Journal of Marine Systems 76 (2009) 1–3

Contents lists available at ScienceDirect

Journal of Marine Systems

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

Preface

Skill assessment for coupled biological/physical models of marine systems

Simulation models coupling physics to biological processes in the ocean are central to many current programs. Ocean physical models have approached a high level of sophistication; the physical relationships are canonical, and modern computational technology for fluid mechanics has advanced steadily for two generations or more. The complexity of biological processes in the ocean presents enormous difficulties beyond physics. There is a recognizable mode of operation wherein ‘complete’ physics is coupled to reduced-complexity biology; and simulations are typically chosen to fit field problems and available data. The upshot of this situation is a great diversity in what is possible in ‘replicating observations’, and even more importantly, in assimilating them into simulations and creating forecast systems.

Many important programs are currently facing the consequences of this. The biological problems being addressed are of immediate human concern, and there is a sense that skillful simulations *can* be constructed. Yet what is meant by “skillful” and “simulation” is typically very different depending on the target problem. Examples include the Joint Global Ocean Flux Study (JGOFS), Ecology and Oceanography of Harmful Algal Blooms (ECOHAB), and Global Ecosystem Dynamics (GLOBEC), all of which have had numerous regional manifestations in terms of target organisms and interactions.

This volume constitutes an effort to develop the theoretical basis for the underlying problem of *skill assessment* in all of its relevant senses—across species and ecosystems, geographical places, and data types. Generic theoretical problems are addressed in specific program contexts; both scholarly and practical aspects are presented across this diverse landscape. It is hoped that presenting this work together will help in finding common ground and the conceptual strength and generality that that leads to.

A scholarly basis of agreement is prerequisite to regulatory progress and sound public advisement. However it is a mistake to focus exclusively on the former, to the neglect of progress in the public sphere where real problems originate and demand attention. Accordingly, coupled to the scholarly advancements herein there must be a parallel effort to embed findings in regulatory practice. Our objective in publishing these papers together is to advance this end.

1. Skill assessment vocabulary

The general problem of vocabulary is greatly compounded by its very importance. The several fields represented in this volume have made significant progress, evolving their own dialects in the process. It is not our intention to change that. However, there are a finite number of critical ideas in these diverse fields.

For the purposes of this volume, *Skill* is fidelity of model behavior to *Truth*. The sense of fidelity is implied in the purpose of the modeling activity. There are several, dealing with state variables, features, or dynamical processes/interactions.

Assessment is a human judgment about skill. Ideally, we have skill and make no mistake about it. There are two types of failure—failure to achieve skill, and failure to recognize it. Each can have distinct consequences.

1.1. Truth, error, misfit

In science we observe, measure, predict, postulate, all on the premise that natural truth is observable and understandable. The understanding we seek may be deterministic, or stochastic, or a blend; yet the presence of natural truth in the object of our study would appear to be noncontroversial.

Both observation and theory—data and model—are approximations to truth. Neither is perfect. Both are separated from truth by errors of fundamentally different origin. Fig. 1 indicates data error ε_d and model error ε_m as distinct entities. Truth is hidden from us by both of these errors. Neither can be known exactly.

It is a classic misconception that data *are* truth. First, data are an incomplete sampling—much of truth goes unobserved. And second, the method of observation is necessarily imperfect and inserts a wedge between data and truth. Models proceed from imperfect theory toward truth estimation; data proceeds the opposite way, from truth through the imperfect filter of the sampling method. Proper interpretation of both requires an understanding of the underlying model and sampling method, and their errors.

0924-7963/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jmarsys.2008.05.002

Since Truth cannot be known, neither can Error. What can be known, with certainty, is the misfit δ, the difference between data and model. A simple manipulation yields

$$\delta = \varepsilon_d - \varepsilon_m \quad (1)$$

If the two errors are uncorrelated, the misfit variance is the simple sum of the two error variances:

$$\overline{\delta^2} = \overline{\varepsilon_d^2} + \overline{\varepsilon_m^2} \quad (2)$$

where the overbar indicates an average. Misfit is by definition limited by observation. Much of the truth goes unobserved, and there is zero misfit if there are no data! A major objective of skill assessment is to make estimates of truth where we *lack* data and to make judgments about those estimates.

Conceptual models of both errors are needed. The simple objective of driving misfit to zero is insensitive to the precision of the observations. Even if data were perfect, zero misfit does not imply zero error unless model error is zero. Necessarily incomplete data require extrapolation if skill is to be assessed away from data points.

