

Ecosystem Model Evaluation Criteria

Introduction

The Integrated Ecosystem Research Program emphasizes the need for quantitative assessments of the impacts of environmental change, specifically the effects on production, distribution, abundance, availability and variability of components of the ecosystem insofar as these are relevant to fishery management. Any successful IERP proposal must pursue modeling that is integrated across components of the ecosystem and coordinated with field work. The demand for vertical integration does not require that this all be accomplished with a single seamless model: it can be accomplished with multiple hard- or soft-linked models; provided the proposal is convincing that the linkage is practical in light of the spatial and temporal resolutions of the respective models.

Proposed models must meet the highest feasible standards of predictive power, performance, and accuracy to be useful in making long-term improvements in understanding of the marine ecosystem and its processes, as well as improving fisheries management, both in terms of ecologically optimal regulation and providing predictions that are useful for economic planning. Toward that end, the NPRB established an Ecosystem Modeling Committee (EMC) to establish criteria for evaluating proposed models, and to work with the successful applicant team to ensure the highest quality in the modeling funded under the IERP. The EMC has developed this discussion paper on models and the evaluation criteria described below. Invited full proposals must address those criteria.

The Need for Models

There is a recognized need for fisheries science to develop methods of forecasting the production of key species in important fisheries and to forecast the dynamics of the protected species that constrain important fisheries, over time scales that encompass environmental trends that are now evident or expected. It is now understood that the forecasts should link the responses of the species of interest among themselves, to other food web components, to habitat, and to oceanography and climate in order to take full account of ecosystem interactions.

Stock assessments routinely incorporate fishing effects, but models are needed that identify the impacts of changes and trends in climate. The reality of natural long term variation in climate is now scientifically accepted, with examples such as the Pacific Decadal Oscillation, the North Atlantic Oscillation or the Arctic Oscillation. Climate change also is generally regarded within the scientific community as having the potential for major impacts on present and future environmental trends. Such changes in climate, if they occur, will certainly affect the biology of marine systems, and thus will affect fisheries. Models will be needed to assess the likely impacts of these changes in climate on the population dynamics of biologically, commercially and socially important species.

Forecasting models are needed to provide decision support for the fishing industry and subsistence users. With the present state of our knowledge, the one certainty is that change will occur. What is needed are predictions which, in order of specificity, tell us what the change will be, when it will occur, or the probability that it will occur, or at least the range of possibilities that are likely. The example of the major ecosystem changes around the 1976/77 regime shift offers ample evidence of the need to expect change. Forecast models are needed to allow managers to be well-informed stewards of our environment. Effective forecasting is the end point of fisheries science. It is an end point that society expects the science to deliver.

For the greatest practical utility, the species population models should predict species population responses to readily measured leading indicators. Examples of variables which might serve such a function are bottom temperature at the time of spawning, or the timing of the spring bloom, or the direction and strength of winds at the time of larval first feeding. Ecosystem level models should link the leading physical indicators to biological factors which, in turn, affect the key harvested and protected species. Likely climate change might be taken into account by developing scenarios for the leading indicators based on a suite of plausible climate scenarios.

The Difficulty with Models

From the practical standpoint of a manager considering whether to use model predictions in making a decision, the crucial difficulty is in knowing the extent to which the predictions should be trusted. Good predictive power is not a foregone conclusion with ecosystem models: the systems are extremely complex, the models themselves are often dauntingly complex, our scientific theoretical knowledge about ecosystems is rudimentary, and the data available for fitting, tuning and testing the models are very sparse relative to the spatial, temporal and taxonomic detail of the models and of the system.

Because of system complexity, and the incomplete state of the underlying theoretical science, it is not credible to base the claim for a model's correctness on the simple assertion that the model incorporates the "right" mechanisms. Even if all the functional forms in the internal representation of mechanisms were correct, the parameter values and initial state description would also have to be correct to ensure correct predictions. The assessment of sensitivity of predictions to errors in parameter values and errors in the initial state description is a technical undertaking in its own right. The estimation of parameter values, and quantification of their uncertainty, is a highly technical statistical matter.

