HHS Public Access

Author manuscript

J Mar Syst. Author manuscript; available in PMC 2017 March 31.

Published in final edited form as:

J Mar Syst. 2009 February 20; 76(1-2): 4–15. doi:10.1016/j.jmarsys.2008.03.011.

Skill Assessment for Coupled Biological/Physical Models of Marine Systems

Craig A. Stow[a], **Jason Jolliff**[b], **Dennis J. McGillicuddy Jr.**[c], **Scott C. Doney**[d], **J. Icarus Allen**[e], **Marjorie A. M. Friedrichs**[f], **Kenneth A. Rose**[g], and **Philip Wallhead**[h]

[a]NOAA, Great Lakes Environmental Research Laboratory, 2205 Commonwealth Blvd., Ann Arbor, MI USA, 734-741-2055 (fax)

[b]Naval Research Laboratory, Stennis Space Center, MS USA, 228-688-4149 (fax)

[c]Woods Hole Oceanographic Institution, Woods Hole MA USA, 508-457-2194 (fax)

[d]Woods Hole Oceanographic Institution, Woods Hole MA USA, 508-457-2193 (fax)

[e]Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH UK, +44 1752 633101 (fax)

[f]Virginia Institute of Marine Science, College of William and Mary, P.O. Box 1346, Gloucester Point, VA USA, 804-684-7889 (fax)

[g]Dept. of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, LA USA, 225-578-6513 (fax)

[h]National Oceanography Centre, Southampton, UK, +44 2380 596485

Abstract

Coupled biological/physical models of marine systems serve many purposes including the synthesis of information, hypothesis generation, and as a tool for numerical experimentation. However, marine system models are increasingly used for prediction to support high-stakes decision-making. In such applications it is imperative that a rigorous model skill assessment is conducted so that the model's capabilities are tested and understood. Herein, we review several metrics and approaches useful to evaluate model skill. The definition of skill and the determination of the skill level necessary for a given application is context specific and no single metric is likely to reveal all aspects of model skill. Thus, we recommend the use of several metrics, in concert, to provide a more thorough appraisal. The routine application and presentation of rigorous skill assessment metrics will also serve the broader interests of the modeling community, ultimately resulting in improved forecasting abilities as well as helping us recognize our limitations.

Keywords

Goodness-of-fit; skill metric; skill assessment; model uncertainty

Correspondence to: Craig A. Stow.

1. Introduction

Quantitative models are widely used in the ocean sciences. Many applications are primarily heuristic; the models serve as “toys to tune our intuition” (Kaufman 1995) allowing users to conduct numerical experiments where real experimentation is infeasible. In these applications model predictions are regarded as testable hypotheses rather than explicit forecasts of future behavior. Thus, when model predictions are inaccurate, the cost of being wrong is low. In fact, erroneous predictions can be informative, affording opportunities for increased understanding of system behavior. But, increasingly models are used as tools to support decision-making, where the stakes can be high and the application of models with limited forecasting accuracy becomes a liability (Pilkey and Pilkey-Jarvis 2007). Particularly, in these high-stakes decision-support applications, information regarding model accuracy or “skill” is essential for decision-makers to consider when weighing forecasts and the possible outcomes of alternative actions.

Given a choice of models to evaluate future management scenarios, a decision-maker is likely to pick the most accurate model. If a model were available that was 100% accurate, this model would be preferable to one that was 75% accurate. With 100% accuracy management actions could be chosen based only on the societal value of the consequences of those actions. Though a model with only 75% accuracy is still informative, applying such a model requires hedging decisions by the relative probabilities of a range of possible outcomes and the societal value of those outcomes (Reckhow 1994). Hence, quantifying model skill provides information useful in both model selection and application.

The definition of model skill is dependent on context-specific factors such as the goals of the modeling exercise and the spatiotemporal scales of importance. Generally when we assess skill we are asking: How well does the model represent truth over a specified range of conditions? However, because truth cannot be measured, we use observations as a surrogate and ask instead: How well does the model fit the data? Both our model predictions and the observations reside in a halo of uncertainty and the true state of the system is assumed to be unknown, but lie within the observational uncertainty (Figure 1a). A model starts to have skill when the observational and predictive uncertainty halos overlap, in the ideal case the halos overlap completely (Figure 1b). Thus, skill assessment requires a set of quantitative metrics and procedures for comparing model output with observational data in a manner appropriate to the particular application. The residual (or misfit) is defined as the difference between the observation and the prediction, and most of the metrics described in this paper are some function of this quantity.

The routine application of rigorous skill assessment techniques is not broadly reflected in the refereed literature. Arhonditsis and Brett (2004) compiled a comprehensive review of 153 aquatic biogeochemical models published from 1990–2002 and found that ~ 30% of the studies reported goodness of fit measures, often a time-series plot of observations vs. model predictions, while ~47% reported some form of model validation. A possible reason for the relatively low skill assessment rate is that consumers of this information (mostly fellow research scientists) seem little affected by the presence or absence of skill information; a follow-up analysis (Arhonditsis et al. 2006) reported no relationship between the level of

skill assessment presented or the accuracy of the model, and the subsequent citation rate of the published paper.

Similarly, we reviewed 142 papers published in five oceanographic journals (Journal of Geophysical Research – Oceans, Deep Sea Research I and II, Journal of Marine Systems, Journal of Oceanography, and Ocean Modeling) between January 1, 2000 and March 31, 2007. We selected only articles presenting ecological or biogeochemical models coupled to a model describing a physical process—in most cases a one or three-dimensional hydrodynamic model. Papers wherein the physical coupling was not explicit (i.e., the 0-dimensional model studies) and papers wherein any direct comparison between model results and observations was absent were excluded. These entries were further sorted by the type of model to observation comparisons made, and emphasis was placed on validation metrics used for the ecological/biogeochemical variables.

Most papers (68.3 %) provided only a basic comparison of model results and observations, usually a visual comparison and occasionally a comparison of ranges, means, and variances. Some of these papers used language such as "*reasonable*" or "*strong similarity*" and "*does a good job in reproducing patterns observed…*" While these statements are consistent with the evidence presented by the authors making them, Allen et al. (2007a) demonstrated that there is no scientific and objective consensus as to what constitutes a "good fit" when model results and observations are visually compared.

Thirteen papers (9.2 %) quantified model and observation misfits (residuals) using linear correlation and difference statistics. An additional 11.3 % of papers reviewed involved data assimilation techniques and summarized model and data misfits using a cost function. Cost functions generally sum the weighted, squared differences between modeled and observed fields over all variables for which data are available. The remaining papers employed various comparison schemes and metrics that ranged from multivariate correlations and scaling techniques (Allen et al., 2002; Allen et al., 2007b) to a comparison of fast Fourier transformations (Powell et al., 2006).

Hence the only model to data comparison metric that is demonstrably the community standard is the basic visual comparison. However, when model predictions bisect a cloud of observations, but fail to mimic the scatter of the data, does this constitute a good fit? Modelers with differing applications and perspectives will offer divergent opinions. Clearly, more specific and quantitative techniques are appropriate, though they may be difficult to prescribe, generally, due to differences in the types of data to which the models are compared, and differences in temporal and spatial scales of comparison. Nevertheless, as the biological-physical modeling community moves to embrace data assimilation techniques (reviewed by Gregg et al. [2007, this issue]) and the stakes contingent on model predictions increase, the presentation of standardized skill metrics, as the OSPAR Commission has recommended (Villars et al. 1998), will become increasingly important.

Herein, we highlight multiple misfit metrics and skill assessment methods, useful for a range of biological-physical modeling applications. First, we examine the simple case of comparing model results for a single prognostic variable with corresponding observations of

same, i.e., univariate comparisons. Second, we present cost functions as a compact method to summarize model performance when multiple types of prognostic variables are compared with corresponding types of observations. Cost functions are also distinct from a collection or summation of univariate metrics in the sense that estimates of the observational error are included in their formulation. We then highlight some additional methods that may be of service to marine ecosystem modelers. Specifically, we discuss ways to quantify patterns between multiple sets of variables (multivariate pattern evaluation), some additional methods to quantify the comparison of modeled and observed spatial maps, and a potentially useful way to quantify the predictive capacity of a model–the Binary Discriminator Test.

