Ocean Model Metrics: What we currently use and how we can advance these methodologies
A workshop held at IPRC, Honolulu, Hawaii
25 February 2006

Large-scale ocean models to be used for synoptic prediction must be capable of simulating high frequency (hours to several months) and mesoscale (10 to 1000 km) processes. With increases in computational power, this has become possible in recent years. In addition, centennial global coupled climate simulations are routinely conducted using lower resolution (one degree or coarser) ocean models. The fidelity of the ocean models directly impacts the accuracy of both short-term forecasts and climate projections. Hence, quantitative metrics need to be constructed that can be used to assess the realism of all these simulations. Metrics should be designed to account for the space-time scales resolved by the model as well as the representation of ocean processes by the model physics at the relevant scales. Diagnostics based on data comparisons, statistical measures, and governing dynamical balances in the ocean all represent means to understand the veracity of these ocean models. Daily products from production centers on which decisions are based also require impact metrics that must relate closely to the performance and accuracy in the decision process.

Although it is not axiomatic that the impact of data assimilation on model solutions is always positive, it has provided representations of the ocean that can be used for a variety of purposes. Through the Global Ocean Data Assimilation Experiment (GODAE - http://www.bom.gov.au/bmrc/ocean/GODAE), a series of metrics have been developed to evaluate assimilative models (http://www.clivar.org/organization/gsop/synthesis/synthesis.php). Such calculations involving fine-resolution, large-scale basins are very computationally intensive (sometimes prohibitively so) hence prognostic ocean models, forced with specified atmospheric states, are still used for ocean studies. Since they are unconstrained by data assimilation they can be evaluated critically against observations. For deterministic processes, to within the error of the observations and forcing fields, we can evaluate prognostic models to assess which aspects of the model physics are poorly represented. There are processes that are not directly predictable given the forcing, and for these we can evaluate the model statistical accuracy.

There are known biases in the representation of physical processes in the ocean by models that results from a variety of sources including truncated physics, initial condition and boundary condition errors, and erroneous sub-grid scale parameterizations. The errors that result can be either random or systematic. For example Wunsch (2006) considers a systematic error of 1 mm/s in the Lagrangian velocity of a fluid parcel. At the end of 100 years of integration, a parcel of fluid will have a position error of 3000km. Random errors also lead to potential accumulation of very large divergence from the true average state even when it is driven by zero-mean fluctuations. The deficiencies of models must be taken into account in any comparison against observations, and the nature of the error from model biases must also be understood.

Near-global data sets include altimeter-derived sea surface height anomalies, velocities from surface drifting buoys, sea surface temperature (SST), upper ocean temperatures from Expendable-Bathythermograph (XBT) data and ARGO float data (http://www.argo.ucsd.edu/) that provides temperature and salinity of the upper 1000m. These data sets have been invaluable to the modeling community and have provided a zero-order measure of the realism of the upper ocean circulation. However, at high resolution, we can move beyond global measures and target specific regions where we know that models exhibit biases. We need to critically evaluate when possible, dynamical and thermodynamical balances in ocean models against observationally derived estimates.

To continue a discussion that was begun by the data assimilation community, a one-day workshop was held at the East-West Center at the University of Hawaii on February 25, 2006. A broad representation of the physical oceanographic community was invited to share in a discussion of how to define metrics for different modeling systems, the best way to communicate the metrics to the community, and what commonalities exist among different model users. The meeting began with a series of half hour presentations about the state of the art of ocean and coupled models as well as talks on data assimilation in ocean models. In the afternoon we divided into three groups: metrics for the operational and climate modeling communities were discussed in two groups while the third-focused on process oriented studies. The goals of the workshop were to establish a framework for progressing towards a consistent and more quantitative evaluation of ocean models, to promote a community wide discussion of model evaluation by both modelers and observationalists, and to discuss whether we can raise the bar on model evaluation.

The working groups were given a series of questions to consider:

1. How does the oceanographic community capitalize on observations to improve forecast and forward models?
2. Can we as a community develop a uniform approach to evaluation of models?
3. Can this effort be accessible to the entire community, be transparent yet evolve as models and observations change?
4. Can we be flexible to allow for novel ways to combine data?
5. Can we develop a process that facilitates intercomparisons?

