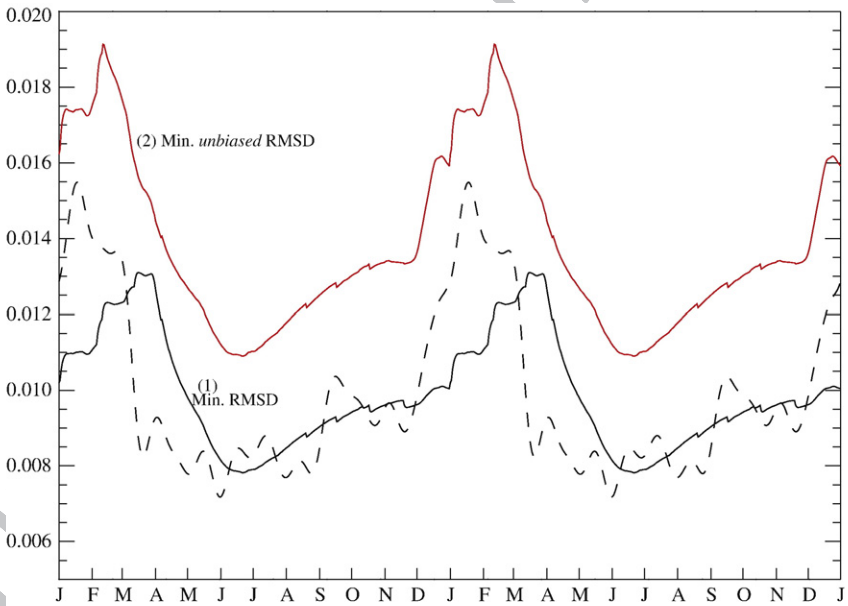

Fig. 4. The reference field phytoplankton absorption (dashed line) is compared to the minimum total RMSD (1 — solid black line) and the minimum unbiased RMSD (2 — red line); these time series correspond to points (1) and (2) in Fig. 3B. As in Fig. 1, two years are shown to emphasize the winter peak and draw emphasis to phase misfits quantified by the linear correlation coefficients. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In both Taylor diagrams presented here, the model points that come closest to the reference point have the smallest unbiased RMSD value (Fig. 3). It would appear that the cluster of model points closest to the reference point may thus provide the closest fit to the data. Here, however, the inclusion of a relative total RMSD color scale, which indicates the range of minimum to maximum total RMSD using a spectral (rainbow) color scaling increment (Fig. 3), reveals that some points nearest the reference point may have larger total RMSD values. This is particularly the case for phytoplankton absorption (Fig. 3B) where the cluster of points closest to the reference point also have the largest total RMSD values. For the phytoplankton absorption field, improvement in the correlation coefficient appears to come at the expense of an increase in the bias, and consequently, the total RMSD. The minimum total RMSD (point 1) and minimum unbiased RMSD (point 2) from the phytoplankton absorption comparisons are also shown as a time series plot (Fig. 4). Clearly, the red line (minimum unbiased RMSD) has a better phase agreement but overestimates the observed values.

In coupled hydrodynamic-ecosystem modeling applications, information about the bias and the total RMSD may be just as important to the investigator as information about the pattern statistics, particularly when evaluating the sensitivity of a model to parameter selection for the purpose of minimizing the magnitude of the misfit between the model and reference fields. Taylor (2001) suggested adding lines of various lengths corresponding to the total RMSD in proportion to the unbiased RMSD onto the Taylor diagram; however, this procedure may result in a confusing diagram when large numbers of model runs are compared. A color scale modification of the Taylor diagram, as shown here (Fig. 3), may also be useful but the overall import of the Taylor diagram may nonetheless be easily misinterpreted.

3.2. Target diagrams

An alternative to the Taylor diagram is to formulate a target diagram that provides summary information about the pattern statistics as well as the bias thus yielding a broader overview of their respective contributions to the total RMSD. The relationship between the bias, unbiased RMSD, and the total RMSD (Eq. (5)) provides a convenient starting point to construct such a diagram. In a simple Cartesian coordinate system, the unbiased RMSD may serve as the *X*-axis and the bias may serve as the *Y*-axis. The distance between the origin

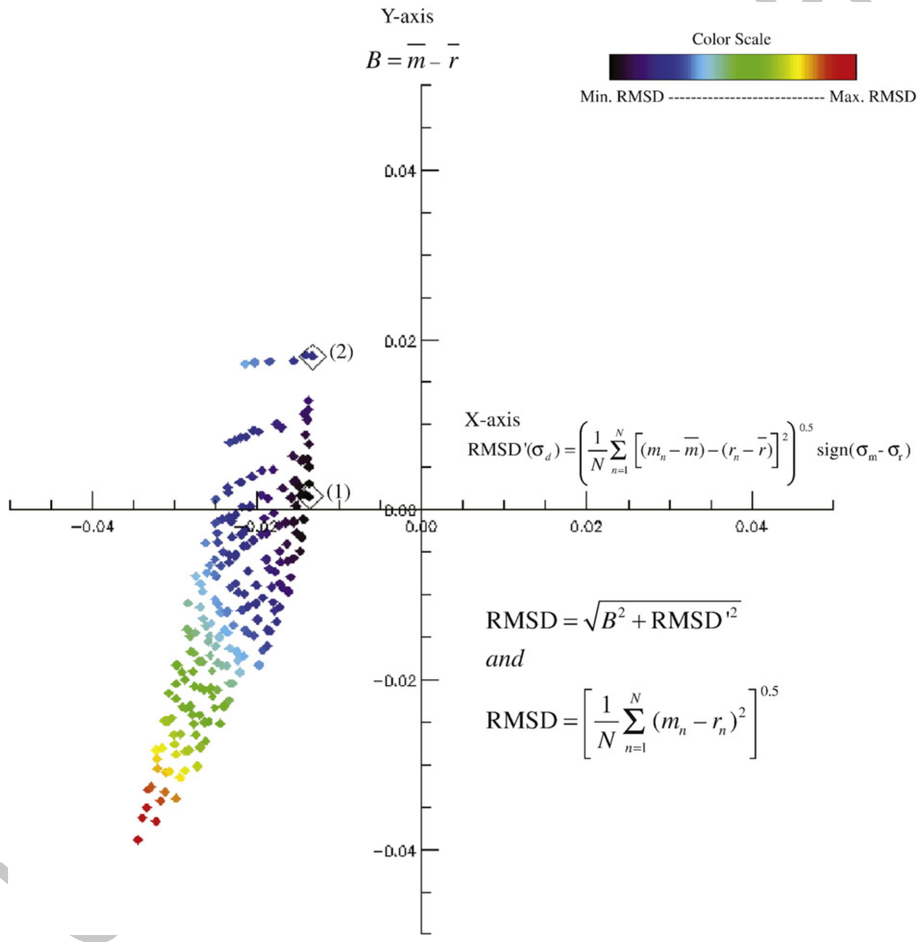

Fig. 5. Target diagram for model chlorophyll-*a* and reference chlorophyll-*a* comparisons. The *Y*-axis corresponds to the bias, the *X*-axis corresponds to the unbiased RMSD multiplied by the sign of the model and reference standard deviation difference (σ_d), and the distance from each point to the origin is proportional to the total RMSD. The minimum total RMSD (1) and the minimum unbiased RMSD (2) are indicated on the plot. The color scaling is the same as in Fig. 3.

and the model versus observation statistics (any point, s, within the X,Y Cartesian space) is then equal to the total RMSD (Fig. 5).

By definition, the X-axis (unbiased RMSD) must always be positive. However, the $X<0.0$ region of the Cartesian coordinate space may be utilized if the unbiased RMSD is multiplied by the sign of the standard deviation difference (σ_d):

$$\sigma_d = \mathrm{sign}(\sigma_m - \sigma_r) \tag{11}$$

The resulting target diagram thus provides information about whether the model standard deviation is larger ($X>0$) or smaller ($X<0$) than the reference field's standard deviation, in addition to a positive ($Y>0$) or negative bias ($Y<0$) (Fig. 5). The units of this diagram are all in chlorophyll concentration (mg m^{-3}), but this may again be addressed by normalizing the quantities by the reference field standard deviation (Fig. 6), such that the distance of each point from the origin is the standard deviation normalized total RMSD:

$$\mathrm{RMSD}^{*2} = B^{*2} + \mathrm{RMSD}^{*\prime\,2} \tag{12}$$

Rendering the diagram in normalized units allows one to better compare the model's chlorophyll performance with other potential areas of performance such as CDM absorption and phytoplankton absorption.

Furthermore, markers within the diagram may be added to provide an additional basis for interpreting model performance. For example, the investigator may wish to gauge how the model's total RMSD compares to the time series mean. In other words, if the first guess is the time series average, does the model provide an overall improvement over the first guess with respect to the minimization of the average misfit between the model and reference fields?