Because of the sparseness of data relative to the effective number of parameters that are being estimated, mere "goodness of fit" is not a reliable guide to the predictive accuracy of such models. For this reason, much more sophisticated statistical testing procedures are required for realistic quantification of the predictive power of such models.

Evaluation of modeling proposals under this initiative will include consideration of the proposed testing of model predictive power. The intention is to encourage projects that do quantify objective measures of model performance, so that these measures can eventually serve as indications of how good the model is, and with what confidence its predictions can be used in decision making. This is not to demand that every component of an integrated ecosystem model be subjected to quantification of predictive performance. The requirement is that the proposed projects each identify at least one component prediction as the prediction from their model which will be of significance to management, and to develop a credible plan for quantifying the anticipated performance of *that* prediction.

The following section identifies salient points that should be considered in describing the plans for testing model performance. Subsequent sections offer more detailed comments about some of the deeper statistical issues. This is not presented in the spirit of a check list that each proposal must address in a one or two sentence response to each question in the outline. Rather the NPRB is emphasizing that the credibility of the treatment of model validation will weigh significantly in the evaluation of modeling proposals and that the EMC will be charged with reviewing that aspect of each modeling proposal.

It is expected that the demand for a model validation plan will increase the effort of preparing a proposal, and will appreciably increase the effort of the project itself. It is believed that this will contribute to the quality of the products.

The material offered below should be useful to the modeling teams in drafting the model validation portion of their proposals. Any modeling group that has difficulty with this section of their proposal might benefit from expanding their team to include more statistical expertise.

Elements of a Model Testing Plan

The following items need to be addressed thoroughly in the modeling section of full proposals to the IERP. Items A, B, C warrant a few sentences each in the pre-proposal, and item G warrants a paragraph in the pre-proposal. Further explanation of these evaluation criteria is given in subsequent sections.

- A. What is the model intended to predict?
- B. What specific aspect of the prediction is anticipated to be of direct value for fisheries management?
- C. What measure of "accuracy" in the prediction is crucial to determining the usability of that prediction to fisheries management?
- D. What alternative models (other mechanisms, greater degrees of spatial and temporal aggregation, simple statistical predictors) are plausible competitors whose performance will be tested against the model being developed?
- E. How will the achieved predictive power of the model be compared against the performance of plausible alternatives, and how will this guide subsequent choices about model form and parameterization?
- F. What data are available (temporal and spatial resolution, time span covered, data quality) to drive, calibrate, and test the model?
- G. How will the existing data be used to quantify model fit and predictive power?
- H. What pertinent future data are anticipated to become available within the time frame of the project?
- I. How will the future data be used to quantify model fit and predictive power?
- J. How has it been determined that the proposed quantity and quality of data can be expected to be sufficient for the intended use in tuning and testing the model?
- K. How will the probabilistic nature of model forecasts be represented in model output, and how will this be communicated to eventual users of the model predictions?
- L. What is the schedule for providing NPRB with specified data files of observations and model output fields, and how does this set of observations and outputs ensure transparency and verifiability?

Statistical Elements of Model Validation

The above list of requested items may not be wholly familiar to the entire community of modeling project proposers. Furthermore, the technical vocabulary for statistical model validation varies considerably among the various application disciplines. This complicates the task of explaining clearly in the RFP what is really wanted in connection with model testing. It may also pose a challenge when technical reviewers are asked to evaluate this aspect of the proposals received from groups from diverse disciplines. As a possible aid to this 2-way communication problem, we offer below an attempt at a unified discussion of some of the technical statistical issues, along with a consistent set of definitions drawing where possible from the technical vocabulary of modeling in the atmospheric sciences and in applied statistical decision theory. The choice of a reference lexicon from the atmospheric sciences is intended to capitalize on the relatively high level of maturity of that discipline's experience with these matters. The framework of statistical decision theory is already common currency in fisheries management, at least in some single species contexts.