This is not an exhaustive list, and many important, closely related topics such as uncertainty analysis (Beck 1987), model selection (Kass and Raftery 1995), model averaging (Hoeting et al. 1999), and scores for probabilistic forecasts (Brier 1950, Katz and Ehrendorfer 2006) are not addressed. Rather, this is an attempt to call attention to some useful skill assessment methods and point out a few that can be misleading. We have chosen not to be overly prescriptive, in the belief that some experimentation and vetting must occur for the most informative metrics to “rise to the surface” and become widely employed. We offer these metrics as a challenge to the community to include their use and presentation as a routine part of model development, publication, and application.

2. Univariate Comparison of Predictions and Observations

Graphically comparing model point predictions with observations can be a useful way to assess model performance. While time-series plots of observations and model predictions seem to be the community standard, bivariate plots of observations vs. predictions are usually more revealing. Additionally, bivariate observed vs. predicted plots can be complemented with supporting quantitative measures such as simple linear regression statistics (Reckhow et al. 1990, Smith and Rose 1995).

Another useful graphical approach is to evaluate the set of differences between corresponding observations and predictions, variously referred to as “misfits” (Evans 2003) or “residuals”. In statistical texts, residual examination is typically the first step to corroborate underlying probabilistic assumptions such as normality and independence of the model error term. However, even if the model is not explicitly contingent on such assumptions, graphical examination of residuals along a logical gradient, such as time, space, or vs. model predictions can reveal systematic biases or a differing ability of the model to capture variability in some regions of space or time (Friedrichs et al., 2007; this issue).

In addition to graphical techniques, there are many simple, quantitative metrics that are useful to assess model skill. Stow et al. (2003) used the six following indices in a side-by-side comparison of three estuarine water quality models of differing complexity:

1. r – the correlation coefficient of the model predictions and observations:

$$r = \frac{\sum_{i=1}^{n}(O_i - \overline{O})(P_i - \overline{P})}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2 \sum_{i=1}^{n}(P_i - \overline{P})^2}},$$

2. RMSE - the root mean squared error (also referred to as root mean squared difference):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}},$$

3. RI – the reliability index

$$\mathrm{RI} = \exp\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{O_i}{P_i}\right)^2},$$

4. AE – the average error (bias)

$$\mathrm{AE} = \frac{\sum_{i=1}^{n}(P_i - O_i)}{n} = \overline{P} - \overline{O},$$

5. AAE – the average absolute error

$$\mathrm{AAE} = \frac{\sum_{i=1}^{n}|P_i - O_i|}{n}, \text{and}$$

6. MEF – the modeling efficiency

$$\mathrm{MEF} = \frac{\left(\sum_{i=1}^{n}(O_i - \overline{O})^2 - \sum_{i=1}^{n}(P_i - O_i)^2\right)}{\sum_{i=1}^{n}(O_i - \overline{O})^2},$$

where n = the number of observations, O_i = the ith of n observations, P_i = the ith of n predictions, and $\bar{O}$ and $\overline{P}$ are the observation and prediction averages, respectively.

The correlation coefficient, r, measures the tendency of the predicted and observed values to vary together. It can range from −1 to 1, with negative values indicating that the observed and predicted values vary inversely. Ideally, this value will be close to one. However, even if the correlation is near one, the predicted and observed values may not match each other;

they could differ by a consistent factor. Additionally, this measure can be dominated by a small proportion of extreme values that may not reflect the behavior of the bulk of the data.

The root mean squared error, average error, and average absolute error are all measures of the size of the discrepancies between predicted and observed values. Values near zero indicate a close match. The average error is a measure of aggregate model bias, though values near zero can be misleading because negative and positive discrepancies can cancel each other.

The average absolute error and the root mean squared error both accommodate the shortcoming of the average error by considering the magnitude rather than the direction of each discrepancy. Together these three statistics provide an indication of model prediction accuracy.

The reliability index (Leggett and Williams 1981) quantifies the average factor by which model predictions differ from observations. For example, an RI of 2.0 indicates that a model predicts the observations within a multiplicative factor of two, on average. Ideally, the RI should be close to one. When the root mean squared error has been calculated for log transformed values of the predictions and observations, then the RI is the exponentiated RMSE.

The modeling efficiency measures how well a model predicts relative to the average of the observations (Nash and Sutcliffe 1970, Loague and Green 1991). It is related to the RMSE according to: $MEF = 1\text{-}RMSE^2/s^2$ where s^2 is the variance of the observations. A value near one indicates a close match between observations and model predictions. A value of zero indicates that the model predicts individual observations no better than the average of the observations. Values less than zero indicate that the observation average would be a better predictor than the model results.

All of these univariate statistics are sensitive to phase errors, in either time or space, in the model predictions relative to the observations. For one-dimensional data sets, one can compute the lagged model-data correlations (or RMSE, RI, etc.). For two-dimensional data sets, some groups have used empirical orthogonal function (EOFs). Graphical comparisons and lagged correlation analysis can be made for observed and predicted EOF spatial maps and their associated principal component time series (Doney et al., in press).

Because they capture different aspects of model performance, it is often useful to use several metrics simultaneously for a thorough skill evaluation. Sometimes it is appropriate to log-transform the observations and predictions before calculating goodness-of-fit statistics so that differences between predicted and observed values will not be highly skewed and dominated by a small proportion of high values.

To illustrate these metrics we compared model derived sea surface temperature (SST) and mixed layer depth (MLD) from a one-dimensional upper ocean model simulation (Doney, 1996) with ship-based CTD data collected at the Bermuda Atlantic Time-Series Site (BATS) in the Sargasso Sea. The BATS station was occupied on a biweekly to monthly time resolution, and within each cruise anywhere from one to more than a dozen CTD casts were

conducted. The data thus include considerable high frequency data (diurnal cycle, internal tides, small-scale spatial heterogeneity) not captured by the model.

Based on the time-series plots (Figures 2 and 3, top left), both sets of model predictions and observations show a *strong similarity*, but the accompanying plots and metrics offer a more comprehensive evaluation. Though the model *does a good job in reproducing observed SST patterns* (Figure 2, top left), it tends to underestimate SST after the middle of 1996 (Figure 2, top right) and for SSTs above about 28 deg. C (Figure 2, bottom left). The diagonal clusters in the model misfit versus observed values (Figure 2, bottom right) reflect data from individual cruises when there is a large spread among different observations but little variation in the model over the short duration of the cruise. Summary statistics (Table 1) show a high correlation between the observed and predicted values, the RMSE, AE, and AAE are all relatively small in comparison to the variability of the data, the RI is close to one and the MEF is fairly close to one. However, the intercept and slope of the predicted vs. observed plot (Figure 2, bottom left) would be judged to be "significantly different from zero and one, respectively, in a classical statistical analysis. The model-data agreement for MLD is *reasonable* (Figure 3, top left) during the summer when MLD is shallow, however tends to be poor during the winter (Figure 3, top right), when small phase shifts in the simulated MLD lead to large misfit values. The predicted vs. observed plot (Figure 3, bottom left) reveals discrepancies more clearly than the time-series plot (Figure 3, top left), showing considerable variability. The patterns apparent in the misfit vs. observation plot (Figure 3, bottom right) are also indicative of individual cruise data. Summary statistics (Table 1) are generally less favorable for MLD than for SST with a lower correlation while the RMSE, AE, and AAE are each relatively large. The MEF is near zero, indicating minimal predictive ability, and the intercept and slope estimates differ from the desired values of zero and one, respectively. Thus, while "*reasonable*", "*strong similarity*", and "*does a good job…"* would probably go unchallenged, additional probing helps reveal the veracity of these assessments.

There are also compact techniques to display potentially large sets of univariate statistics on summary diagrams. For example, Taylor (2001) described a method to exploit relationships between variance, correlation, and RMSD statistics in order to display these quantities on a single summary diagram, i.e., the Taylor diagram. These diagrams have begun to appear in the coupled model literature as a convenient way to quantify and communicate model performance to both modelers and non-modelers as the model is modified or aspects of model output are delineated by variable type or geographic region (e.g., Gruber et al. 2006; Raick et al., 2007). More detailed presentations of the Taylor diagram and other compact methods useful to graphically convey information are also found in other papers in this special volume (Jolliff et al., 2007; Friedrichs et al., 2007b).