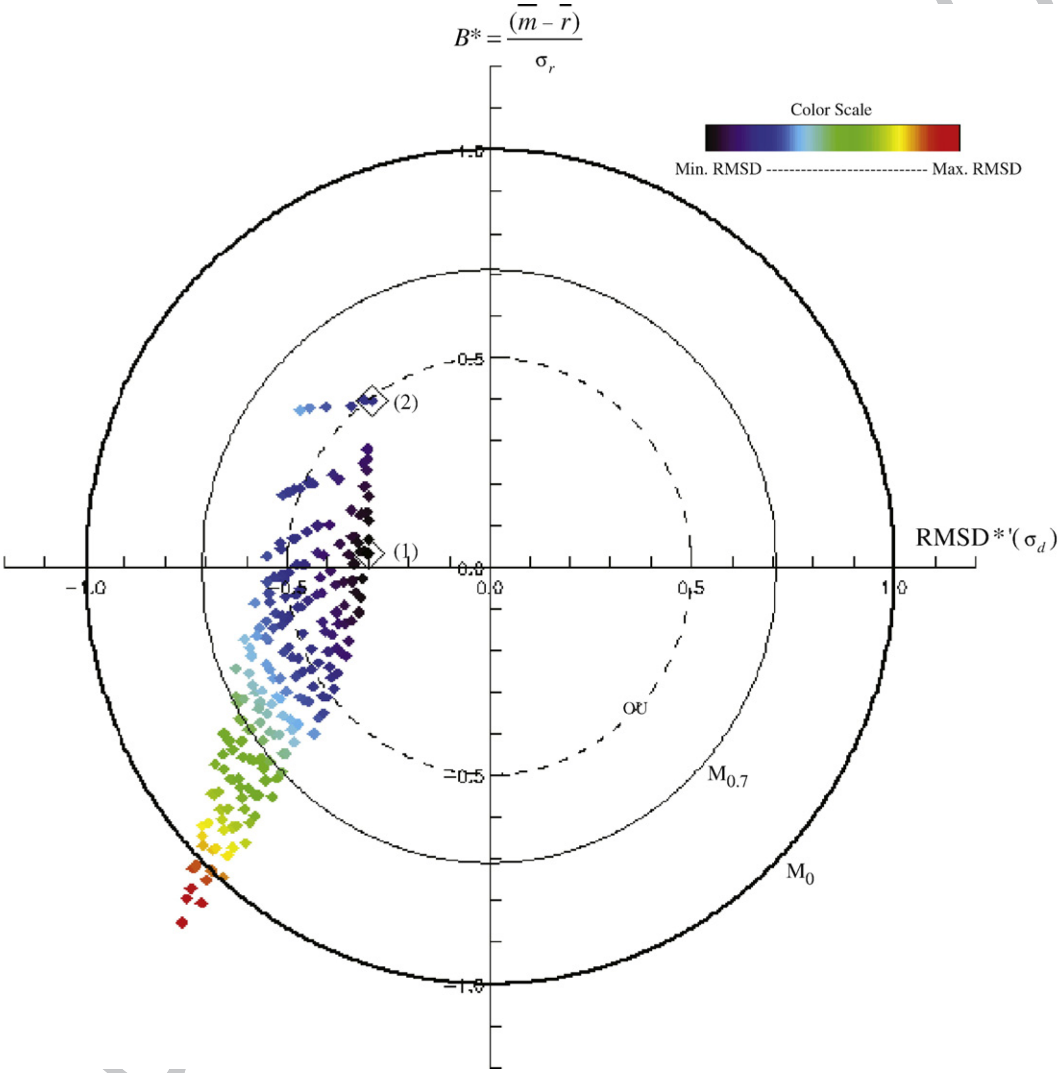

Fig. 6. Normalized target diagram for model chlorophyll-a and reference chlorophyll-a comparisons. The axes are the same as in Fig. 4, only they are normalized by the reference field standard deviation (indicated by *). The thick line (M_0) corresponds to a normalized total RMSD of 1.0, the thin line ($M_{0.7}$) corresponds to RMSD* = 0.71. The significance of these markers is explained in the text. The dashed line represents the threshold of observational uncertainty (OU). The minimum total RMSD (1) and the minimum unbiased RMSD (2) are indicated on the plot. The color scaling is the same as in Figs. 3 and 5.

The total RMSD between the reference field and the reference field mean is simply the reference field's standard deviation. Since the diagram is in standard deviation normalized units, a normalized total RMSD value of 1.0 provides a convenient performance marker (marker M_0, Fig. 6). If the investigator is concerned with the total RMSD, and not merely the pattern statistics, then any points greater than RMSD* = 1 may be considered poor performers since they offer no improvement over the time series average.

It is also interesting to note that the normalized total RMSD (RMSD*) is related to the modeling efficiency (MEF) metric presented in Stow et al. (submitted for publication) via the relationship: MEF = 1 − RMSD*^{2}. The MEF may be used to discern how well a model performs as a predictor of the data compared to the mean of the data (Stow et al., 2003; Nash and Sutcliffe, 1970). This underscores the significance of the RMSD* = 1 (M_0) marker within the normalized target diagram since points between it and the origin also have a better than average MEF score.

A weakness of the target diagram is that it does not provide explicit information about the correlation coefficient. However, there are certain limits inherent in the statistics summarized by the diagram that one may use to make some inference about the correlation coefficient. For example, recall the relationship between the correlation coefficient, the normalized standard deviation, and the normalized and unbiased RMSD (Eq. (8)). It can be shown that for values of R (where $-1.0 \le R < 0.0$) the minimum value of RMSD*′ for all potential values of σ^* (where $0.0 < \sigma^* < \infty$) approaches 1.0. Thus no model/reference comparison points that appear on the target diagram within the range of $-1.0 < X < 1.0$ can be negatively correlated. Since the square of the normalized bias must always be positive, then by extension all points where RMSD* < 1.0 must also be positively correlated. In other words, the first marker at RMSD* = 1.0 (marker M_0, Fig. 5) also establishes that all points between it and the origin are positively correlated. Positively correlated results may appear outside this marker; however, these points will have a large magnitude of difference from the observations due to either a significant bias, a difference in variance, or some combination thereof. This relationship may be formally expressed as follows:

$$\text{for } \forall s \in \{\mathrm{RMSD}* \mid \mathrm{RMSD}* \le 1.0\} \rightarrow R > 0.0 \qquad (13)$$

where s is a notation for any point on the target diagram. Similar such markers based upon the correlation coefficient may be established closer to the origin for values of R where $R > 0.0$. In accordance with Eq. (8), the minimum value of RMSD*′ occurs for any positive value of R where $\sigma^* = R$. Thus if one wants to determine the minimum unbiased RMSD value possible (M_{R1}) given a specific correlation value, $R1$, then the solution may be expressed as:

$$M_{R1} = \min\left(\mathrm{RMSD}*'\right) = \sqrt{1.0 + R1^2 - 2R1^2} \qquad (14)$$

Since the minimum total RMSD must also occur where the bias is equal to 0.0, M_{R1} is also the minimum total RMSD value for a given correlation coefficient value, $R1$. For the general case where $R1 > 0.0$:

$$\text{for } \forall s \in \{\mathrm{RMSD}* \mid \mathrm{RMSD}* \le M_{R1}\} \rightarrow R \ge R1 \qquad (15)$$

For example, Fig. 6 shows the second marker towards the origin for $R1 = 0.7$. Thus all points between this marker ($M_{0.7}$) and the origin are indicative of a correlation coefficient greater than 0.7.

The color scale in Fig. 6 is redundant: both the distance from the origin and the color index are proportional to the total RMSD. The color variable is thus left as a free variable that may be used to also explicitly indicate the correlation coefficient, or it may be used to indicate any supplemental information regarding the simulations that are displayed in the diagram (Friedrichs et al., submitted for publication). In our example, the sensitivity analysis is focused upon the grazing parameters. We may define an aggregate index of phytoplankton grazing stress (AI) as the sum of the three Ivlev parameters and display this index using the color scale, as in Fig. 7. Clearly, the AI most appreciably impacts the bias: as aggregate grazing stress increases the simulations consistently underestimate the satellite-based observations of surface chlorophyll. Furthermore, the lowest aggregate grazing stress corresponds to the highest bias (point 2, Fig. 7).

Diagrams that summarize repeated comparisons of model results and data should also make some indication of uncertainties that exist within the data. One may define data as truth plus some unknown observational uncertainty. The advantage of using a satellite climatology based upon a large number of spatial means, as in this case, is that one may choose to assume that the ensemble average observational uncertainty approaches zero as the total number of observations becomes very large ($\sim n > 1000$). One approach might be to state that assumption and forego any further indication of observational uncertainty. A note of caution must also be applied insofar as this approach assumes that the observational uncertainty is also unbiased.

Nevertheless, for the more general case there exists a large sum of potential observational uncertainties arising, in part, from measurement error. For satellite data, these errors may arise from imperfections in the satellite sensor, errors in the algorithms applied, atmospheric correction errors, and numerous other areas beyond the scope of this paper. It is therefore reasonable to assume that there must be some average minimum threshold value for the total RMSD below which further improvement in model/data agreement may not be significant. The dashed line in Fig. 5 is an estimate of this observational uncertainty (OU) threshold. Points that fall between this limit and the origin are all within the range of estimated observational uncertainty.

To be sure, observational uncertainty is a potentially complicated and contentious subject. Our objective here is to simply represent some estimate of this uncertainty on the target diagram so as to indicate where further efforts towards improved model to data agreement may not be a prudent use of time and resources. While it is entirely reasonable and appropriate to assume that observational uncertainty does provide an upper-limit upon potential improvements in model performance, our tentative estimates of this average uncertainty should be regarded as preliminary and much more work in this area needs to be done.