Predictive Power in Context

The central concept is the quantification of what has been loosely called predictive power, performance, or accuracy in the discussion above. To make this concrete, there first must be a specification (really a choice) of what is being predicted (see item A in the above list of model evaluation criteria). The presumption is that the knowledge of the quantity being predicted would directly affect decisions by agency managers, industry planners or subsistence users (item B). Understanding how the prediction will be used for decisions should reveal how errors or inaccuracies in the prediction will lead to incorrect decisions if the prediction is acted upon, which in turn should suggest the most telling measure of accuracy appropriate for evaluating a model that is used for making this kind of prediction in this kind of decision context (item C).

Model Skill

In atmospheric modeling, the statistically assessed measure of a model's predictive accuracy, in the context of the planned use of the prediction, is called the model "skill." Several formulations are available, as appropriate for the kind of quantity being predicted and its use in a decision process. More particularly the skill can refer to different aspects of a model's performance--notably hindcast skill, forecast skill, artificial skill, ensemble skill, conditional skill (not all of which are mutually exclusive)--which will be defined in subsequent sections.

Skill Score

Regardless of whether the skill is hindcast, forecast, artificial, ensemble, or conditional, there must be a defined metric for quantification. The "skill score" (SS) is useful for evaluating predictions of numerical variables that vary continuously or approximately continuously, such as recruitment to a specified fish population. It compares a forecaster's root-mean-squared or mean-absolute prediction errors, $E_{\{f\}}$, over a reference scenario, with those of a reference alternative prediction, $E_{\{refr\}}$, such as forecasts based entirely on the past average, a simple regression relationship, or persistence, which involve fewer variables, parameters, or mechanisms than the model being evaluated:

$$SS=1-(E_{\{f\}}/E_{\{refr\}})$$

If $SS>0$, the forecaster or technique is deemed to possess some skill compared to the reference technique, and it may be concluded that the addition of variables, parameters, or mechanisms beyond those of the reference technique was worthwhile.

Skill Quantified by Classification Error Rates

For binary, yes/no kinds of forecasts or detection techniques, such as the determination of a regime shift, the probability of detection (POD), false alarm rate (FAR), and critical success index (CSI) may be useful evaluators. For example, if A is the number of forecasts that rain would occur when it subsequently did occur (forecast=yes, observation=yes), B is the number of forecasts of no rain when rain occurred (no, yes), and C is the number of forecasts of rain when rain did not occur (yes, no), then

$$\text{POD} = A / (A + B)$$

$$\text{FAR} = C / (A + C)$$

$$\text{CSI} = A / (A + B + C)$$

For perfect forecasting or detection, $\text{POD} = \text{CSI} = 1.0$ and $\text{FAR} = 0.0$.

POD and FAR scores should be presented as a pair. A decision theoretic assessment will present the full contingency table of probabilities of true positive, true negative, false positive, false negative, conditioned on a prior and data-driven inference of the underlying probability that the system state actually is "positive" or "negative."

Hindcast Skill

A model is usually tuned to some observed data set. That is, the model parameters are adjusted to try to best replicate the observations, or are estimated from the observations by a statistical procedure. The model's ability to replicate these data is generically called the "goodness of fit" which may be measured as "hindcast skill," for instance the percent of variance of the observations that the model can capture. The observations used in this process are termed "dependent" or the "training data."

Data Assimilation

A special case of hindcast skill assessment is employed in diagnosis of the internal consistency of the model's representation of the system. A deliberately diverse and comprehensive set of data, possibly sampled at different times and intervals and different locations, may be combined by use of the model into a unified and consistent and complete state description of the system. This is called "data assimilation." Poor fit in data assimilation may reveal shortcomings of the model structure, and warrants close examination from the perspective of temporal, spatial and taxonomic aggregation in the model, known omission of mechanisms, and measurement error in the data themselves. At a minimum, it is important to be able to account for the poorness of fit, if such is seen. Good fit, in a data assimilation, may be less diagnostic.