3. Multivariate Comparison of Predictions and Observations

For models with multiple response variables, independent, univariate comparisons of each response with its corresponding observations may still be informative, but it is often appropriate to compare responses and observations across all of the response variables simultaneously (Friedrichs et al., 2006). A cost function, J, is a single metric of overall

model performance defined for applications such as objective analysis and data assimilation, where attempts are made to minimize model-data misfit against some set of observations (e.g., Wunsch, 1996; Kasibhatla et al., 2000; Kalnay, 2003). Cost functions combine the model-data misfit across incommensurate variables with differing units and uncertainties, and thus are also useful for characterizing the overall misfit across a suite of model simulations.

The most straightforward cost function is the weighted sum of squares of individual, point to point model data misfits:

$$2\mathrm{J}(\mathbf{x}_\mathrm{P})=[\mathbf{x}_\mathrm{P}-\mathbf{x}_\mathrm{O}]^\mathrm{T}\mathbf{R}^{-1}[\mathbf{x}_\mathrm{P}-\mathbf{x}_\mathrm{O}] \quad (1)$$

$\mathbf{x}_\mathrm{O}$ and $\mathbf{x}_\mathrm{P}$ are vectors of length n of the observations and corresponding model prediction values for all variables at all available points in time and space, the superscript T refers to the transpose of a vector, and $\mathbf{R}^{-1}$ is the inverse of the $n\times n$ error covariance matrix. The form of the cost function in Equation 1 is equivalent to a weighted sum of squares of model-data misfits and thus is a generalization of the RMSE. This form can be derived from both maximum likelihood and Bayesian approaches, for the case where the model-data misfits in **R** are normally distributed.

Much of the art in the construction of the cost function involves developing the covariance matrix **R** that weights the contributions of individual data points to the total cost function. If the misfits are independent, as is commonly assumed, the off-diagonal terms in **R** are zero, the diagonal elements of **R** can be estimated using the misfits, denoted $\boldsymbol{\varepsilon}_\mathrm{i,}$ as $\mathrm{R_{ii}}=\boldsymbol{\sigma}^2_\mathrm{ii}=\boldsymbol{\varepsilon}_\mathrm{i}^\mathrm{T}\boldsymbol{\varepsilon}_\mathrm{i}$, and the elements of the inverse are $\mathrm{R}^{-1}_\mathrm{ii}=1/\boldsymbol{\sigma}^2_\mathrm{ii}$. The $\boldsymbol{\varepsilon}_\mathrm{i}$ may represent observation error, model error or both. In some cases “errors of representativeness” may be included to account for the presence in the observations of subgrid-scale variability that is not captured at the grid-scale of the model (Kalnay, 2003). Off-diagonal elements can arise when the observational data contain regional or temporal biases that are correlated across observations. Note that we are discussing here correlations among the observation errors, not correlations in the observations themselves.

The form of the cost function, is identical (barring the conventional factor of 2 in J) with the chi-squared statistic (Press et al., 1986; Bevington and Robinson, 2002):

$$\chi^2=\sum\nolimits_\mathrm{i}(\mathrm{P_i}-\mathrm{O_i})^2/\varepsilon_\mathrm{i} \quad (2)$$

and the related quantity the reduced chi-squared:

$$\chi^2_\upsilon=1/\upsilon\sum\nolimits_\mathrm{i}(\mathrm{P_i}-\mathrm{O_i})^2/\varepsilon_\mathrm{i} \quad (3)$$

where υ is the number of degrees of freedom in the observations. The reduced chi-squared metric would have a value of about 1 if the model fit the observations within about the

observational error and if all of the data were independent. Values of χ^2_υ significantly greater than 1 indicate that the model is a poor fit to the observations, and there are statistical tests to assess the model goodness of fit (Press et al. 1986). Spatial, temporal, and variable-variable correlations in the observations and model results lower the number of degrees of freedom υ, which can be quantified using autocorrelation and cross-correlation estimates (Emery and Thomson 1998).

In some situations, the model does not directly predict the quantity that is observed, and the model variables need to be transformed using an observation operator H($\mathbf{x}_{\mathrm{mod}}$). The cost function J would then be written:

$$2\mathrm{J}(\mathbf{x}_\mathrm{P})=[\mathrm{H}(\mathbf{x}_\mathrm{P})-\mathbf{y}_\mathrm{O}]^\mathrm{T}\mathbf{R}^{-1}[\mathrm{H}(\mathbf{x}_\mathrm{P})-\mathbf{y}_\mathrm{O}] \quad (4)$$

where the observations are denoted as $\mathbf{y}_\mathrm{O}$ as a reminder that they are different quantities than in the model. In some formulations, H is used to denote the fact that the model points compared with the observations are a sub-sample of the model state space. More interestingly, the need for an observation operator can arise, for example, when comparing model and observed spatially or temporally integrated quantities (e.g., vertically integrated primary production) or for quantities that are not directly predicted by the model but which can be diagnosed from model variables (e.g., acoustic backscatter, biooptical properties). There may be additional error terms that need to be added to **R** associated with the observation operator H.

The cost function J is not limited strictly to point to point comparisons, and additional terms can be added to equation 1 to reflect model skill with aggregate model behavior and model patterns with respect to observations:

$$2\mathrm{J}(\mathbf{x}_\mathrm{P})=[\mathbf{x}_\mathrm{P}-\mathbf{x}_\mathrm{O}]^\mathrm{T}\mathbf{R}_\mathrm{X}{}^{-1}[\mathbf{x}_\mathrm{P}-\mathbf{x}_\mathrm{O}]+[\mathbf{z}_\mathrm{P}-\mathbf{z}_\mathrm{O}]^\mathrm{T}\mathbf{R}_\mathrm{z}{}^{-1}[\mathbf{z}_\mathrm{P}-\mathbf{z}_\mathrm{O}] \quad (5)$$

For example, the vector **z** could include terms related to the model-data misfit in the total flow of a current through a section, the integrated biological production for a basin, or a biological diversity index for an ecosystem, irrespective of the exact agreement of the model and observed patterns. Regularization terms can also be added to the cost function to express prior knowledge about the nature of the solution, such as imposing smoothness constraints by penalizing gradients in predicted fields.

4. Multivariate Pattern Evaluation

Univariate and multivariate metrics are useful measures to summarize model skill. However, considerable information can be lost when complex multivariate information is reduced to a single numerical index. Multivariate approaches that allow the simultaneous examination of the ways in which numerous variables vary in relation to each other spatially and temporally are also helpful to evaluate model skill. Marine ecologists commonly use these approaches to interpret complex data sets and marine ecosystem modelers are beginning to use them to

investigate patterns and modes of variability in model outputs (Allen et al., 2002; Blackford et al., 2004, Schrum et al., 2006, Allen et al., 2007a, Allen and Clark, 2007).

If we have a set of multivariate observations available for model validation we can subject them to multivariate analysis. If we then reconstruct a data set from the model by taking the nearest equivalents in space and time, we can subject them to the same analysis and compare the results. By definition, if the observations are the truth then the perfect model should exactly reproduce the observed multivariate patterns. Multivariate analysis allows us to explore complex relationships by reducing the dimensionality of the problem. Allen and Somerfield. (this volume) have demonstrated the applicability of a range of techniques (Principal Component Analysis (PCA, e.g. Chatfield and Collins 1980), Multi Dimensional Scaling (MDS e.g. Clark 1993) and cluster analysis e.g., Clark and Gorley 2006) and shown that the dimensions of the problem can be reduced and multivariate and univariate goodness of fit measures, in terms of both magnitude and trend determined.

5. Binary Discriminator Tests

This is a class of tests which assess the predictive power of a binary classification system to evaluate how useful a model is in a decision-making process. These tests can reveal the following about a model: a) whether or not the fit between model and observations is better or worse than we would obtain if the model was replaced with a random number generator, and b) how well it quantifies skill as a function of threshold using a binary discriminator, i.e. if an algal bloom is defined as being above a certain concentration of chlorophyll, what is the probability that our model predicts a bloom?