In this case, an average observational uncertainty was assumed for the satellite time series based on literature values for chlorophyll algorithm accuracy in optically deep waters (Bailey and Werdell, 2006; McClain et al., 2006) without any further consideration of the uncertainty within the measurements to which the satellite data are compared. If the average observational uncertainty (α) is expressed as a percent, then $\alpha\overline{r}$ may be used as an estimate for the average value of uncertainty for the time series. For example, a α value of ±15% and an average chlorophyll-*a* observation of 0.2 mg m^{-3} would yield an average uncertainty of ±0.03 mg m^{-3}. A model to reference field total RMSD of <0.03 mg m^{-3} is within the average observational uncertainty threshold and further improvement (model to data misfit reduction) may not be meaningful.

This assumed OU limit may be placed on the target diagram by normalizing $\alpha\overline{r}$ by the reference field standard deviation (dashed line, Fig. 7). The normalization procedure effectively means that the assumption of average observational uncertainty (α) is divided by the coefficient of variation, which is the reference field standard deviation divided by the reference field mean. The coefficient of variation is a common measure of the dispersion within a distribution. It is beyond the scope of this paper to further examine how the dispersion, in turn, may be impacted by the observational uncertainty, but we recognize that they are not necessarily independent.

In summary, the target diagram displays the model to reference field bias (*Y*-axis) and the model to reference field unbiased RMSD (*X*-axis). The distance between any point, *s*, and the origin is then the value of the total RMSD. All of the quantities may be normalized by the reference field standard deviation to remove the units of measurement. The outermost marker (M_0=RMSD*=1.0) establishes that all points between it and the origin represent positively correlated model and reference fields, and also have a better than average MEF score. A second marker may be added to indicate another positive *R* value, such as *R*=0.7, for which all points between it and the origin are greater than *R*. Finally, a dashed line indicates the estimate of average observational uncertainty and further model to data agreement for points between this marker and the origin may not be meaningful.

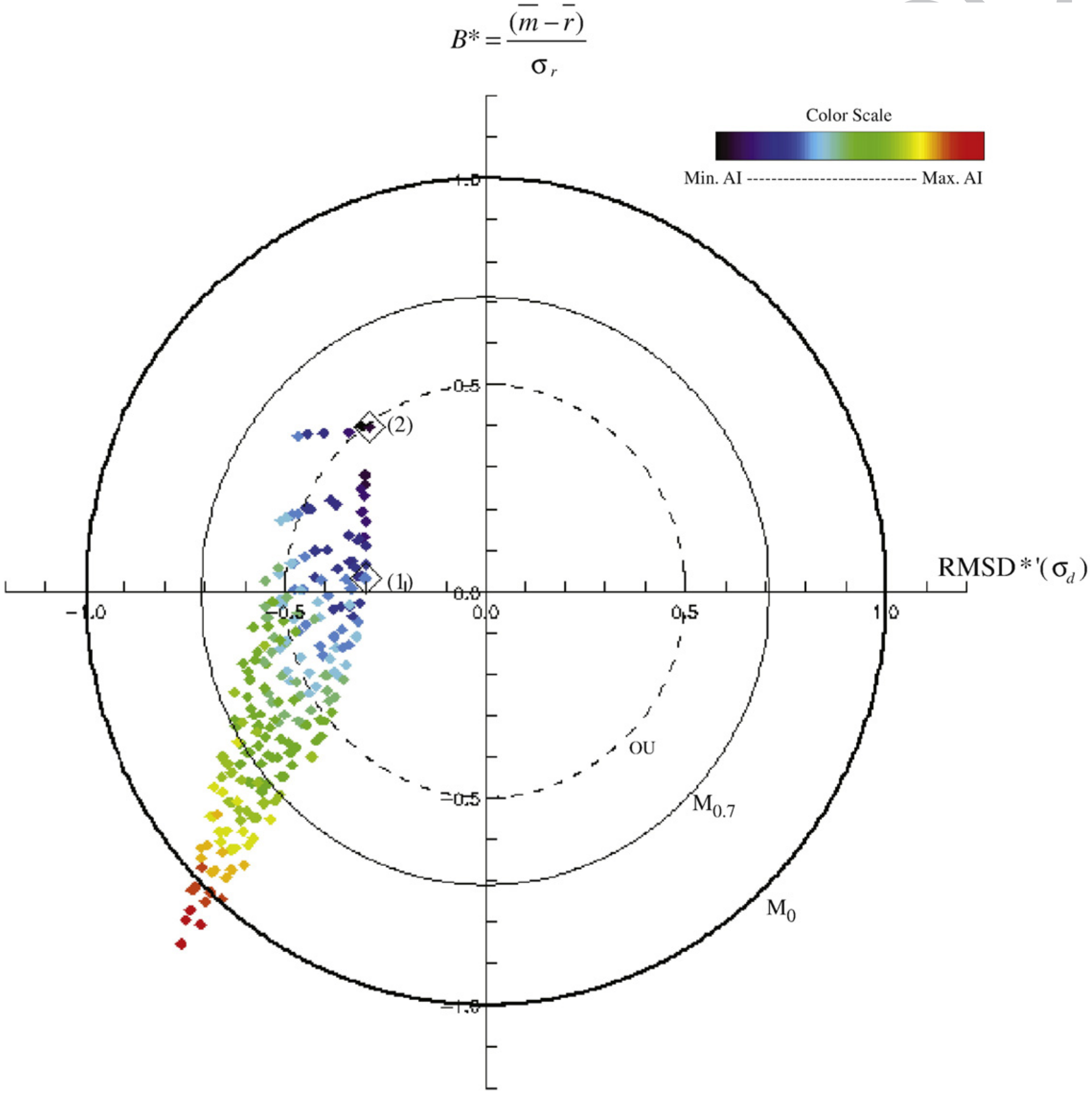

Fig. 7. Normalized target diagram for model chlorophyll-*a* and reference chlorophyll-*a* comparisons. The axes and the markers are the same as in Fig. 6. The color scaling has been changed to indicate the aggregate index (AI) for grazing stress, as explained in the text.

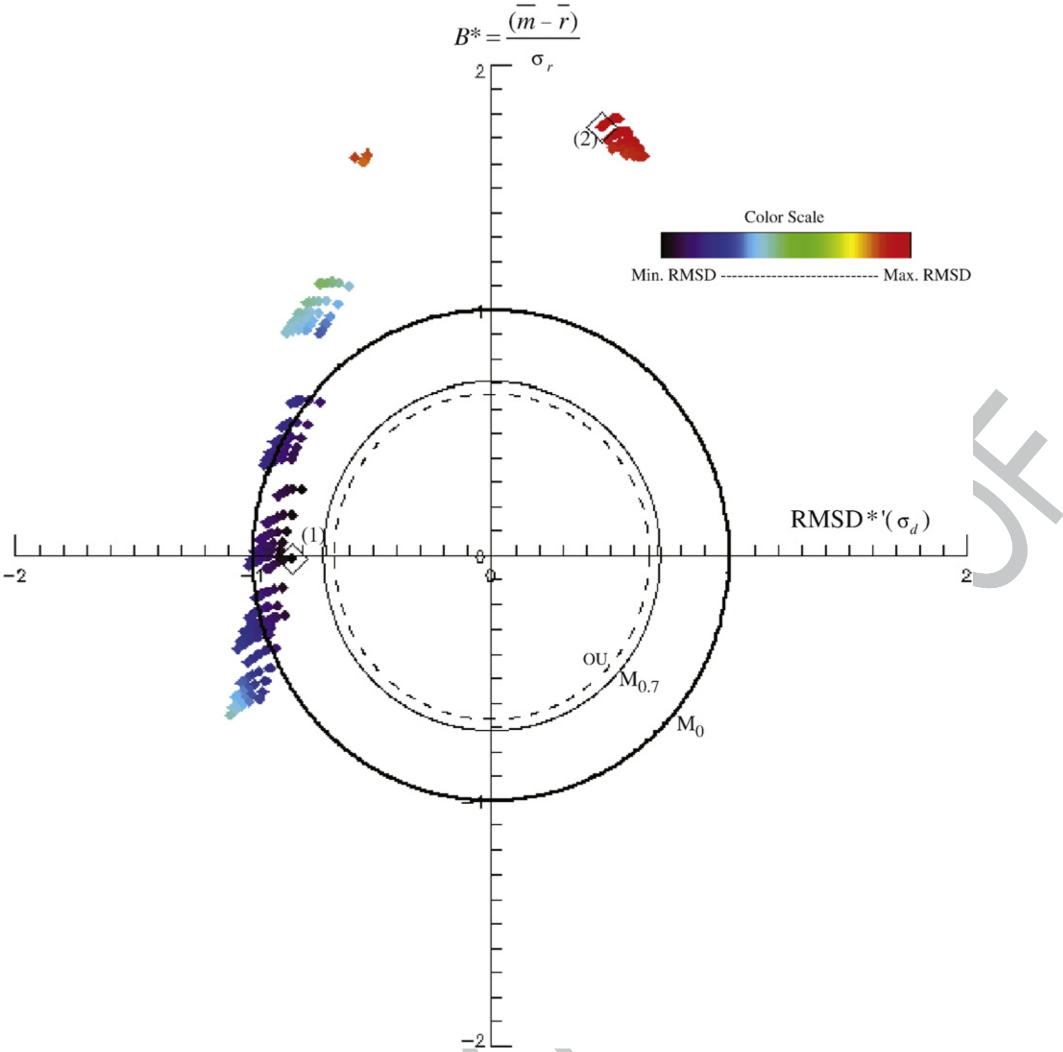

Fig. 8. Normalized target diagram for model/reference phytoplankton absorption fields. The axes are normalized by the reference field standard deviation (indicated by *). The thick line (M_0) corresponds to a normalized total RMSD of 1.0, the thin line ($M_{0.7}$) corresponds to RMSD* = 0.71. The significance of these markers is explained in the text. The dashed line represents the threshold of observational uncertainty (OU). The minimum total RMSD (1) and the minimum unbiased RMSD (2) are indicated on the plot.