Forecast Skill

By contrast, "forecast skill" is a measure of the model's ability to reproduce observations that have in no way been involved in the model tuning process. These are "independent" data. The forecast skill can be estimated from entirely "new" data obtained after the model was tuned, i.e., real forecasts of the future. Alternatively, the forecast skill might be estimated from "holdout data" by application to a subset of data that were initially set aside, or via jackknife cross-validation.

Degrees of Freedom

In an unconstrained dynamic system, the number of independent variables required to specify completely the state of the system at a given moment is called the "degrees of freedom" of the system. If the system has constraints--that is, definitional, kinematic or geometric relations

between the variables--each such relation reduces by one the number of degrees of freedom of the system.

In fitting a model to data, as in estimating the model parameters from the data via a likelihood procedure, the number of degrees of freedom available for the statistical operation is the number of independent observations (observations that cannot be calculated exactly from the other observations), whose joint probability is affected by all of the parameters, minus the number of independent parameters in the model (parameters whose value cannot be calculated exactly from the values of the other parameters). Generally, the precision of the parameter estimates resulting from such a statistical operation will increase with the available degrees of freedom; and if the available degrees of freedom is zero or negative, the parameters cannot be resolved uniquely at all (though it may be possible to estimate some combinations of parameters, such as ratios or sums). When the joint distribution of the observations can be accounted for by a reduced set of the parameters of interest (such as a subset of the parameters, or combinations such as sums or ratios), the original model is said to be "non-identifiable," and estimation of the full set of original parameters is impossible without the provision of supplemental information (such as theoretical constraints, ancillary statistical analyses on other data, or prior distributions) not contained in the data.

Artificial Skill

Fitting a model to data can always be done, even though the model may be completely "wrong" and have little or no forecast skill. In the limit that the effective number of tunable model parameters equals or exceeds the number of degrees of freedom in the system (or when the number of available degrees of freedom for the statistical fitting goes to zero or less), the model may have nearly perfect hindcast skill, even if the real or forecast skill is nil. The "artificial skill" is the difference of hindcast skill minus forecast skill, and is a measure of the degree to which the hindcast skill would prove illusory as an indicator of forecast skill.

Parsimony and Alternative Models

Ecosystem models, like most large scale environmental models, as a class, are under pressure of degrees of freedom problems eroding their forecast skill. In practice, as the number of tuned model parameters increases, the models start to fit the noise rather than fitting the signal, and forecast skill degrades accordingly. This places a premium on effective strategies for keeping the number of model parameters to a minimum. In statistical modeling (e.g., multiple regression, GAM), there are standard procedures (e.g., stepwise significance testing, AIC) for making decisions about including model terms or adding predictor variables. Parsimony in large ecosystem models needs to be approached in a similar spirit. In part this may be addressed by a thorough exploration of the performance of simpler alternative models as the reference models for scoring forecast skill (items D and E).

Persistence models deserve special attention as candidates for simple alternatives. In meteorology and oceanography, red noise processes, such as the simple first order autoregressive, often fit the existing data about as well as much more elaborate models. Red noise models are important alternatives, because they make very few assumptions about underlying mechanisms or the state space, and they imply very low predictive power for the distant future even though the short term (and possibly medium term) predictive power is often good. Such short term predictive power can be useful in a decision process that only needs to look a short time span ahead of the ongoing monitoring, with the knowledge that adaptive corrections are feasible with each revision of the forecast with each increment of ongoing monitoring information.

Conditional Skill

Model skill (hindcast or forecast) may be measured on some conditional basis. This might be conditional on one or more of the state variables being in a specified range, or it might be conditional on more elaborate scenarios involving the system state or trajectory. For a simple example, model performance might be scored only when an ENSO index is above a certain threshold. The skill in the defined limited set of situations is termed "conditional skill."