For each modeling effort, in order to define the metrics, it is necessary to define clearly the purpose of the model that is being evaluated and to determine whether specific metrics will aid in the realization of the purpose of the modeling exercise. In addition, once a metric is defined in terms of observations, it must be easily repeatable and accessible to the user. It should also be possible for others to evaluate the model besides the modelers themselves.

Metrics for ocean models used in climate studies

The establishment of metrics for large-scale ocean only and the ocean component of climate models are more advanced than for other ocean model frameworks (see for instance Large and Danabasoglu, 2006). Existing metrics of large-scale model performance have largely focused on the mean state or climatology of the model, because of large common biases that still persist. Volume transports through key inter-basin exchanges such as the Drake Passage, the Indonesian Throughflow (ITF), and the Florida Straits, mean sea surface temperature and sea surface salinity, meridional heat transport, and seasonal sea ice extent are commonly used. Comparisons are made with the gridded data such as the WOA (World Ocean Atlas, http://www.nodc.noaa.gov/OCL/indprod.html) that gives a global three-dimensional temperature and salinity climatology. The great value of these data sets is that there is a value globally at every grid point at every time in a monthly climatology or in the annual mean. However, there are disadvantages: for example, WOA has data at every space-time grid point even when there were no data at that location. This can cause the density to be statistically unstable at the surface in one-third of the world's oceans. In the tropics, owing to horizontal averaging - the equatorial thermocline is over stratified owing to meridional averaging of the data (Large and Danabasoglu, 2006).

Errors are also inherent in the other metrics. The best transport record for the Gulf Stream is from cable data through the well-defined passage between Florida and the Bahamas, but definitive transport estimates from other locations are more difficult to obtain and rely on sparsely sampled current meter observations or one-time hydrographic sections. The best record for the Indonesian Through Flow is from the World Ocean Circulation Experiment (WOCE) IX1-XBT line between Java and Australia (Meyers, 1996). It provides long-term (over two decades) monitoring; however only the top 800m of the water column is sampled. Similarly, the complete transport is not available even for the relatively geostrophically constrained Drake Passage. For meridional heat transport the situation is even worse with estimates constructed from one-time sections with significant errors and aliasing. For each estimate, it is important to keep in mind that the period during which the data were collected is representative of a particular climatic regime.

The response of an ocean model to changes in the atmospheric surface forcing can be compared with those of the real ocean to give insight into the fidelity of the modeled dynamics, although both model error and forcing error are always present and may even be compensating. The best observed ocean variables are SST (sea surface temperature) and SSH (sea surface height). The relationships between simulated SST or heat content and heat flux or wind stress can be compared with those from observations on various time scales for particular regions such as in Western Boundary Currents and the tropical current systems. Also the correlation between sea level height and heat content could be compared with the correlation obtained from observations provided such correlations are from comparable time-scale space. Such correlations on a global basis could reveal deficiencies in the surface forcing and upper ocean processes such as advection and vertical mixing. Since sea surface temperature in ocean models can often be correct owing to the use of bulk formulae for turbulent heat fluxes which can hide model deficiencies, upper ocean heat content and mixed-layer depth may be a better test of model performance than SST.