The target diagram was also constructed for the phytoplankton absorption field (Fig. 8). In order to display the entire set of model versus reference comparisons for phytoplankton absorption, the scale for the target diagram (Fig. 8) had to be expanded to encompass RMSD* = 2. Note that the simulations with the best pattern statistics (Fig. 3B) also have a very large positive bias (red cluster, Fig. 8). In this particular case, the target diagram better delineates poor performing model executions than the Taylor diagram since the model is prone to a large bias for this field.

3.3. The skill target diagram

Additional alternatives to the Taylor diagram for summarizing pattern statistics as a measure of model skill may be preferable since there is a subtle discrepancy between improving the unbiased RMSD and improving the individual correlation coefficient and standard deviation statistics, and there may be circumstances where this consideration is important. For example, consider that there may be fundamental limits to the expected agreement between a model and a reference field. Even if all model inaccuracies and observational uncertainties could be eliminated, there may yet remain unforced oscillations that prevent exact model/reference field agreement. Suppose that an estimate of this uncertainty yields a maximum potentially attainable correlation coefficient of 0.65. As stated in Section 3.1, the minimum value of the unbiased RMSD occurs where $\sigma^* = R$ for positive values of R.

This relationship may be demonstrated on a Taylor diagram (Fig. 9). For $R = 0.65$ the minimum RMSD*′ value occurs where $\sigma^* = 0.65$. The three sets of pattern statistics correspond to the waveforms in Fig. 9B. The minimum average difference is the smallest amplitude pattern, but if amplitude and phase are weighed equally, as in a potential alternative measures of model skill, then the waveform where $\sigma^* = 1$ may be the most skillful.

This example demonstrates the implicit contradiction between minimizing the RMSD and improving σ^* towards an ideal value of 1.0. If the goal is to improve the total RMSD then σ^*

Please cite this article as: Jolliff, J.K., et al., Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, Journal of Marine Systems (2008), doi:10.1016/j.jmarsys.2008.05.014

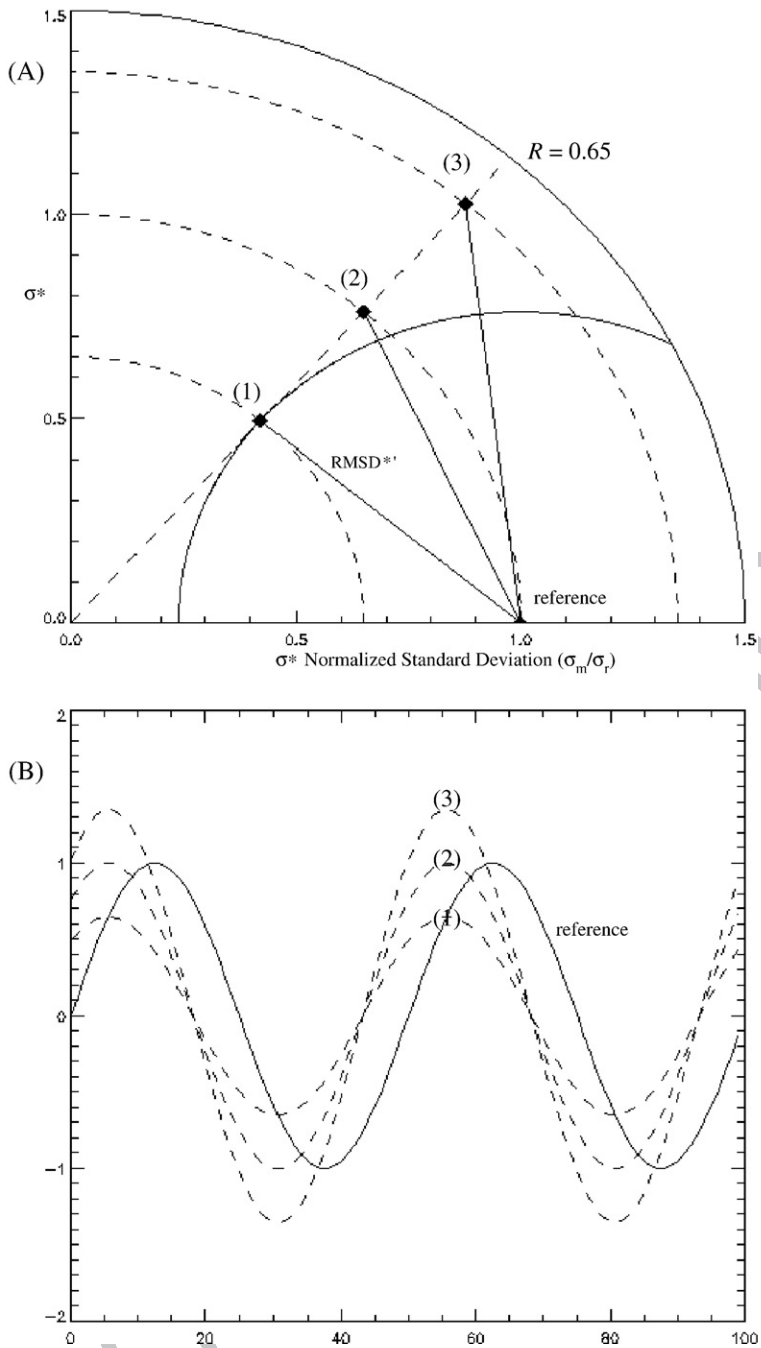

Fig. 9. (A) A Taylor diagram is shown for three model to reference field comparisons where $R=0.65$ and (1) $\sigma^*=0.65$, (2) $\sigma^*=1.0$, and (3) $\sigma^*=1.35$. An example of three sinusoidal waveforms and a reference field corresponding to the statistics in (A) is shown in panel (B).

values <1.0 are preferable. Clearly, if the two signals are out of phase, then reduction in the model variance to a threshold value diminishes the total RMSD value. However, if the goal of the investigation is to independently move R and σ^* as close to an ideal value of 1.0 as is possible then it may be inappropriate to use the total or unbiased RMSD as a model validation metric.

This is an important point since many model and observation comparison exercises may involve RMSD-based metrics. For example, Wallhead et al. (submitted for publication) use the term "skillful" to refer to model predictions that minimize mean-square differences. Sheng and Kim (submitted for publication) use RMSD metrics and Taylor diagrams as part of their water quality model evaluation scheme. Smith et al. (submitted for publication) use an RMSD-based cost function as part of a data assimilation scheme. Indeed, RMSD-based metrics of model performance are likely to continue to be used in a wide variety of contexts and investigators should at least be cognizant of how RMSD-based functions or

Please cite this article as: Jolliff, J.K., et al., Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, Journal of Marine Systems (2008), doi:10.1016/j.jmarsys.2008.05.014

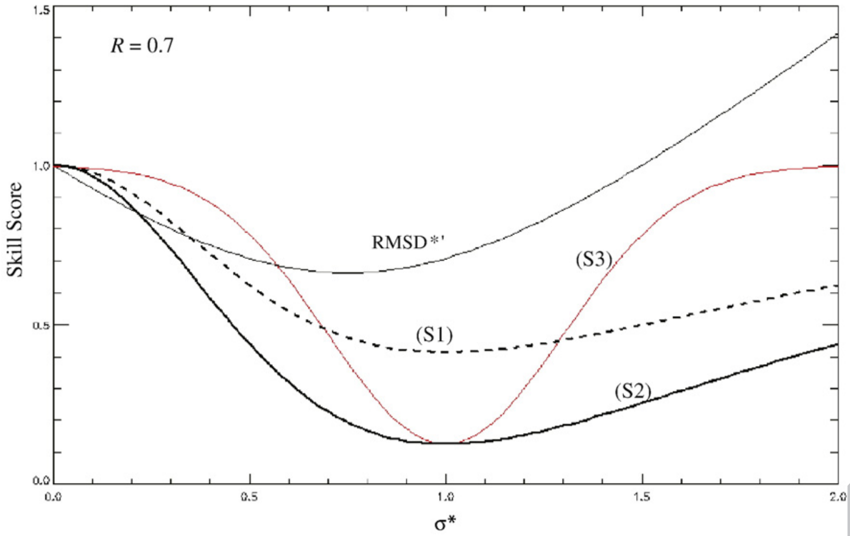

Fig. 10. The unbiased RMSD and skill scores $S1$–$S3$ are shown for $R = 0.7$ and σ^* over the range [0, 2].

skill scores quantify mismatches in variance when correlation coefficients are less than unity.

Alternative metrics of model skill (skill scores) have been proposed (Murphy and Epstein, 1989; Taylor, 2001), such as:

$$S1 = 1.0 - \left[\frac{2(1+R)}{(\sigma*+1/\sigma*)^2}\right] \tag{16}$$

and

$$S2 = 1.0 - \left[\frac{(1+R)^4}{(\sigma*+1/\sigma*)^2 4}\right] \tag{17}$$

The prevailing convention is to have the skill score range between 0.0 (for poor skill) and 1.0 (for superior skill). This convention is reversed here since our objective is to build a summary skill target diagram similar to the one developed in Section 3.2.