The best known example is the Receiver Operator Characteristic (ROC) devised during the Second World War for radar operators to correctly differentiate hostile and friendly aircraft. These techniques are now widely used in a number of fields, particularly medical research. Brown and Davis (2006) provide a detailed and accessible tutorial of the use of ROC curves and related metrics. We outline the methods below, following the nomenclature of Brown and Davis (2006).

At the heart of the test is a simple yes/no decision, based on the comparison of two independent information sets (in our case observations and model) with respect to a threshold value. Each trial has four possible outcomes, either correctly positive (CP), correctly negative (CN), incorrectly positive (IP) and incorrectly negative (IN), these are also known as Type 1 and Type 2 errors (Figure 4a). We can use this approach to make an analysis of similarity of how well the model the fits the data. The perfect model is one where all the points in a scatter diagram of model vs. data lie on the x = y line (Figure 4a). If we set a threshold criterion (T_D) dividing the data into two sets and then compare it with the model using the same threshold (T_M, Figure 4a) we can assess model data similarity at that threshold, effectively assessing the model ability to discriminate that threshold. The perfect model will only give CP and CN outcomes; the more scatter there is in the model-data relationship the more IP and IN conditions will occur and the worse the model performance. Because we are interested in model performance we want to assess how well the model resolves the data across the whole range of data. By allowing T_D to co-vary with T_M, we

obtain a non-parametric measure of the model's ability to simulate a given variable, which can be compared directly with other simulated variables. The decision process can be further assessed by calculating the correct negative fraction (CNF) and the correct positive fraction (CPF).

$$CNF=\frac{CN}{CN+IP} \quad \text{(a)}$$

$$CPF=\frac{CP}{CP+IN} \quad \text{(b)}$$

CNF and CPF are independent of the actual numbers of positive and negative events in the trials and express the fraction of negative and positive events, which are correctly determined. A curve which illustrates model performance can then be calculated by plotting CPF_i on the vertical axis and 1-CNF_i on the horizontal axis for i=1, k threshold values (Figure 4b). These values are sometimes referred to as the sensitivity and specificity, where the sensitivity (CPF) is the probability that case X classified correctly as above the threshold and the specificity (1-CNF) is the probability that X classified correctly as below the threshold. The perfect model corresponds to a point in the top left hand corner of the Y axis (i.e. CNF = 1 and CPF = 1), the top right (CPF=1, CNF=0) and bottom left (CPF = 0 and CNF = 1) of the diagram correspond to the extremes of the decision process where every trial is always deemed either positive or negative. A completely random predictor (by definition CP = IP and CN = IN) gives a straight line at an angle of 45° from the horizontal. This is because as the threshold rises equal numbers of true and false positives occur. Results below this line suggest the model gives consistently incorrect results.

Decisions based on CPF and CNF are estimators of probabilities of decisions contingent on events: if a positive event has occurred what is the probability I will make the correct decision. While these probabilities are useful they do not address the fundamental question, if I make a positive decision what's the probability that the decision is correct. The positive predictive value (PPV) and negative predictive value (NPV) can be expressed as (see Brown and Davis (2006) for the theoretical background and derivation).

$$PPV=\frac{CP}{CP+IP} \quad \text{(c)}$$

$$NPV=\frac{CN}{CN+IN} \quad \text{(d)}$$

Values of PPV and NPV can range between 0 and 1, reflecting the intrinsic power of the decision; high values indicating a decision can be trusted, low values suggesting the decision should be regarded with skepticism.

As an illustration some examples are shown in Figure 5. Employing the ROC technique, Figure 5 indicates that the model has some predictive skill for both temperature and chlorophyll. Unsurprisingly temperature (Figure 5, top) shows a very high skill level while chlorophyll-a, has limited skill at low concentrations. These are confirmed by the respective pearson correlation scores for each data set (T, r = 0.96, Chl r = 0.24, Allen et al 2007b). Figure 5c,d shows the probabilities that a positive or negative decision is correct at a particular threshold for temperature and chlorophyll. Temperature (Figure 5c) is clearly the most reliable variable, with a greater than 90% probability that both positive and negative decisions are correct over the range 8–16 °C. For chlorophyll the negative predictive values are in excess of 0.9 over substantial ranges of the data range, but the ability to discriminate a positive event is poor, if the chlorophyll concentration is above 1mg m^{-3}, effectively indicating this simulation is poor at predicting bloom events. Following from this, if we have a large spatio-temporal data set (e.g. satellite ocean color chlorophyll) we can plot of map of the model skill at predicting algal blooms (Allen et al in press).

6. Comparison of Spatial Maps

Evaluation of many hydrodynamic and biogeochemical models involves comparison of model-generated and data-derived spatial maps of key variables (e.g., water velocities, chlorophyll-a). The spatial maps are often presented on the x and y (latitude and longitude) dimensions with the continuous variable of interest as the height (z-variable), or the variable of interest categorized into intervals that are color-coded (e.g., Hashioka and Yamanaka 2007; Wiggert et al. 2006). Three-dimensional models (i.e., include a vertical dimension) are reduced into the x and y dimensions of a map by taking a slice in the vertical dimension (e.g., fixed depth interval) or by integrating over the water column. The presentation of these side-by-side spatial maps are often accompanied by statements such as “the model captures the major features” and other visually-oriented qualitative statements. With the increasing use and application of coupled biophysical models, there is a clear need to formalize the comparison between model and data spatial maps.

Fortunately, quantifying the patterns in spatial maps, and the question of how to compare two or more spatial maps, have been long-term problems inherent in the fields of image analysis (Foody 2002), pattern recognition (Duda et al. 2001), and landscape ecology (Gustafson 1998); however, the bad news is that, despite the great interest and effort, the problem has not been completely solved. Typically, the focus of comparisons in marine ecosystem modeling is the appearance of specific features or patterns in the model and data maps, such as areas of high mixing, nutrient gradients, and patches of high phytoplankton concentrations. The major difficulty is that the model map may resemble the data map but with the features of interest offset slightly in the x or y directions, rotated, or compressed or dilated. The challenge for skill assessment is determining at what point does one say the two maps are similar or different, and how does one quantify how similar the two maps are in an objective manner?

Approaches for comparing two spatial maps can be broadly grouped into those that compare composition only (e.g., frequency histogram of cells binned by their magnitude), and those that also include the degree of agreement in configuration (i.e., include agreement of the spatial arrangement). Approaches in the later broad category that include configuration can be divided into those that are based on a pure cell-by-cell comparison, those that allow for some fuzziness or relaxation of the strict cell-by-cell comparison (i.e., looks at nearest neighbors or moving windows of cells to see how well the maps agree), and those that compare higher-order properties (e.g., fractal dimension) between the maps.

We highlight a few of the many possible approaches for comparing spatial maps in order to raise awareness in the oceanographic community that quantitative methods exist for comparing spatial maps. One commonly used approach is to compute measures of misfit or residuals between predicted and observed values cell-by-cell, and then display the misfit on the same spatial grid as a misfit or difference map. An example of a statistical approach is the Kappa statistic that uses the classification error (or confusion) matrix to determine the percent improvement in the agreement between the two maps on a cell-by-cell basis over that would be expected by randomness (Lillesand et al. 2004). There exists a fuzzy variation of the Kappa test statistic that allows for the information in neighboring cells and for differences in the similarities between adjacent categories in the map's legend to count towards the fit of the model map to the data map (Hagen-Zanker et al. 2005). Methods for map comparison can become quite complicated, such as the approach of Fewster and Buckland (2001) who use a windowing approach and compute the switches needed (mutations) of the variable of interest between cells within each widow on one map to get it to agree as close as possible with the other map. Higher-order properties try to capture features of the spatial heterogeneity (composition and configuration), such as the fractal dimension, lacunarity (Gustafson 1998), and the state probability function that is a categorical variable version of a variogram (Phillips 2002). Rose et al. (this issue) compare several of these approaches using maps generated with known features. Rose et al. (this issue) include a variation on the cell-by-cell comparison borrowed from multivariate statistics (procrustes analysis, Krzanowski 1990) that allows for the model-generated map to be rotated, dilated, and shifted relative to the data map to determine what adjustments are needed to get the model map as close as possible to the data map.