Heat content can be calculated using a variety of data sets, each with its own set of errors. XBT sampling is biased to specific regions and tracks in the world ocean and thus has high errors. Only in recent years have profiling float data allowed better estimates, but the times series is still short for climate investigations and for both data sources, only the upper ocean is sampled. Altimetry also can provide estimates of ocean volume, however altimetry does not relate directly to heat content but rather reflects both heat and salinity (fresh water) variations, and salinity changes can confound its interpretation. Combinations of profiling float and XBT data, together with the remotely sensed observations are optimal. In addition, large-scale heat content can be obtained from acoustic thermometry. The wind stress products to be used in these comparisons must be considered carefully as there are multiple products, each with its own temporal and special sampling. For instance, Gille (2005) compared five different global gridded wind products with the Quick Scatterometer (QuikSCAT) swath winds and found that the JPL winds were somewhat low in energy compared with other gridded scatterometer fields. To capture the more extreme events in the wind forcing it may therefore be necessary to use a blended product that merges the high-wavenumber information available from observations with high-frequency numerical weather prediction fields. Reanalysis wind products from the atmospheric operational models such as the National Center for Environmental Prediction (NCEP1 and NCEP2 http://www.cpc.ncep.noaa.gov/products/wesley/reanalysis.html) and the European Center for Medium Range Forecasts (ERA40, http://www.ecmwf.int/products/data/archive/descriptions/e4/index.html) are often used to force models. Their longer time series along with consistent atmospheric thermodynamic variables make them useful for climate studies; however, NCEP2 winds are weak but better than the NCEP1, and one should be cautious using the heat fluxes that are calculated directly by the models. In each case, there must be a reliable observational estimate that can be used to evaluate the model, and the source of the model error must be determined, whether it comes from inaccurate model physics, or biases in the forcing fields.

Many ocean models do not correctly simulate coastal or equatorial upwelling, likely owing, at least in part to the lack of spatial resolution. Many issues arise when assessing the realism of simulated upwelling: how do we define it in a consistent way in ocean models, and how do we quantify an observationally based index that can be used as a model metric? How do we identify whether the ocean model is in error owing to the wind forcing, or to the model physics? To what do we compare the upwelling? Traditionally, the Bakun upwelling index (Bakun, 1973), based on longshore winds only, is used to define upwelling, but it can be noisy. Scatterometer data could be utilized as an additional wind field. An upwelling index related to SST could be very useful as upwelling is important not only in its own right, but also because of its impacts on heat fluxes.

Because of the opportunity of increased computing capacity, additional tests will have to be made on climate models as they move to eddy permitting and eddy resolving frameworks. These tests include eddy variability, decorrelation time scales and other eddy characteristics, and evaluation of the model against hydrographic sections. Some quantities can be obtained from the global drifting buoy data set at 15 m and from sea surface height derived from altimetry. Eddy statistics should be calculated from geostrophic velocity calculated from along-track sea surface height (SSH) data rather than an optimally interpolated product. While it is tempting to set up a set of standard metrics, the scientific question that is being address must always be taken into account when models are being evaluated, and the accuracy of the relevant data must also be taken into consideration.

Hydrography data, such as the repeat and one-time lines collected during WOCE, have a great deal of inherent variability and do pose a challenge for comparisons with eddy-resolving models. However, careful co-located extractions in time and space will still yield useful information, with error estimates based on the presence of eddies in both fields. Hydrography can be compared more directly with non-eddy resolving ocean simulations, once the hydrography is smoothed to the scales appropriate to the model.

A comparative measure of a particular quantity from two or more models can be made using a Taylor diagram (Taylor, 2001) that requires a correlation and standard deviation. It is a good technique to integrate into mainstream ocean model analysis.

Metrics for process oriented models and experiments

The aim of process studies in physical oceanography is to elucidate fundamental physical processes at work in the ocean, most often through a combination of models and observations. To achieve the goals of process oriented experiments it is useful to define metrics before process oriented experiments are designed. The requirements for model/data comparisons will vary greatly depending on the questions being asked and what specific dynamical regime is of interest and the metrics will not rely on climatological data, but rather will be tied to the time and the place of the particular experiment. At the core of the issue is the recognition that communication between modelers and observationalists requires easy access to both model and data. It was recognized that “best practices”, test cases, and metrics and diagnostics should be collated in a general repository to allow consistent evaluation of models. In order for the relationship to be successful, the benefits of participation for all parties must be articulated, as well as the requirements for their participation. For modelers, ease of comparison with observations will help to lend credibility to any modeling study and will help with the identification of model biases and weaknesses. For observationalists on the other hand, a more direct means of investigating the underlying dynamics of the ocean can be possible. In addition, better integration between models and observations will allow more effective design of experiments. Data sets derived from various satellite measurements and from direct measurements of the ocean are complex and diverse. Distillation of these different sets into a format that is relatively simple to use would facilitate comparisons among models and between models and observations.