Ensemble Forecasts

A model may be run numerous times with the same trajectory of deterministic forcing, but varying initial conditions, or independently sampling presumed real stochastic processes in the forcing, or sampling presumed real stochastic processes that are part of the model, or sampling parameter uncertainty. Each run is termed a realization and the group of all runs an ensemble. The statistics of the ensemble (mean, variance, etc) are obtained from the set of realizations. A variety of measures of the ensemble distribution can be used to estimate model sensitivity (to numerical precision, to uncertainty about initial conditions, to uncertainty about parameters, or to uncertainty about the future unfolding of real stochastic forcing) and precision. The relationship of the ensemble distribution to a known independent value can be used for a robustly conditioned estimate of forecast skill.

Skill and Forecast Confidence in Perspective

With this broad spectrum of available skill metrics for quantifying model predictive power it becomes clear why the choice of metric (item C) needs to be tailored to the proposed decision-making use (item B) of the promised useful prediction (item A). We see that high hindcast skill is not very telling, though low hindcast skill might serve some diagnostic function.

One might select a very stringent forecast skill metric like percent variance accounted for, but a less demanding forecast metric like a simple n-tile contingency approach would be appropriate when the anticipated use of the prediction is for detection or prediction of a relatively discrete state, such as a defined ecosystem configuration. Some effort should be made to explain how the skill metric is a good match to the prediction use. A conditional skill metric might be entirely appropriate if it is conditioned on a scenario that is of genuine concern, and if there is a plausible decision process that will use the prediction, in the way that it is packaged by the model, if the defining scenario arises.

Model predictions are inherently probabilistic, owing to process variation represented explicitly in the forcing and in the modeled mechanisms, and owing to parameter uncertainty. Evaluation of the probabilistic nature of the predictions can generally be accomplished through ensemble forecasts exploring the correct universe of process variation and parameter uncertainty. This should be taken into account in the estimation of skill. Ultimately, the confidence in the predictions that are delivered to the users of the prediction, should be communicated in terms of quantification both of the probabilistic nature of the prediction (item K) and the measured forecast skill appropriate to the use (possibly a conditional forecast skill).

Data Sources, Use, and Sufficiency

Because quantification of forecast skill is an empirical statistical enterprise, it is crucially dependent on data, which coincidentally, in the case of biological data for ecosystem modeling, are generally in very short supply and subject to many possible questions of sampling error and measurement error. For this reason it is important to be very explicit and detailed about what data really are, or will be, available (items F and H), and exactly how they will be used for the various phases of the planned model testing (items G and I), noting that different subsets of the data might get used for different parts of the process, such as tuning, data assimilation, or cross-validation.

The sparseness of ecological and fisheries data, relative to the scale and complexity of the system represented in an ecosystem model, raises the question whether the amount of data is sufficient to draw the conclusions that are promised (item I). Parameter estimation and forecast skill quantification both are formal statistical operations whose resolution (precision) depends on, among other things, the "sample size" of the data employed.

In planning simpler statistical exercises there are standard techniques, generally called "power analysis," for quantifying in advance the expected resolution of the estimates from a specified statistical operation on a specified quantity of data of defined quality. This allows an evaluation of whether the planned study is likely to deliver results that are sufficiently conclusive to answer the motivating question or to be useful in the decision process which will rely on them. The same logic may be pursued for much more elaborate statistical operations, such as ecosystem modeling, by simulating data sets corresponding to the size and quality that is planned, using hypothetical values of the key parameters to drive the simulation, submitting these simulated data sets to the planned analysis, and reporting the statistics of the power of the planned analysis, in context of its intended use, based on the ability of the analyses to recover the known hypothetical parameter values.

Indicators

There are striking differences in the scales and intensity of sampling, methods of measurement, and levels of resolution, as they relate to the different kinds of physical and biological data that will be combined in an ecosystem model. Often enough the biological quantities of greatest interest for decision purposes are among those with the weakest or sparsest measurements. Sometimes the quantity which is the focus of policy interest is not even well defined, such as "ecosystem health."