An important feature to consider is how these potential skill scores proportionally penalize underestimates or overestimates of the standard deviation. For example, given a constant R value of 0.7, the normalized and unbiased RMSD, $S1$, and $S2$ are shown for $0.0 \le \sigma^* \le 2.0$ in Fig. 10. Minimum skill scores occur where $\sigma^* = 1$, consistent with our stated skill score convention. However, $S1$ and $S2$ appear to penalize underestimates of the variance more than proportional overestimates, and are thus opposite of the RMSD*′ statistic that rewards variance underestimates. A potential alternative to these measures is a Gaussian function that penalizes proportional overestimates and underestimates of σ^* equally over the range [0, 2]. Multiplication by a scaled correlation score may then constitute a measure of model skill:

$$S3 = 1.0 - \left(e^{-\frac{(\sigma*-1.0)^2}{0.18}}\right)\left(\frac{1+R}{2}\right) \tag{18}$$

This measure of skill may now be incorporated into a diagram similar to the one developed in the previous section. Here, however, the emphasis is on the comparison of one model to another more than the misfit between the model and the data. Accordingly, a relative measure of bias may be given as:

$$B_m = \frac{B_i}{|\mathrm{Max}\{B_{i=1,2,3\,\ldots\,n}\}|} \tag{19}$$

that is, the maximum normalized bias of the ith model execution is its bias divided by the maximum magnitude bias from the total set of n model to data comparisons.

If B_m serves as the Y-axis and $S3$ times the sign of the standard deviation difference (σ_d) serves as the X-axis, then the resulting skill target diagram renders distances from the origin that are proportional to:

$$\mathrm{ST} = \sqrt{B_m^2 + S3^2} \tag{20}$$

The contrast between the ST score and the total RMSD is that the skill score does not reward underestimates of the variance for correlation values less than one. Markers for the skill target diagram are based on the percentile ST score of the models. For example, in this case the mean ST score ($\overline{\mathrm{ST}}$) is 0.51 and the standard deviation (σ_{ST}) is 0.28, thus the 90th percentile (assuming a normal score probability density function and recalling our skill convention rewards low scores instead of high scores) corresponds to $\overline{\mathrm{ST}} - 1.28\,\sigma_{\mathrm{ST}}$ or ST = 0.15. A similar marker for the 50th percentile ($\mathrm{ST} = \overline{\mathrm{ST}}$) is shown on Fig. 11. In this case, the most skillful simulation (point 2, Fig. 11) is yet again different from the minimum total RMSD simulation (point 1, Fig. 11).

The discrepancy between minimum skill and RMSD scores is exaggerated for the phytoplankton absorption field (Fig. 12). The minimum unbiased RMSD score, as would appear to be the best fit in a Taylor diagram, is also indicated (point 3, Fig. 12). These three model fields are presented against the reference field in Fig. 13. Evidently, the minimum unbiased RMSD model field is unacceptable due to the large positive bias. In contrast, the minimum RMSD (point 1, Fig. 12) and superior skill model fields (point 2; Fig. 12) are less biased but are out of phase with

Please cite this article as: Jolliff, J.K., et al., Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, Journal of Marine Systems (2008), doi:10.1016/j.jmarsys.2008.05.014

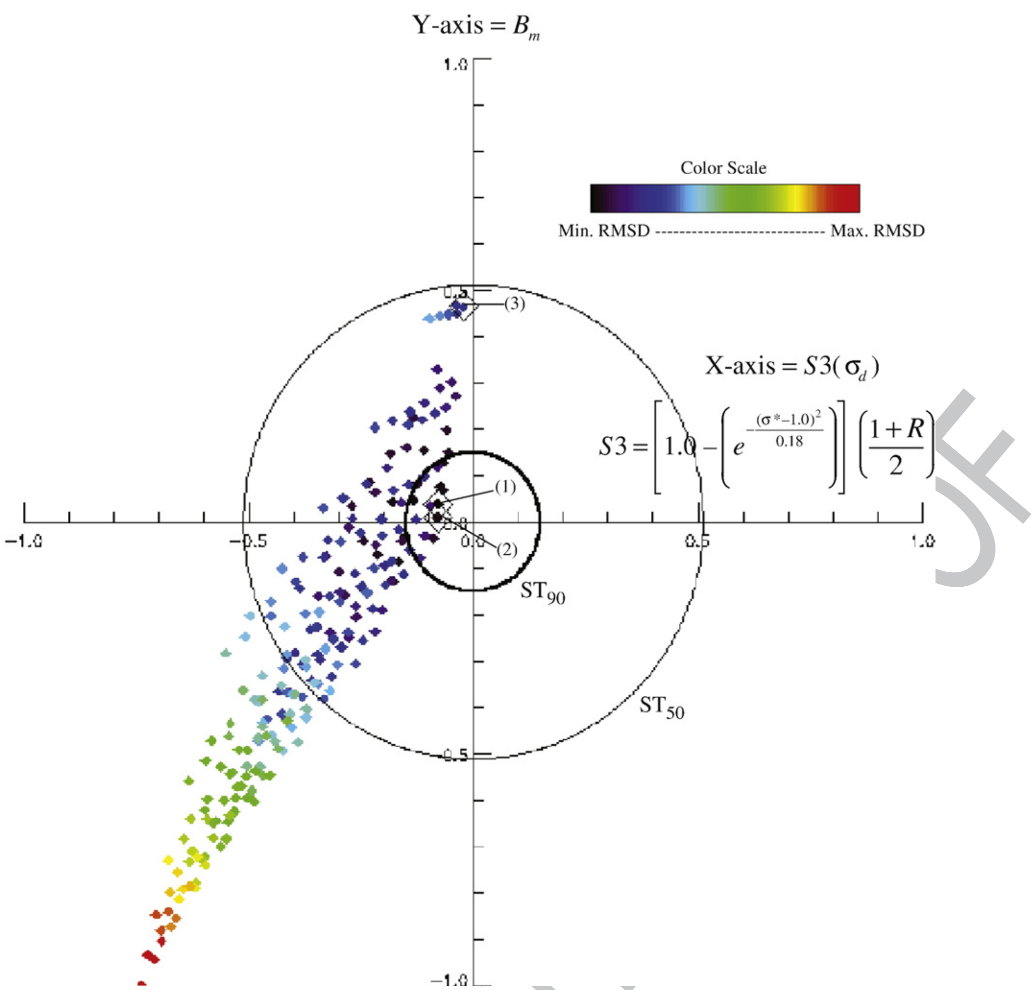

Fig. 11. Skill target diagram for model to reference chlorophyll-*a* field comparisons. The minimum total RMSD (1), minimum skill score (2), and minimum unbiased RMSD (3) are indicated on the plot. The markers indicate the 50th and 90th percentile total skill scores (ST) for the total set of model to reference comparisons, as explained in the text. The *X*-axis is the *S*3 skill score multiplied by the sign of the standard deviation difference. The *Y*-axis is the maximum normalized bias. The color scale indicates the total RMSD values.

the reference field by several months (Fig. 13). All three results provide information potentially useful to the investigator; other parameters may potentially be adjusted to either reduce the phase error for fields (1) and (2), or the bias may be reduced in (3), which is better correlated with the reference field. The salient point to be made here, however, is that for multiple model executions the skill target diagram may identify potential contrasts between minimum RMSD and other measures of model skill.

4. Discussion

An important point mentioned elsewhere in this special volume (Stow et al., submitted for publication) is worthy of reiteration here: different statistical quantities (i.e., skill metrics) may capture different aspects of model performance, and a thorough assessment of model skill may require use of multiple types of skill metrics simultaneously. Accordingly, it is important to recognize the relationships that exist between various statistical quantities and how they represent related but differentiable aspects of model performance. Linear correlation coefficients and variance comparisons help to identify similarities of pattern, and they may be combined in a way that is equivalent to the unbiased RMSD score (Eq. (7)), which succinctly quantifies pattern agreement. In our example of a one-dimensional time series, we related these aspects of model performance to the similarity of phase and amplitude between two time-dependent and sinusoidal-like patterns, but this concept may be generalized to describe the shape (such as the pattern of potential contour lines) of multidimensional property fields.