7. Conclusions

The continuing development of deployed observing systems such as moored arrays and autonomous underwater vehicles will provide scientists with a new wealth of data that may potentially be used to constrain and evaluate model performance. These advances in observing systems, computational power, and the now frequent practice of coupling three-dimensional hydrodynamic models with complex ecosystem models will also provide new opportunities to test hypotheses regarding the structure and function aquatic ecosystems. Simultaneously, these complex and potentially powerful modeling tools will likely continue to seduce management agencies and decision-makers into requesting prognostic model products that transition from the realm of scientific experiment into part of a policy-making matrix of probabilities and consequences.

Accordingly, assessment of a model's prognostic skill should specify whether the model is being evaluated against calibration or verification data. Calibration data refers to the data set used to estimate or optimize model parameter values, while verification data (variously referred to as validation, confirmation, or corroboration data) are independent of model calibration. Calibration-based metrics are likely to indicate the best possible model performance, particularly if the metrics were used as criteria for parameter estimation (Friedrichs et al. 2006). However, using calibration-independent data for skill assessment provides a much more rigorous test of prognostic model capabilities; ideally the verification data should represent conditions different from those represented in the calibration data set.

It is inevitable that complex models of marine systems will increasingly be used to forecast future conditions. To the extent that such models contain an ecosystem component (or alternately referred to as a biogeochemical component) they presume to quantitatively describe the totality of the interactions between organisms and their environment and, moreover, how these interactions ultimately manifest in time and space as emergent properties that we may observe. Given the overwhelming complexity of this task, it is reasonable to assume that all of the models are severely handicapped by deficiencies in our knowledge of how ecosystems function.

Thus quantitative metrics that assess model performance are required on both scientific and policy-making fronts. First, model improvement, and ultimately, knowledge of how emergent properties arise in complex systems, is aided by incorporating quantitative metrics into a hypothesis testing cycle that involves both model results and observations. Second, a record of model data misfits should not be used to boast of predictive power, but should instead be used to remind the scientist, decision-maker, and the public that the equations within the model represent hypotheses about how the system works, and by omission, hypotheses about which processes are likely to be unimportant–and as is true for any hypothesis, they may be wrong. This harkens us back to the prophetic words of Hedgpeth (1977) who presciently warned against overconfidence in our computations.

Acknowledgments

JIA was funded by theme 9 of the NERC core strategic Oceans2025 program. This manuscript is GLERL contribution number 1464.

References

Allen JI, Smyth TJ, Siddorn JR, Holt M. How well can we forecast high biomass algal bloom events in a eutrophic coastal sea? Harmful Algae. 2007 In press.

Allen JI, Holt JT, Blackford J, Proctor R. Error quantification of a high-resolution coupled hydrodynamic-ecosystem coastal-ocean model: Part 2. Chlorophyll-a, nutrients and SPM. Journal of Marine Systems. 2007b; doi: 10.1016/j.jmarsys.2007.01.005

Allen JI, Somerfield PJ, Gilbert FJ. Quantifying uncertainty in high-resolution coupled hydrodynamic-ecosystem models. Journal of Marine Systems. 2007a; 64:3–14.

Allen JI, Somerfield PJ, Siddorn J. Primary and bacterial production in the Mediterranean Sea: a modelling study. Journal of Marine Systems. 2002; 33–34:473–495.

Allen JI, Clarke KR. Effects of demersal trawling on ecosystem functioning in the North Sea: a modelling study. Marine Ecology Progress Series. 2007; 336:63–75.

Arhonditsis GB, Brett MT. Evaluation of the current state of mechanistic aquatic biogeochemical modeling. Marine Ecology Progress Series. 2004; 271:13–26.

Arhonditsis GB, Adams-VanHarn BA, Nielsen L, Stow CA, Reckhow KH. Evaluation of the current state of mechanistic aquatic biogeochemical models: citation analysis and future perspectives. Environmental Science & Technology. 2006; 40:6547–6554. [PubMed: 17144276]

Beck MB. Water-quality modeling - a review of the analysis of uncertainty. Water Resources Research. 1987; 23:1393–1442.

Bevington, P., Robinson, KD. Data Reduction and Error Analysis for the Physical Sciences. McGraw-Hill Science; New York: 2002. p. 336

Blackford JC, Allen JI, Gilbert F. Ecosystem dynamics at six contrasting sites: a generic modelling study. Journal of Marine Systems. 2004; 52:191–215.

Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review. 1950; 78:1–3.

Brown CD, Davis HT. Receiver operating characteristics curves and related decision measures: A tutorial. Chemometrics and Intelligent Laboratory Systems. 2006; 80:24–38.

Chatfield, C., Collins, AJ. Introduction to Multivariate Analysis. Chapman and Hall; London: 1980.

Clarke, KR., Gorley, RN. PRIMER v6: User manual/tutorial. PRIMER-E Ltd; Plymouth: 2006. p. 192

Clarke KR. Nonparameteric multivariate analyses of changes in community structure. Australian Journal of Ecology. 1993; 18:117–143.

Doney SC. A synoptic atmospheric surface forcing data set and physical upper ocean model for the U.S. JGOFS Bermuda Atlantic Time-Series Study (BATS) site. J Geophys Res Oceans. 1996; 101:25,615–25,634.

Doney SC, Yeager S, Danabasoglu G, Large WG, McWilliams JC. Mechanisms governing interannual variability of upper ocean temperature in a global hindcast simulation. J Phys Oceanogr. in press.

Duda, RO., Hart, PE., Stork, DG. Pattern Classification. John Wiley and Sons; New York: 2001. p. 654

Emery, WJ., Thomson, RE. Data Analysis Methods in Physical Oceanography. Pergamon Elsevier; Oxford UK: 1998. p. 634

Evans GT. Defining misfit between biogeochemical models and data sets. Journal of Marine Systems. 2003; 40–41:49–54.

Fewster RM, Buckland ST. Similarity indices for spatial ecological data. Biometrics. 2001; 57:495–501. [PubMed: 11414575]

Foody GM. Status of land cover classification accuracy assessment. Remote Sensing of Environment. 2002; 80:185–201.

Friedrichs MAM, Carr M-E, Scardi M, Barber R, the PPARR team. Assessing the uncertainties of model estimates of primary productivity in the tropical Pacific Ocean. Journal of Marine Systems. 2008 this volume.

Friedrichs MAM, Dusenberry JA, Anderson LA, Armstrong R, Chai F, Christian JR, Doney SC, Dunne J, Fujii M, Hood R, McGillicuddy D, Moore JK, Schartau M, Spitz YH, Wiggert JD. Assessment of skill and portability in regional marine biogeochemical models: the role of multiple planktonic groups. J Geophys Res Oceans. 2007; in press. doi: 10.1029/2006JC003852

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

Gregg W, Friedrichs MAM, Robinson AR, Rose K, Schlitzer R, Thompson KR. Skill assessment in ocean biological data assimilation. Journal of Marine Systems. 2008 this volume.

Gruber N, Frenzel H, Doney SC, Marchesiello P, McWilliams JC, Moisan JR, Oram JJ, Plattner G-K, Stolzenbach KD. Eddy-resolving simulation of plankton ecosystem dynamics in the California Current System. Deep-Sea Research I. 2006:1483–1516.

Gustafson EJ. Quantifying landscape spatial pattern: what is the state of the art? Ecosystems. 1998; 1:143–156.

Hagen-Zanker A, Straatman B, Uljee I. Further development of a fuzzy set map comparison approach. International Journal of Geographical Information Science. 2005; 19:769–785.

Hashioka T, Yamanaka Y. Seasonal and regional variations of phytoplankton groups by top-down and bottom-up controls obtained by a 3D ecosystem model. Ecological Modelling. 2007; 202:68–80.

Hedgpeth JW. Models and muddles. Helgoland Marine Research. 1977; 30:92–104.

Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. Statistical Science. 1999; 14:382–401.

Allen JI, Somerfield PJ. A multivariate approach to model skill assessment. Journal of Marine Systems. 2008 this volume.