The requirements for participation in any comparison study were outlined. For the observationalists, data that was used to generate figures should be available to anyone to facilitate reproduction of figures for comparisons, and error information must be given. In addition, the raw data should be made available as soon as possible after the data were collected in standard formats and along with analyses such as mean, variance etc. The benefits of data types and observing systems should also be outlined. For modelers, it should be possible for others to take a quick peak at model output including a list of past and coming model runs. In the description of the model runs modelers should make it a practice to explain the model design strategy, and what trade offs were made. Because the determination of physical processes often requires the determination of the balance of terms, there should be a forum for suggestions of what should be saved in model runs to facilitate future comparisons.

Operational Metrics

Production centers are chartered to provide products to a wide range of people who make decisions on possible courses of action. These decisions include short time scales such as where search and rescue operations must be conducted today to long time scales such as what crops to plant based on expected El Niño conditions. There are commonalities in the metrics for all time and space scale products.

First, the production center must have confidence that the model represents the dynamical mechanisms properly. This is the first step in model validation for a specific purpose. Historical observations are used in this process in much the same manner as those used in process studies. Operational metrics are used to identify shortcomings in the model physics, forcing fields and forecast approaches. Case studies of specific events need to be conducted when extensive observations are available, as this will enable a more complete evaluation of model deficiencies. For specific problem areas in a model, observations must be identified that can help to identify problems in model physics that once corrected could lead to increased model capabilities.

As numerical representation becomes more complex and as models represent a wider range of physics, it is necessary to encompass larger sets of data to validate a system for a range of applications. Once a numerical model has been validated for a specific purpose, production centers must understand the accuracy to which the system represents the physical mechanisms. The accuracy will certainly change from day to day depending on the environmental conditions. Some understanding of expected accuracy under particular conditions can be gained through historical observations. However, a daily monitoring of system performance relative to returned observations is necessary. The metrics at this point are physical performance metrics. That is, the metrics provide the production center with information on how well the system is representing the physics for which it has been validated.

On the opposite end of the spectrum of metric evaluations are impact metrics. These provide information on the accuracy of decisions based on products from the production centers. These metrics are crucial to the production centers from a number of aspects. The end impact guides the center forecasters to examine the model accuracy from the user perspective. In the operational community, an important task is to communicate model performance to the end user, and metrics should be tuned to the specific application in question (for example, determining the depth at which tuna hooks should be set based on the temperature structure throughout the water column). The user must be informed regarding the strengths and weaknesses of the particular product in question. Finally, the user must communicate back to the production center when the model is sufficiently accurate to be useful for their application.

The end impact is vital for the production center to demonstrate its relevance and maintain advocacy for continued operations. The end impact metrics also guide forecasters to understand which of the physical performance metrics are strongly related to the end impact metrics. This information then passes to numerical model developers. Focusing efforts to correct deficiencies in physical representation that relate to end impact constructs a solid case for future development work.

Thus, metrics in an operational setting are used as instrumentation in a complex system to understand the accuracy of information as it flows from numerical model physical representation to production center forecaster to end decision. This instrumentation provides the necessary data to understand the entire system and identify critical areas where there are shortfalls.

For operational applications, the demonstrated capabilities and expertise developed within GODAE should be used. An important aspect of operational metrics requires that model output and observations be available that can be read with standard software. This implies tools exist for forecasters to easily access model output and observations as well as construct metrics based on the two. To expand on GODAE (Global Ocean Data Assimilation Experiment http://www.usgodae.org/), the suite of metrics for model evaluations needs to be extended and observations that come from new observing systems such as IOOS (Integrate and Sustained Ocean Observations http://www.ocean.us/) and ORION (Ocean Research Interactive Observatory Networks http://www.orionprogram.org/) need to be made available quickly and be easily accessible. An example of such as effort is described in Chassignet et al (2006). The current metrics that are used in present systems should be collated and all models for particular applications should be submitted to the same standard metrics. The challenge for the operational community is to have those providing observational analyses that are used for model evaluation work closely with the modelers to facilitate ease of comparison.