Pattern agreement is an important aspect of model performance, and there may be instances where this aspect is of particular or exclusive concern to the investigator. For example, Li et al. (2007) use Taylor diagrams to compare modeled and observed distributions of soil moisture and precipitation. Since the average values from the simulations were adjusted to agree with observed averages, the pattern information was

Please cite this article as: Jolliff, J.K., et al., Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, Journal of Marine Systems (2008), doi:10.1016/j.jmarsys.2008.05.014

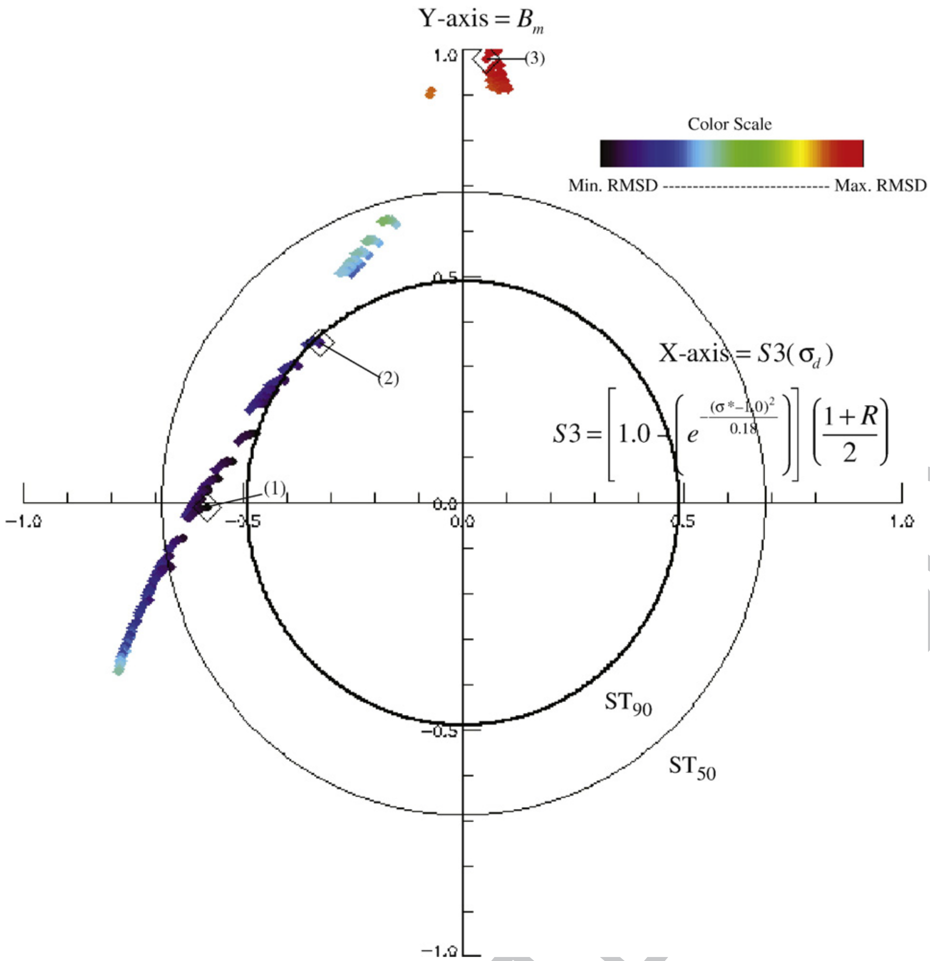

Fig. 12. Skill target diagram for model to reference phytoplankton absorption field comparisons. The minimum total RMSD (1), minimum skill score (2), and minimum unbiased RMSD (3) are indicated on the plot. The markers indicate the 50th and 90th percentile total skill scores (ST) for the total set of model to reference comparisons, as explained in the text. The *X*-axis is the *S*3 skill score multiplied by the sign of the standard deviation difference. The *Y*-axis is the maximum normalized bias. The color scale indicates the total RMSD values.

the primary aspect of interest from their climate model's performance. In such cases, Taylor diagrams are useful skill assessment tools insofar as they provide summary information about how the linear correlation coefficient and the variance comparisons each contribute to the unbiased RMSD on a two-dimensional diagram. Indeed, the pattern information may often be the primary area of interest for many climate model studies.

Nevertheless, in cases where the magnitude of the model results are not adjusted *a posteriori*, the usefulness of the Taylor diagram (and the statistical quantities it summarizes) as a skill assessment tool may be incomplete since it often provides no information about other aspects of model performance such as the bias (the comparison of mean values) or the total RMSD (a metric for overall model and data agreement). One way to remedy this omission is to modify Taylor diagrams via the addition of a color dimension indicating the magnitude of either the bias or the total RMSD. An example of this style of modification is given here and has been previously shown elsewhere (Orr, 2002).

More generally, however, information about the bias introduces the aspect of scale or magnitude to the model skill assessment process. For example, two surface chlorophyll fields may have a perfect correlation score and identical variances but the model field may still be an order of magnitude larger than the observations. This would suggest that too much nitrogen or carbon, for example, resides within the phytoplankton compartment and the ecosystem model may be inappropriately parameterized or structurally inadequate. In many ocean ecoystem (or biogeochemical) model applications, the time-dependent flux of materials from one reservoir to another may be constrained by the magnitude of the observations, rather than merely the pattern information. This is particularly pertinent to the biological aspects of coupled models because the overall magnitude of biological productivity is a critical aspect of ecosystem function. Furthermore, while the unbiased RMSD may effectively quantify pattern agreement, it is seldom used as a metric for overall model and data agreement, whereas the total RMSD is more frequently applied to this task.

Please cite this article as: Jolliff, J.K., et al., Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, Journal of Marine Systems (2008), doi:10.1016/j.jmarsys.2008.05.014

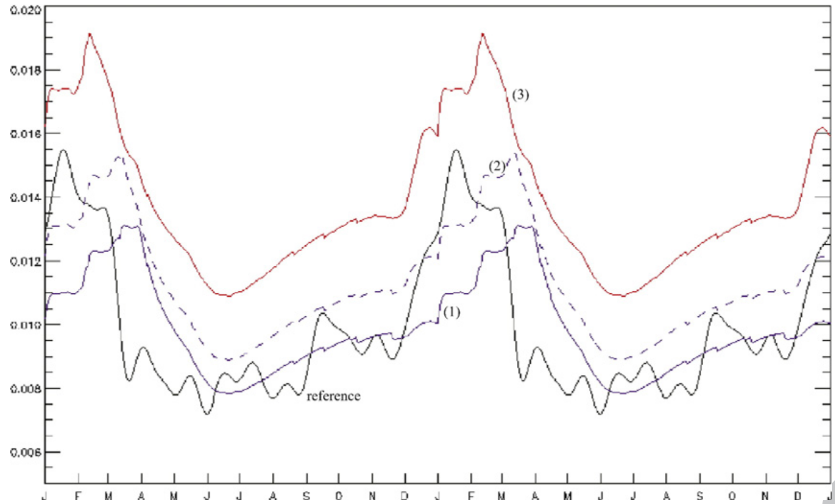

Fig. 13. The model and reference fields are plotted for the results indicated in Fig. 12: the minimum total RMSD (1), minimum skill score (2), and minimum unbiased RMSD (3; red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For these reasons, we have developed the target diagram, a Cartesian coordinate plot that provides summary information about how the magnitude and sign of the bias and the pattern agreement (unbiased RMSD) each contribute to the total RMSD magnitude. Markers may be added to the diagram in order to: (1) help identify limits based upon the correlation coefficient; (2) provide an assessment of model performance compared to an observational average (marker M_0) ; and (3) indicate potential limits to model performance improvement when the average observational uncertainty has been estimated. The observational uncertainty marker creates a "bull's-eye" for the target diagram that may very effectively communicate the estimated limits of model performance to other investigators.

For example, in our sensitivity analysis of grazing parameter selection, 216 model fields may be compared to three reference field categories for a total of 648 sets of model to reference field statistics. These may all be summarized on a single target diagram (Fig. 14). Cursory inspection of this summary diagram reveals that phytoplankton absorption is the most sensitive field and CDM absorption is the least. The phytoplankton absorption field is also prone to a large positive bias. The chlorophyll field appears to achieve the minimum magnitude for total difference statistics, but further improvement would be within the estimated range of average observational uncertainty.

To be sure, the purpose of both the Taylor and target diagrams is to compactly summarize statistical quantities that serve to aid in the skill assessment of model performance. The utility of either approach is dependent upon the aspects of model performance the metrics they summarize adequately capture. For the specific application to ocean ecosystem modeling, we suggest that target diagrams may better summarize the overall agreement between model and data since aspects of pattern agreement and magnitude (bias) are given equal weight and one may clearly visualize how they each contribute to the total RMSD.

It would be inappropriate, however, to suggest that skill assessment must always be implicitly synonymous with finding the lowest RMSD value amongst an ensemble of model results or an acceptably low RMSD values for a single model result. A potential deficiency in both the Taylor and target diagrams stems directly from a peculiarity of the RMSD metrics: the RMSD values may improve for correlations less than unity ($R<1.0$) where the normalized standard deviation is equal to the correlation ($\sigma^*=R$) instead of an ideal value of one ($\sigma^*=1.0$).

Another way to conceive of this behavior: if the correlation between a modeled and observed field is imperfect, i.e., in some areas the modeled values increase where or when the observed values decrease, then the average magnitude of this misfit may be reduced by diminishing the observed field's variance (assuming the bias is not a significant source of mismatch). For example, suppose a three-dimensional coupled model of phytoplankton growth and ocean circulation appears to adequately reproduce the observed details of chlorophyll patterns within a mesoscale eddy, only the eddy is in the wrong location when compared to the observations (a common type of mismatch for coupled models since modeled velocity fields are imperfect and advection is a time-integrative process). Given this spatial mismatch, the RMSD-based metrics of model/data misfit may improve if the details (i.e., the variance) of the modeled chlorophyll field are diminished or smoothed over. Would the investigator prefer a blurred modeled field over the one where the exclusive source of model/data disagreement appears to be dislocation?