1.2. Prediction

We can go further by recognizing that a prediction typically is a blend of model and data, illustrated in Fig. 2. There is a prediction error ε_p that is ideally smaller than the other errors. The general field of data assimilation can be viewed as forming predictions that join data and model, and ideally getting closer to truth. The prediction procedure is itself a theoretical construction. By extension of Fig. 1, we can define misfits between prediction and data; and between prediction and model.

Error, like truth, is real and unknowable. At best we can expect a statistical description of error. A good prediction is a credible combination of data and model, invoking known or hypothesized statistics of their errors. The resultant prediction error ε_p is therefore a blend of ε_d and ε_m, achieved by selecting model inputs. A good prediction will have valid statistical basis for the selection, and will result in

- small, noisy misfits;
- small, smooth inputs deduced (e.g., from an inversion); and
- credible features of the prediction.

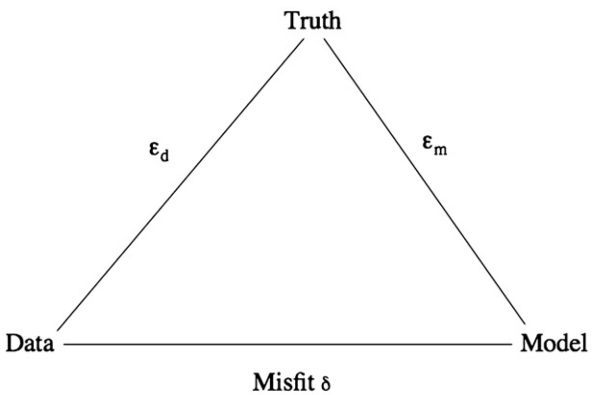

Fig. 1. Conceptual diagram of truth, Observed Data and Models Results, and associated departures from each other. Note the distinction between Error (ε) and Misfit (δ). Refer to text for definitions.

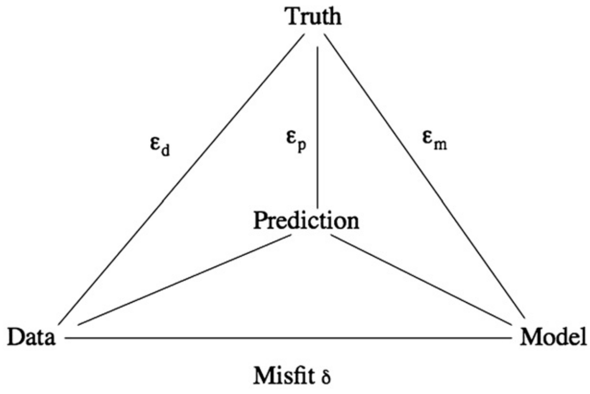

Fig. 2. Prediction and Prediction Error; remaining terms as in Fig. 1.

One presumes a range of recognizable possibilities, giving context to the words small, noisy, smooth and credible. This knowledge of credibility must be available to support a prediction, whether formally or informally. It is often realized in the form of a "Best Prior Estimate", against which adjustments to inputs, control parameters, or other deductions are judged; and the probability (credibility) of departures from it. Related is the concept of overfitting, in which misfits are rendered arbitrarily small by electing unrealistically large or noisy fluctuations in the free parameters. In so doing, overfitting reduces predictive skill.

The contributions to this volume explore aspects of misfits, errors and other metrics one faces when assessing model skill. We look forward to continued development of all of these ideas and hope that the collection of papers presented herein will set the stage for future progress in this important field.

2. Coverage and topics

Authors of the papers in this volume convened for a series of two workshops at University of North Carolina, Chapel Hill.[1] Invitees were drawn from several sources reflective of the state-of-the-art in coupled marine modeling in four different application areas: plankton ecosystems and biogeochemistry, harmful algal blooms, food webs, and water quality. Interaction amongst the groups resulted in papers focused on the cross-cutting themes of skill metrics (Stow et al.) and skill assessment in the context of data assimilation (Gregg et al.). Rose et al. introduce some new tools for quantitative comparison of spatial maps using simulated data and identical twin experiments to evaluate their effectiveness.