The use of more readily measured (and defined) surrogate variables is attractive, but raises its own set of questions that need to be addressed in the course of model validation when such "indicators" are used in the modeling. Basically, there needs to be a clear definition of what the indicator is supposed to indicate, there needs to be a set of independent measurements of that ground truth, and there needs to be a statistical analysis of how well the indicator predicts that sample of ground truth. This is essentially an assessment of "simulcast skill," in that the surrogate is being used to predict the contemporaneous ground truth. If the indicator is a leading indicator, then the claim is a forecast, and its performance should be evaluated in terms of forecast skill.

Transparency and Data Access

Science is a social activity, where quality control and consensus are the products of independent verification of claims made about a shared empirical reality. In a science such as ecosystem modeling, where the data are so sparse and hard won that replication of the data themselves would be prohibitively expensive at best, and where the prominence of historical data trajectories sometimes makes renewed measurement of the key data an impossibility, independent verification of assertions about model performance requires common access to common sets of data about which there is substantive scientific agreement concerning what was measured, where it was measured, when it was measured, and how it was measured.

NPRB will require that data sets used in the modeling projects it funds as part of the IERP, and numeric output fields from data assimilation and critical model runs, be made available and thoroughly documented on the NPRB database. The proposals should commit to a schedule of delivery of specified data files and specified files of model output, and the proposals should

explain how these files would constitute an adequate set for independent attempts at verification of skill and key relationships revealed by the modeling (item L). Progress in meeting this schedule will be one of the elements of annual reviews by the EMC of funded projects. The NPRB will encourage the funded modeling projects to develop and document consensus among themselves about the quality of pertinent data.

Kinds of Models

Models developed for purposes of “explanation” may not appear to fit the paradigm of models developed for purposes of prediction, as described above. But from the standpoint of the NPRB demand for high standards of documented useful performance in the models funded under the IERP, an explanatory model needs to make some verifiable significant prediction. Testing whether some hypothesized mechanisms represented in a model account for an observed pattern in data via examining the fit to those data, is essentially an evaluation of hindcast skill, which is a weak test for complex models with sparse data, as discussed above. The credibility of the hypothesized mechanism should be evaluated by identifying some corollaries, specific to the hypothesized mechanisms, and not automatically entailed in the data used for tuning the model, and then treating these as forecasts to be evaluated by the statistical machinery discussed above.

The assumptions and predictions of explanatory models require testing and assessment of predictive power that is as rigorous as that required of models intended for a direct management prediction. Thus, field research should support the testing of assumptions and predictions of an explanatory model. Such testing is necessary before the explanatory model can contribute toward the development and testing of the quantitative predictions that are the ultimate goal of the IERP. One should note that some extant conceptual models, such as the Oscillating Control Hypothesis, in many cases have yet to have their assumptions or predictions rigorously tested. Such tests are important, and could be an appropriate activity within the IERP.

EMC Role in Reviewing Modeling Proposals and Projects

Review of Full Proposals

The EMC will be involved with reviewing the modeling components of full proposals to the IERP. Full proposals will be expected to address all the elements of a model testing plan (A through L), as discussed in this document on ecosystem model evaluation criteria. The usual NPRB standards for conflict of interest will apply for recusal of EMC members from review of particular projects where conflict might arise.

Proposal Revisions

There is a possibility that the model validation portion of an otherwise meritorious proposal may not meet the expectations of the EMC. If that proposal is selected for funding, the successful team will be expected to meet with the EMC and make adjustments in their modeling approach as reasonable and appropriate within the funding and time limitations. If the funded team and EMC cannot agree on suitable revisions, the matter in dispute will be elevated to the Science Panel and then to the full Board for resolution.

Annual Review and Direction

The EMC will meet annually with the successful team to appraise the modeling effort and recommend adjustments as necessary.