Jolliff JK, Kindle JC, Shulman, Penta, Friedrichs MAM, Helber R, Arnone RA. Summary diagrams and skill assessment for coupled hydrodynamic-ecosystem model performance: Modifications and alternatives to the Taylor diagram. Journal of Marine Systems. 2008 this volume.

Kalnay, E. Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press; Cambridge UK: 2003. p. 341

Kasibhatla, P.Heimann, M.Rayner, P.Mahowald, N.Prinn, RG., Hartley, DE., editors. Inverse Methods in Global Biogeochemical Cycles. Vol. 114. American Geophysical Union; Washington DC: 2000. p. 324Geophysical Monograph Series

Kass RE, Raftery AE. Bayes Factors. Journal of the American Statistical Association. 1995; 90:773–795.

Katz RW, Ehrendorfer M. Bayesian approach to decision making using ensemble weather forecasts. Weather and Forecasting. 2006; 21:220–231.

Kaufman, S. At Home in the Universe. Oxford University Press; NY: 1995.

Krzanowski, WJ. Principles of Multivariate Analysis: A User's Perspective. Clarendon Press; Oxford: 1990. p. 563

Leggett RW, Williams LR. A reliability index for models. Ecological Modelling. 1981; 13:303–312.

Lillesand, TM., Kiefer, RW., Chipman, JW. Remote Sensing and Image Interpretation. John Wiley and Sons; New York: 2004.

Loague K, Green RE. Statistical and graphical methods for evaluating solute transport models: Overview and application. Journal of Contaminant Hydrology. 1991; 7:51–73.

Nash JE, Sutcliffe JV. River flow forecasting through conceptual models, Part 1 – A discussion of principles. Journal of Hydrology. 1970; 10:282–290.

Phillips JD. Spatial structures and scale in categorical maps. Geographical and Environmental Modelling. 2002; 6:41–57.

Pilkey, OH., Pilkey-Jarvis, L. Useless Arithmetic: Why Environmental Scientists Can't Predict the Future. Columbia University Press; NY: 2007.

Powell TM, Lewis CVW, Curchitser EN, Haidvogel DB, Hermann AJ, Dobbins EL. Results from a three-dimensional, nested biological-physical model of the California Current System and comparisons with statistics from satellite imagery. Journal of Geophysical Research. 2006; 111doi: 10.1029/2004JC002506

Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical Recipes. 1986:818.

Raick C, Alvera-Azcarate A, Barth A, Brankart JM, Soetaert K, Gregoire M. 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. 2007; 65:561–583.

Reckhow KH. Importance of scientific uncertainty in decision-making. Environmental Management. 1994; 18:161–166.

Reckhow KH, Clements JT, Dodd RC. Statistical evaluation of mechanistic water-quality models. Journal of Environmental Engineering-ASCE. 1990; 116:250–268.

Rose KA, et al. Skill assessment approaches for comparing spatial maps from coupled biophysical models. Journal of Marine Systems. 2008 this volume.

Schrum C, St John M, Alekseeva I. ECOSMO, a coupled ecosystem model of the North Sea and Baltic Sea: Part II. Spatial-seasonal characteristics in the North Sea as revealed by EOF analysis. J Mar Sys. 2006; 61:100–113.

Smith EP, Rose KA. Model goodness-of-fit analysis using regression and related techniques. Ecological Modelling. 1995; 77:49–64.

Stow CA, Roessler C, Borsuk ME, Bowen JD, Reckhow KH. A comparison of estuarine water quality models for TMDL development in the Neuse River Estuary. Journal of Water Resources Planning and Management. 2003; 129:307–314.

Taylor KE. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research. 2001; 106:7183–7192.

Villars, M., de Vries, I., Bokhorst, M., Ferreira, J., Gellers-Barkman, S., Kelly-Gerreyn, B., Lancelot, C., Menesguen, A., Moll, A., Patsch, J., Radach, G., Skogen, M., Soiland, H., Svendsen, E., Vested, HJ. Report of the ASMO modelling workshop on eutrophication Issues, 5–8 November 1996. The Hague; The Netherlands: 1998. p. 90

Wiggert JD, Murtugudde RG, Christian JR. Annual ecosystem variability in the tropical Indian Ocean: Results of a coupled bio-physical ocean general circulation model. Deep-Sea Research II. 2006; 53:644–676.

Wunsch, C. The Ocean Circulation Inverse Problem. Cambridge University Press; Cambridge UK: 1996. p. 442

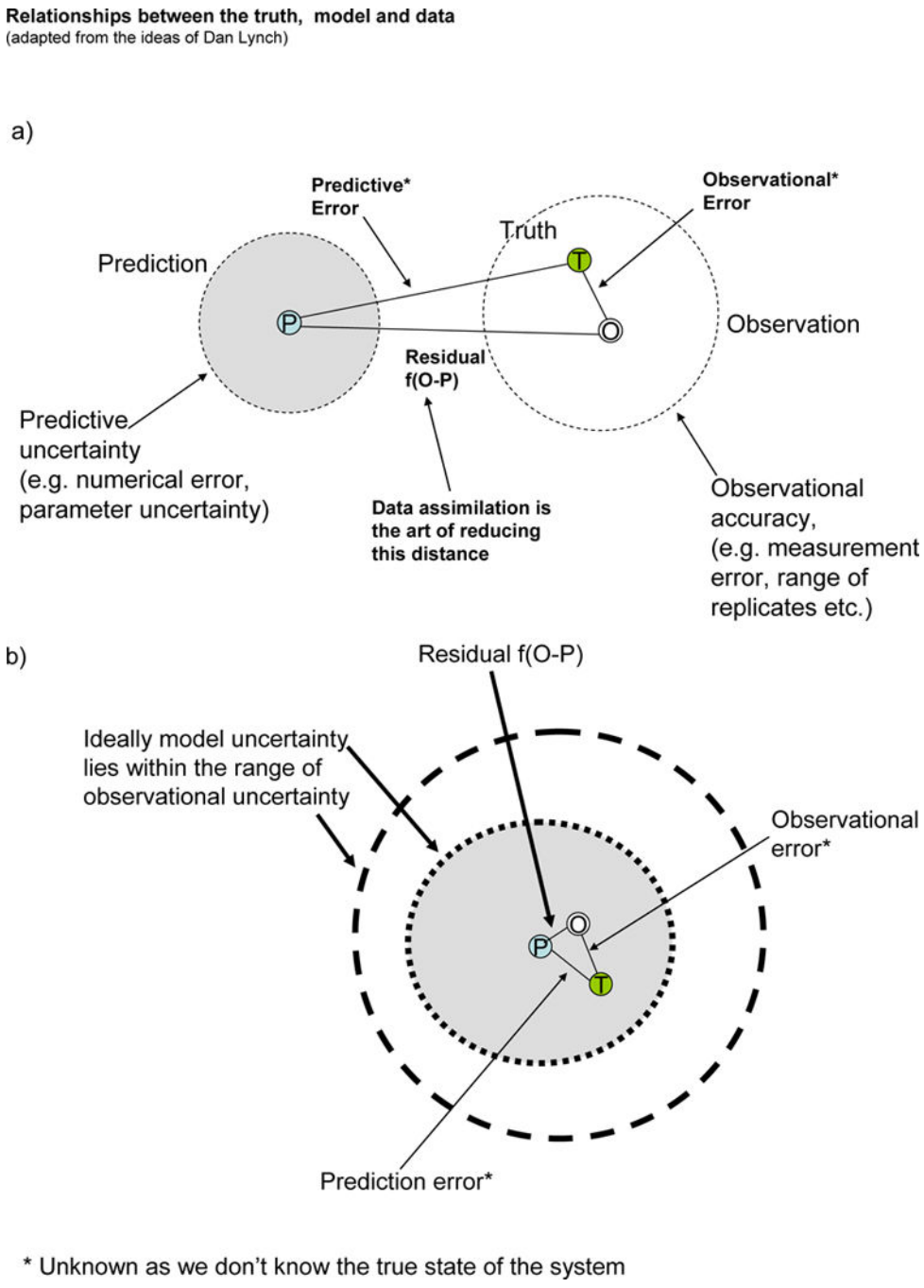

Figure 1.