This circumstance may be clearly demonstrated using satellite ocean color patterns from areas of complex mesoscale variability, such as Moderate Resolution Imaging Spectroradiometer data for the Mozambique Channel off the southwest coast of Madagascar (Fig. 15A). The complex pattern of apparent surface chlorophyll within mesoscale eddies and fronts (Fig. 15A) may potentially be mimicked by a coupled model, but imperfectly so with respect to spatiotemporal agreement. We approximate this kind of disagreement by reversing the array order (Fig. 15B) such that the hypothetical modeled field is effectively a mirror image of the data. The means and variances of the two fields are identical, but the correlation between them is quite low ($R=0.09$) and this

Please cite this article as: Jolliff, J.K., et al., Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, Journal of Marine Systems (2008), doi:10.1016/j.jmarsys.2008.05.014

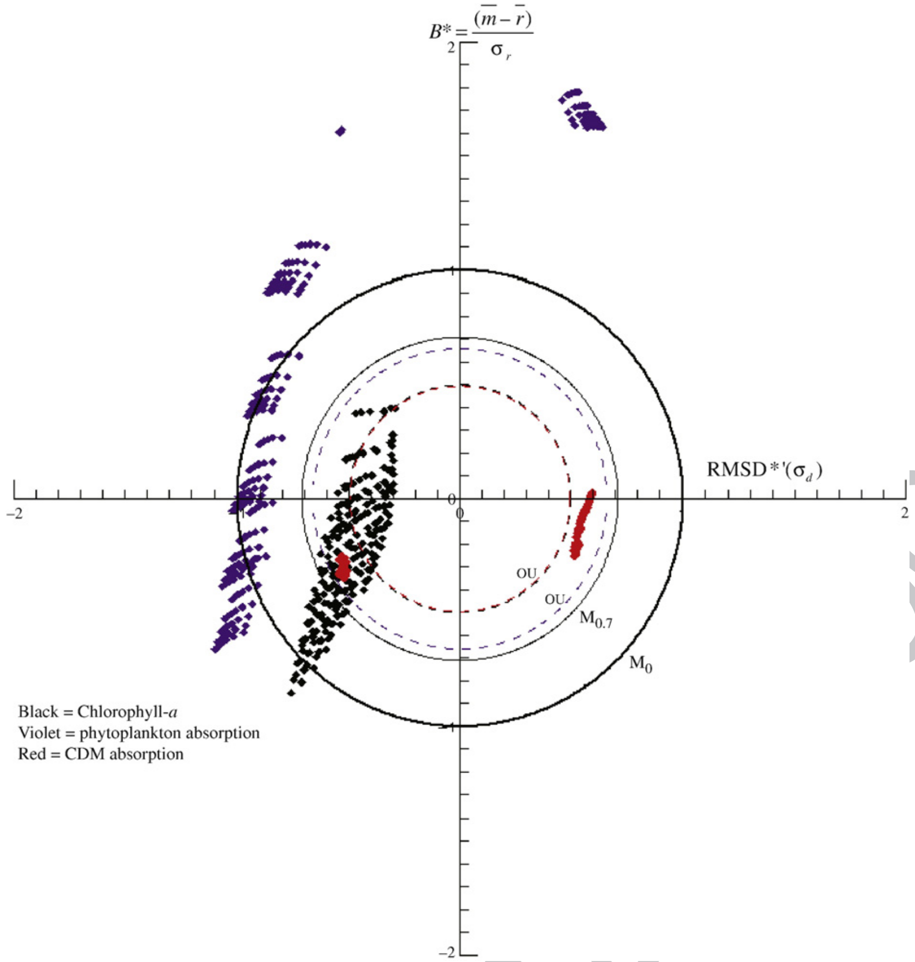

Fig. 14. Summary target diagram for all three types of model to reference field comparisons: chlorophyll-*a* (black), phytoplankton absorption (violet), and CDM absorption (red). The dashed lines indicate the estimated observational uncertainty (OU) threshold (corresponding to the field color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

results in high RMSD scores (RMSD*′=RMSD*=1.35). These scores may be artificially improved by simply reducing the variance of the hypothetical model field (Fig. 15C) until the threshold criterion $\sigma^*=R$ is met. As a result of this procedure, complex spatial details of the modeled chlorophyll field have been significantly diminished (Fig. 15B and C) yet the RMSD scores have certainly improved (RMSD*=0.99). Another way to demonstrate this property of RMSD-based metrics is to begin with the original field (Fig. 15A) and simply apply a large smoothing filter (Fig. 15D). Of the three hypothetical modeled fields (Fig. 15B,C, and D), one may be inclined to select B as the most skillful, though RMSD scores run contrary to this inclination.

Thus there are indeed cases where a distinction may be appropriately made between reducing RMSD statistics and increasing model skill. An alternative skill scoring system and skill target diagram was developed and presented for such a contingency. The advantage of this system is that for $R<1.0$ the minimum value skill score instead occurs where $\sigma^*=1.0$. In our example, the *S*3 skill score, Eq. (18), would indicate that field (B) is indeed the most skillful (Fig. 15). There are potentially many other creative ways to combine correlations, variances, and other metrics into composite skill scores that have properties distinctly different from RMSD-based metrics. Our intent is not to promote a specific solution but, rather, to point out that a contradiction may arise between minimum RMSD scores and other potential definitions of model skill.

In summary, model skill assessment ultimately requires specification about which quantitative metrics should be applied and how they should be interpreted to constitute "good" or "bad" model performance. The "skill" portion of skill assessment may be mathematically defined, but the "assessment" will invariably rely upon the value judgments of the investigator. Our analysis has focused upon some widely known statistical quantities (linear correlation coefficients, means, and variances) and ways that they may be combined mathematically and graphically to describe RMSD-based measures of model/data misfit. Taylor diagrams are polar coordinate plots that focus upon pattern agreement, whereas the target diagrams developed here summarize both the aspects of pattern agreement and magnitude (bias) and how

Please cite this article as: Jolliff, J.K., et al., Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, Journal of Marine Systems (2008), doi:10.1016/j.jmarsys.2008.05.014

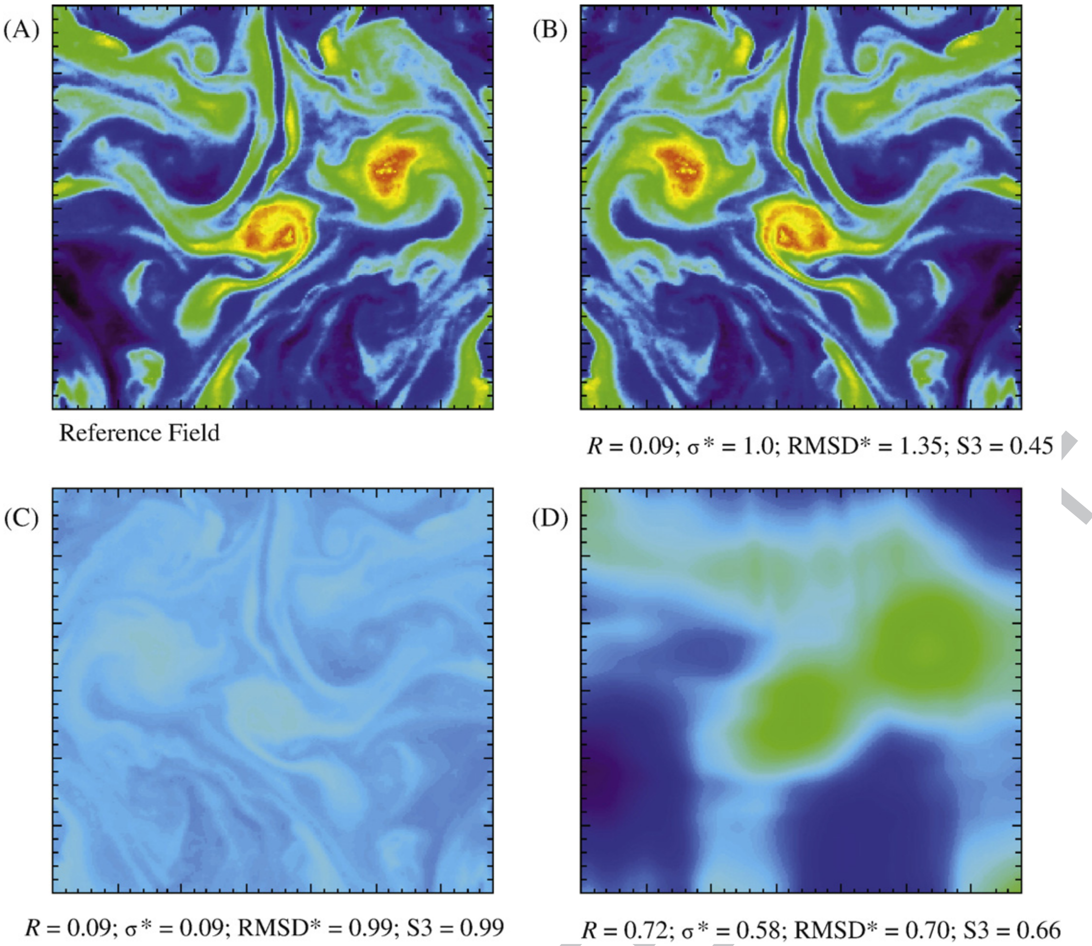

Fig. 15. A pattern of ocean color data is shown in panel A (surface chlorophyll fields; Moderate Resolution Imaging Spectroradiometer image 25 July 2007; data provided by NASA from their website at http://oceancolor.gsfc.nasa.gov/). To make a hypothetical model field wherein the misfit arises exclusively from spatial incoherence, the data array in (A) was reversed and is shown in panel (B) as a hypothetical modeled field. The resulting correlation is low but the mean and variance are the same. The field in panel (B) was further manipulated so that the normalized standard deviation (σ^*) is equal to the correlation coefficient ($\sigma^* = R$). This field is shown in panel (C). As a final comparison, the field in panel (A) was smoothed using a moving average filter. The correlation (R), normalized standard deviation (σ^*), normalized total root-mean-square difference (RMSD*), and skill score ($S3$) are shown beneath each panel for the comparison to the reference field (A). Panel (D) has the lowest RMSD* score and panel (B) has the lowest skill score.

they each contribute to the total RMSD, a common metric of overall model/data agreement. Investigators should be cognizant of the aspects of model performance summarized by each of these aforementioned statistical and graphical approaches before making claims of "model validation." Furthermore, both methods presume that RMSD-based metrics are sufficient criteria upon which to base model skill assessments, and this may not always be the case.