2.1. Plankton ecosystems and biogeochemistry

Applications related to the area of ocean biogeochemistry were quite diverse, including an assessment of a radiative transfer model used as forcing for coupled physical–biogeochemical models (Gregg and Casey). Jolliff et al. introduce a new tool for skill assessment, the "target error diagram," using it to test an ocean ecosystem model with satellite-based ocean color data. Allen and Somerfield describe principal component analysis and nonparametric multivariate approaches to assessing

[1] http://www-nml.dartmouth.edu/Publications/internal_reports/NML-06-Skill/.

an ecosystem model of the North Sea. Doney et al. present a generalized framework for assessing the skill of global upper ocean ecosystem–biogeochemical models, utilizing a variety of metrics including model–data residuals, time–space correlation, root mean square error, and Taylor diagrams. Advanced methods for skill assessment were applied in two different regional studies. Friedrichs et al. tested a set of thirty primary production models with an extensive data set in the Equatorial Pacific. Wallhead et al. use data subsetting and Monte Carlo simulation to test statistical and dynamical models of areal-mean chlorophyll on Georges Bank.

2.2. Harmful algal blooms

Stumpf et al. evaluate the skill of an operational harmful algal bloom forecast model for *Karenia brevis* on the west Florida shelf, using quantitative metrics to assess how well the system performs in five aspects: identification, intensification, transport, extent, and impact. A harmful algal bloom issue in the Gulf of Maine is addressed by Smith et al. who use a Monte Carlo ensemble smoother approach to inverting for initial conditions and mortality in a spatially explicit physical–biological model of the toxic dinoflagellate *Alexandrium fundyense*.

2.3. Food webs

Fennel grapples with the issue of coupling plankton ecosystem models with models of fish production, examining the impact of model truncation and parameterization on skill. Steele uses inverse methods to fit a linear food web model to observations, critically evaluating the ecological assumptions underlying these optimization strategies via comparison to application in nonlinear dynamic models.

2.4. Water quality

Fitzpatrick provides a review of metrics used to assess the skill of water quality models, which are becoming increasingly important in setting policy on total maximum daily loads of nutrient discharge in many areas of the coastal ocean. Sheng and Kim use a variety of quantitative metrics to evaluate a water quality model of the Indian River Lagoon. Stow and Scavia utilize a Bayesian framework for parameter estimation that yields both model forecasts and probabilistic estimates of forecast uncertainty, a key input into policy decision-making.

3. Summary outcomes

This volume is intended to represent the state-of-the-art in coupled physical–biological model skill assessment. Included is theory, practice, and data assimilation. From these papers and discussion at the two workshops, a number of recommendations emerged for the future of operational physical–biological models to be used for management purposes:

Conceptual

- Encourage the use of probabilistic model results—mean and variance—and the expression of this in simple ways to a general audience, backed by rigorous analysis.
- Encourage the formalization of the best prior estimate—at the least, the mean and variance of all relevant prior quantities.
- Always examine the posterior: a) the remaining misfit and b) the departure from the prior. There is information in both.
- Ensemble modeling approaches, specifically the use of an ensemble of *different* models, are appealing in the context of operational physical–biological models.

Practical

- It is essential to facilitate access to real-time data streams. This includes networking, servers, and people.
- Encourage communication and interaction between data providers and modeling activities.
- Similarly, encourage partnership between physical modeling and biological modeling.

Organizational

- Recognize the importance of organizational structure. Encourage regional expertise in regional centers; and networking of these relative to technical and scientific generalities.
- Encourage a blend of Government/University/Industrial activity.
- Use the existing centers and cooperative programs to their fullest. There is much opportunity in these for cross-fertilization. Avoid creating new organizations if extant ones can be made to work.
- Recognize the importance of small steps toward a larger goal.
- Focus on system integration—of models, theory, and observation—as an overarching goal.

Acknowledgements

We gratefully acknowledge the support of the NOAA Center for Sponsored Coastal Ocean Research, through the Cooperative Institute for Climate and Ocean Research at WHOI. Thanks to Sue Stasiowski for organizational assistance and the staff of the Friday Center at UNC-Chapel Hill for providing an excellent venue for the workshops.

DRL gratefully acknowledges support from the National Science Foundation (CMG program). DJM gratefully acknowledges support from the National Science Foundation, National Oceanic and Atmospheric Administration, National Aeronautics and Space Administration, and the National Institute of Environmental Health Sciences. FEW gratefully acknowledges support from the Office of Naval Research (SEACOOS) and the National Ocean Partnership Program (HYCOM).

Daniel R. Lynch
Dartmouth College, Hanover, NH 03755, USA
Corresponding author.
E-mail address: Daniel.R.Lynch@dartmouth.edu.

Dennis J. McGillicuddy, Jr.
Woods Hole Oceanographic Institution,
Woods Hole, MA 02543, USA

Francisco E. Werner
Marine Sciences Department, University of North Carolina,
Chapel Hill, NC 27599, USA