Schematic diagram of the relationships between model prediction (P), observations (O) and the true state of the system (T). Both P and O are assumed to have a halo of uncertainty. Fig 1a) shows the case for a model with no skill and b) shows the case for the ideal model, with inner circle representing model uncertainty and outer circle representing observational uncertainty.

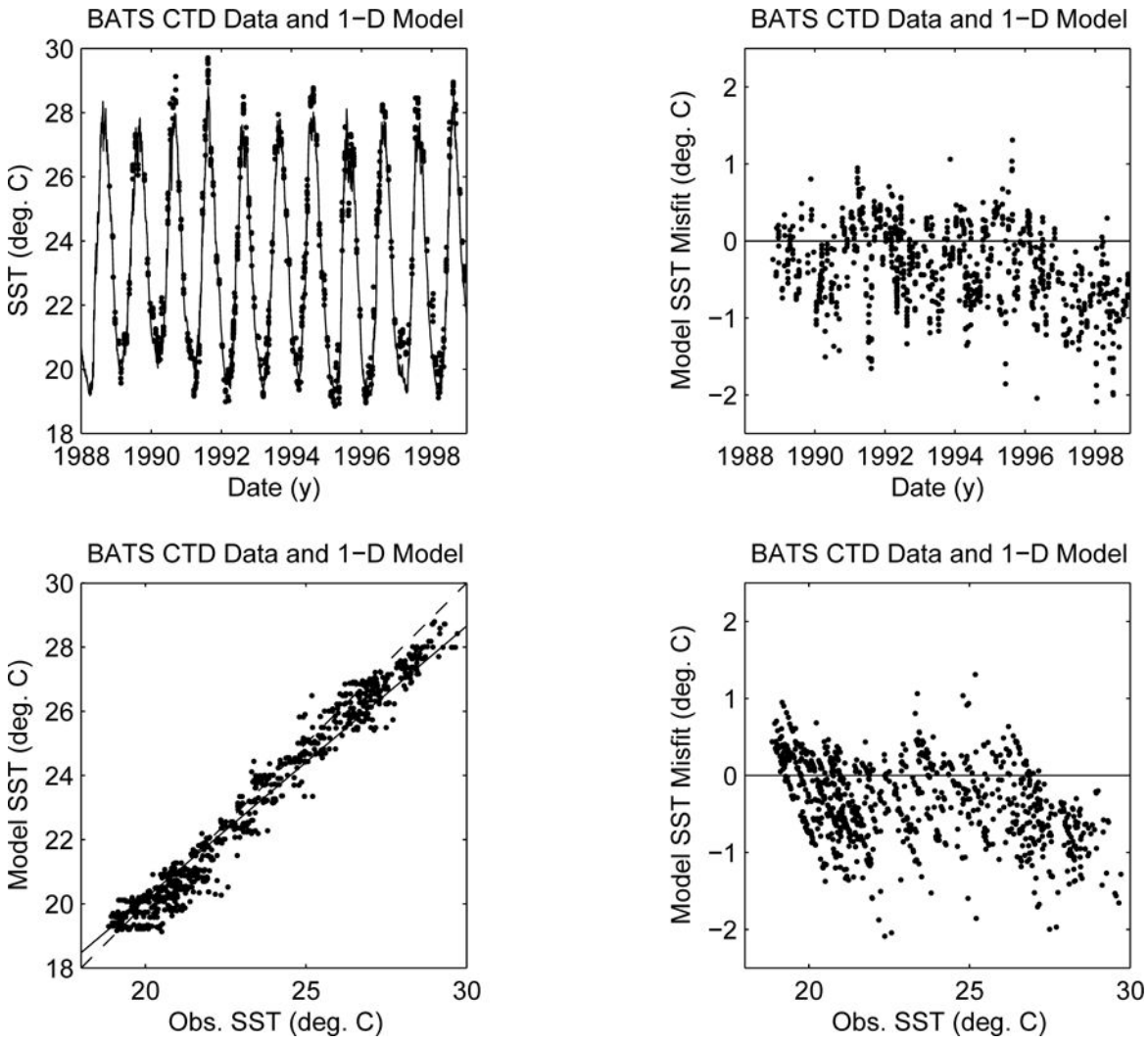

Figure 2.

Observed (points) and predicted (line) variables versus time (top left), the model misfit versus time (top right), the predicted versus observed values (bottom left), and the model misfit versus observed values (bottom right). The linear regression for the model versus predicted value is plotted as a solid line in the bottom left panel relative to the 1:1 line (dashed line).

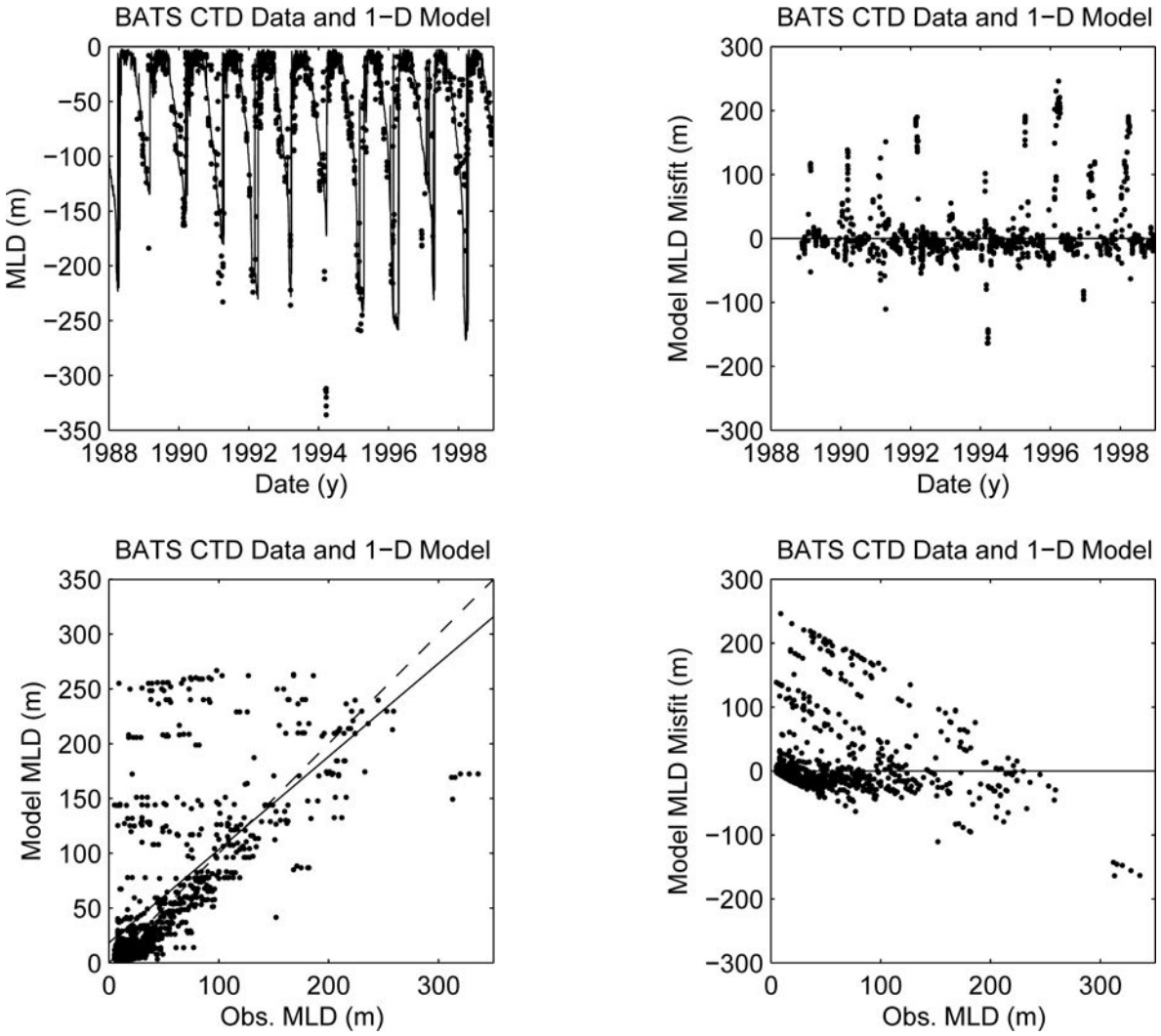

Figure 3.