Acknowledgements

This research is a contribution to the Naval Research Laboratory 6.1 project, "Coupled Bio-Optical and Physical Processes in the Coastal Zone" under program element 61153N sponsored by the Office of Naval Research. This research was supported by the National Research Council's Post-Doctoral Research Associateship Program, and partially supported by the Office of Naval Research, grant number N0001405WX20735. Paul Martinolich provided assistance with SeaWiFS data processing and C. N. Barron and Clark Rowley provided assistance with the MODAS system. We also would like to thank two anonymous reviewers whose helpful comments certainly improved this manuscript.

References

Allen, J.I., Somerfield, P.J., Gilbert, F.J., 2007. Quantifying uncertainty in high-resolution coupled hydrodynamic-ecosystem models. Journal of Marine Systems 64, 3–14.

Bailey, S.W., Werdell, P.J., 2006. A multi-sensor approach for the on-orbit validation of ocean color satellite data products. Remote Sensing of Environment 102, 12–23.

Bretherton, F.P., Davis, R.E., Fandry, C.B., 1976. A technique for objective analysis and design of oceanographic experiments applied to MODE-73. Deep-Sea Research 23, 559–582.

Bricaud, A., Claustre, H., Ras, J., Oubelkheir, K., 2004. Natural variability of phytoplankton absorption in oceanic waters: influence of the size structure of algal populations. Journal of Geophysical Research 109, C11010. doi:10.1029/2004JC002419.

Fox, D.N., Teague, W.J., Barron, C.N., Carnes, M.R., Lee, C.M., 2002. The Modular Ocean Data Assimilation System (MODAS). Journal of Atmospheric and Oceanic Technology 19, 240–252.

Please cite this article as: Jolliff, J.K., et al., Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, Journal of Marine Systems (2008), doi:10.1016/j.jmarsys.2008.05.014

Franks, P.J.S., Chen, C., 2001. A 3-D prognostic numerical model study of the Georges bank ecosystems. Part II: biological-physical model. Deep-Sea Research II 48, 457–482.

Friedrichs, M.A.M., Dusenberry, J., Anderson, L.A., Armstrong, R.A., Chai, F., Christian, J.R., Doney, S.C., Dunne, J., Fujii, M., Hood, R., McGillicuddy, D.J., Moore, J.K., Schartou, M., Spitz, Y.H., Wiggert, J.D., 2007. Assessment of skill and portability in regional marine biogeochemical models: role of multiple planktonic groups. Journal of Geophysical Research 112, C08001. doi:10.1029/2006JC003852.

Q1 Friedrichs, M.A.M., Carr, M.-E., Scardi, M., Barber, R., submitted for publication. Assessing the uncertainties of model estimates of primary productivity in the tropical Pacific Ocean. Journal of Marine Systems.

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

Gruber, N., Frenzel, H., Doney, S.C., Marchesiello, P., McWilliams, J.C., Moisan, J.R., Oram, J.J., Plattner, G.-K., Stolzenbach, K.D., 2006. Eddy-resolving simulation of plankton ecosystem dynamics in the California Current System. Deep-Sea Research I 53, 1483–1516.

Holt, J.T., Allen, J.I., Proctor, R., Gilbert, F.G., 2005. Error quantification of a high resolution coupled hydrodynamic-ecosystem coastal ocean model: Part 1. Model overview and hydrodynamics. Journal of Marine Systems 57, 167–188.

Ivlev, V.S., 1961. Experimental Ecology of the Feeding of Fishes. Yale University Press, New Haven, Connecticut. 302 pp.

Jochens, A.E., DiMarco, S.F., Nowlin Jr., W.D., Reid, R.O., Kennicutt II, M.C., 2002. Northeastern Gulf of Mexico Chemical Oceanography and Hydrography Study: Synthesis Report. Technical Report, U.S. Department of the Interior, Minerals Management Service, Gulf of Mexico OCS Region, New Orleans, Louisiana. 586 pp.

Jolliff, J.K., Kindle, J.C., 2007. Naval Research Laboratory Ecological-Photochemical-Bio-Optical-Numerical Experiment (Neptune) Version 1: a portable, flexible modeling environment designed to resolve time-dependent feedbacks between upper ocean ecology, photochemistry, and optics. NRL Technical Memorandum, NRL/MR/7330-07-9026, Naval Research Laboratory, Stennis Space Center, Mississippi. 49 pp., http://stinet.dtic.mil/.

Kindle, J.C., DeRada, S., Arnone, R.A., Shulman, I., Penta, B., Anderson, S., 2005. Near real-time depiction of the California Current System. American Meteorological Society Sixth Conference on Coastal Atmospheric and Oceanic Prediction and Processes, San Diego, CA.

Lee, Z.P., Carder, K.L., Arnone, R.A., 2002. Deriving inherent optical properties from water color: a multiband quasi-analytical algorithm for optically deep waters. Applied Optics 41, 5755–5772.

Li, H., Robok, A., Wild, M., 2007. Evaluation of Intergovernmental Panel on Climate Change Fourth Assessment soil moisture simulations for the second half of the twentieth century. Journal of Geophysical Research 112, D06106. doi:10.1029/2006JD007455.

McClain, C., Hooker, S., Feldman, G., Bontempi, P., 2006. Satellite data for ocean biology, biogeochemistry, and climate research. EOS Transactions, American Geophysical Union 87, 337.

Millan-Nunez, E., Sieracki, M.E., Millan-Nunez, R., Lara-Lara, J.R., Gaxiola-Castro, G., Trees, C.C., 2004. Specific absorption coefficient and phytoplankton biomass in the southern region of the California Current. Deep-Sea Research II 51, 817–826.

Murphy, A.H., Epstein, E.S., 1989. Skill scores and correlation coefficients in model verification. Monthly Weather Review 117, 572–581.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, Part 1 — A discussion of principles. Journal of Hydrology 10, 282–290.

O'Reilly, J.E., Maritourena, S., Mitchell, B.G., Siegal, D.A., Carder, K.L., Garver, S.A., Kahru, M., McClain, C., 1998. Ocean color algorithms for SeaWiFS. Journal of Geophysical Research 103, 24937–24953.

Orr, J.C., 2002. Global Ocean Storage of Anthropogenic Carbon (GOSAC). Final Report (December 1, 1997 to March 31, 2001). EC Environmental and Climate Programme (Contract ENV4-CT97-0495). IPSL/CNRS, France. 116 pp.

Pacanowski, R.C., Philander, S.G.H., 1981. Parameterization of vertical mixing in numerical models of the tropical oceans. Journal of Physical Oceanography 11, 1443–1451.

Raick, C., Alvera-Azcarate, A., Barth, A., Brankart, J.M., Soetaert, K., Gregoire, M., 2007. Application of a SEEK filter to a 1D biogeochemical model of the Ligurgian Sea: twin experiments and real in-situ data assimilation. Journal of Marine Systems 65, 561–583.

Sheng, P., Kim, T., submitted for publication. Skill assessment of an integrated modeling system for shallow coastal and estuarine ecosystems. Journal of Marine Systems. Q2

Smith, K.W., McGillicuddy, D.J. Jr., Lynch, D.R., submitted for publication. Parameter estimation using an ensemble smoother: the effect of the circulation in biological estimation. Journal of Marine Systems. Q3

Stow, C.A., Jolliff, J.K., McGillicuddy, D.J. Jr., Doney, S.C., Allen, J.I., Rose, K.A., Wallhead, P., submitted for publication. Skill assessment for coupled biological/physical models of marine systems. Journal of Marine Systems. Q4

Stow, C.A., Roessler, C., Borsuk, M.E., Bowen, J.D., Reckhow, K.H., 2003. A comparison of estuarine water quality models for TMDL development in the Neuse River Estuary. Journal of Water Resources Planning and Management 129, 307–314.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research 106, 7183–7192.

Wallhead, P.J., Martin, A.P., Srokosz, M.A., Franks, P.J.S., submitted for publication. Predicting the bulk plankton dynamics of Georges Bank: model skill assessment. Journal of Marine Systems. Q5

Walsh, J.J., Weisberg, R.H., Dieterle, D.A., He, R., Darrow, B.P., Jolliff, J.K., Lester, K.M., Vargo, G.A., Kirkpatrick, G.J., Fanning, K.A., Sutton, T.T., Jochens, A.E., Biggs, D.C., Nababan, B., Hu, C., Muller-Karger, F.E., 2003. The phytoplankton response to intrusions of slope water on the West Florida shelf: models and observations. Journal of Geophysical Research 108 (C6), 3190. doi:10.1029/2002JC001406.

Please cite this article as: Jolliff, J.K., et al., Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, Journal of Marine Systems (2008), doi:10.1016/j.jmarsys.2008.05.014