The Next Steps

The discussions that took place during this workshop heightened our awareness of the importance of establishing metrics to assess the fidelity of the ocean models we are using in our particular application. Our climate models and short-term forecasting systems (even with a data assimilative capability) require the realistic representation of ocean processes.

The design of process oriented experiments must take into consideration the relevant questions being asked and the relevant scales. KESS (Kuroshio Extension Sytem Study, http://www.po.gso.uri.edu/dynamics/KESS/) is an example of such an experiment that was designed from the outset with model comparison and evaluation in mind. KESS included moored profilers, bottom mounted Inverted Echo Sounders with current meter and pressure sensor (CPIES), a surface mooring, floats, satellite measurements and ship surveys. The ship surveys included both oceanic (ADCP and CTD) and atmospheric (sounding) measurements. Understanding the processes that govern the variability of and the interaction between the Kuroshio Extension and the recirculation gyre is the goal of this study. Processes coupling the baroclinic and barotropic circulations are being examined by case studies of the local dynamical balances, particularly during strong meandering events. The mechanisms by which water masses are exchanged and modified as they cross the front will be characterized. The objective is to determine the processes governing the strength and structure of the recirculation gyres in relation to the meandering jet. It is the interaction between the models and the observations that will allow the questions to be answered, without either component, the experiment will not be successful.

There are several differences between monitoring systems and process studies. The main difference however is that in general, the process study includes several different variables, with sufficient spatial & temporal resolution and duration to resolve the process. The monitoring system generally focuses on a reduced (set) of variables, necessary for parameterizing a process or monitoring a response, and likewise typically trades off spatial resolution for increased duration. Many other experiments of successfully design process oriented experiments exist, and they have very different character from the design of monitoring observational systems that would be appropriate for many aspects of climate model evaluation.

Workshop participants suggested creating a website dedicated to the development and use of metrics for ocean-model evaluation. The site would feature, among other things, descriptions and links to observationally derived data sets as well as a comparison between metrics derived from ocean models and from observations.

Acknowledgements

The workshop was organized by **LuAnne Thompson**, University of Washington**, and Julie McClean**, Scripps Institution of Oceanography, and sponsored by the Office of Naval Research and the International Pacific Research Center. We would also like to thank Jim Potemra of IPRC as the local organizer, and Bill Large (National Center for Atmospheric Research), Gregg Jacobs (Naval Research Laboratory), and Sarah Gille (Scripps Institution of Oceanography) who helped lead the discussions.

References

Bakun, A. , 1973: Coastal upwelling indices, west coast of North America, 1946-71, *Tech. Rep. NMFS SSRF-671*, 103pp., Natl. Oceanic and Atmos. Admin., Seattle, Washington.

Chassignet, E. P., Harley, Hurlburt, O. M. Smedstad, G. R. Halliwell, P. J. Hogan, A. J. Wallcraft, and R. Bleck, Ocean Prediction with the Hyrbrid Coordinate Ocean model (HYCOM), 2006: in *Ocean Weather Forecasting: An Integrated View of Oceanography*. Chassignet, E.P., and J. Verron (Eds.) Springer, 577 pp.

Gille, S. T., 2005: Statistical Characterization of Zonal and Meridonal Ocean Wind Stress, *J. Atmos. Ocean. Tech.* **22**(9), 1353-1372.

Large,W. G. and G. Danabasoglu, 2006: Attribution and impacts of upper-ocean biases in CCSM3, *J. Clim.*, 19, 2325-2346.

Meyers, G., Variation of Indonesian throughflow and the El Niño-Southern Oscillation, 1996: *J. Geophys. Res.*, *101*, *12*, 255-12,263, 1996

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106,** 7183-7192

Wunsch, C. 2006: The Past and Future Ocean Circulation from a Contemporary Perspective. http://ocean.mit.edu/~cwunsch/papersonline/present_pastocean_19july.pdf