Observed (points) and predicted (line) variables versus time (top left), the model misfit versus time (top right), the predicted versus observed values (bottom left), and the model misfit versus observed values (bottom right). The linear regression for the model versus predicted value is plotted as a solid line in the bottom left panel relative to the 1:1 line (dashed line).

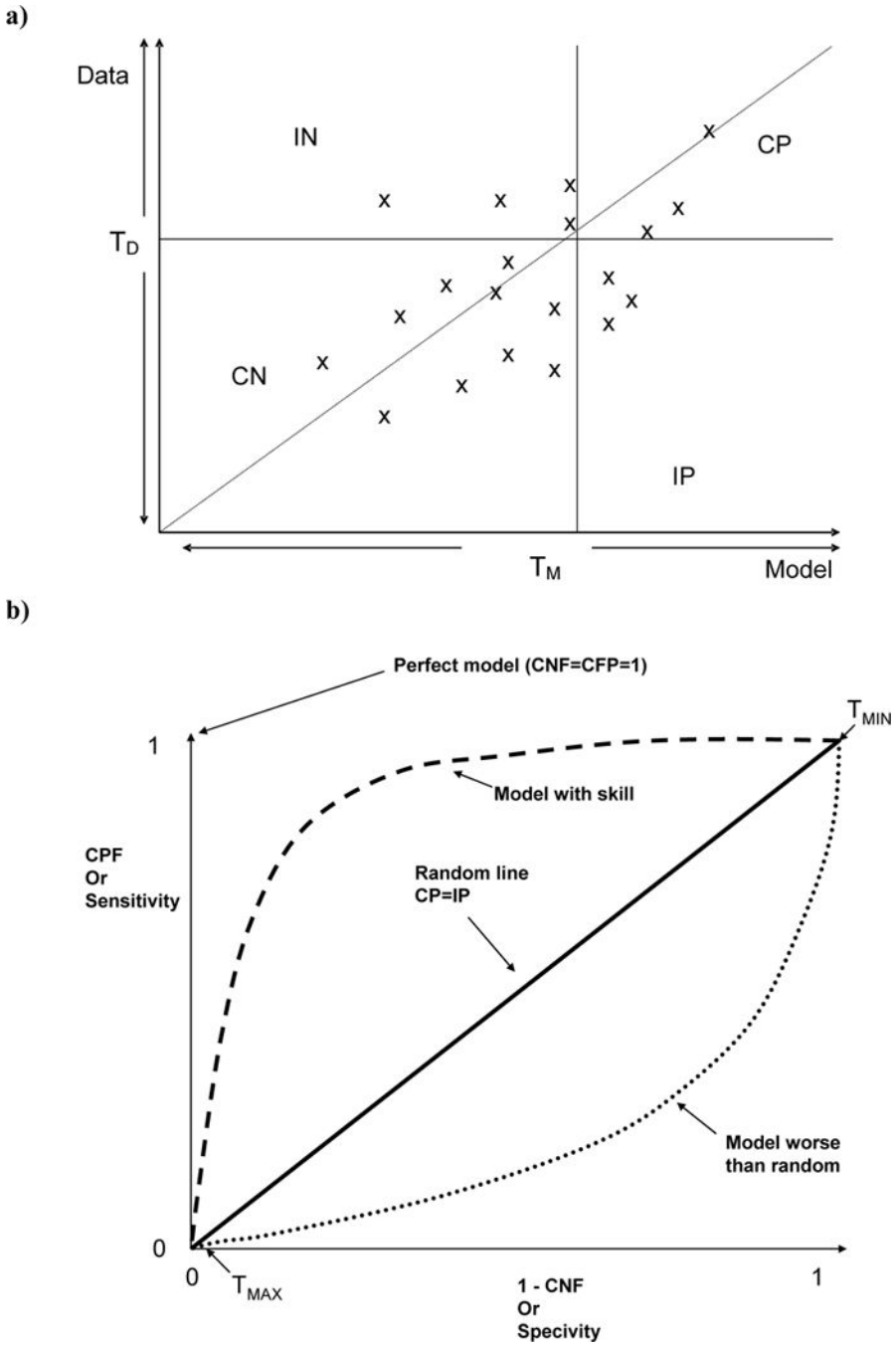

Figure 4.

Schematic diagrams of a) the discrimination analysis and b) the binary discrimination skill assessment curves.

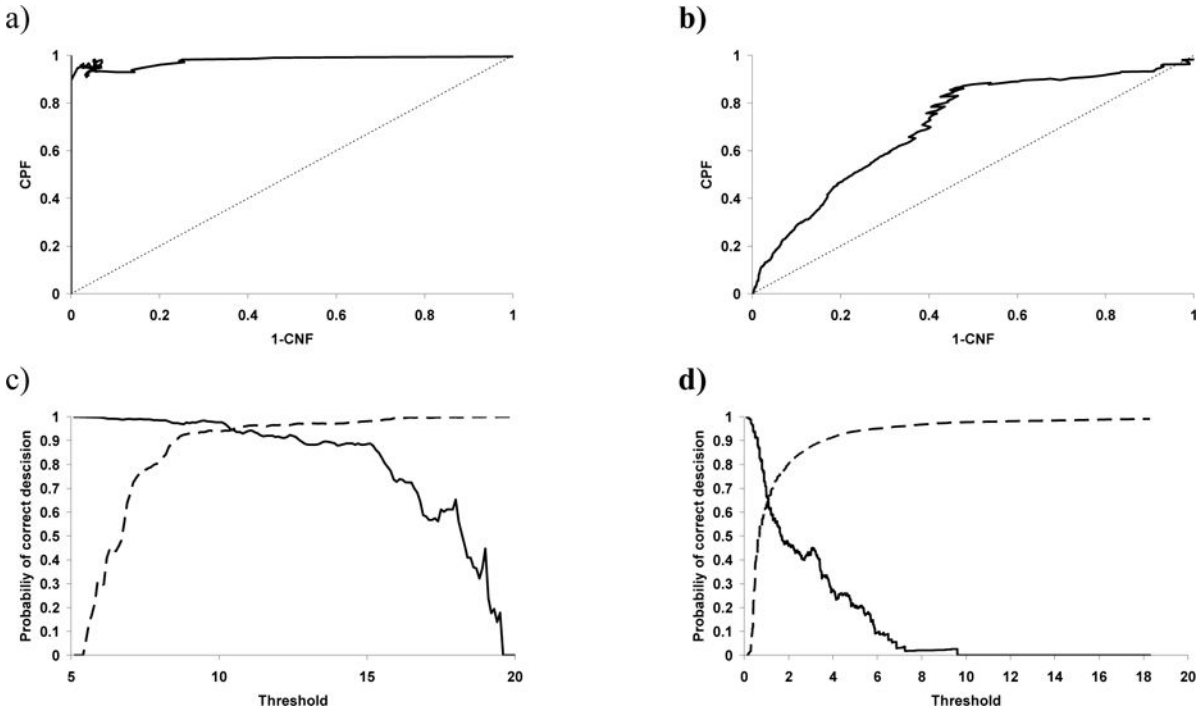

Figure 5.

Binary discrimination analysis plots of model performance, a) temperature, b) chlorophyll. Dots indicate threshold point calculated lowest threshold is top right, highest bottom left. The probability that a positive or negative decision is correct as the discrimination threshold is varied, c) temperature, d) chlorophyll. Positive predictive value = solid line, Negative predictive value = dashed line. Figures are taken from Allen et al 2007b.

Table 1

	Sea Surface Temperature (C)	Mixed Layer Depth (m)
n	951	940
r	0.94	0.67
RMSE	1.10	55.4
RI	1.03	1.93
AE	−0.30	9.86
AAE	0.55	30.12
MEF	0.88	0.062
Intercept / S.E.	3.2 / 0.23	18.26 / 2.46
Slope / S.E.	0.85 / 0.0097	0.85 / 0.031