DESIGNING MULTIMODEL ENSEMBLES REQUIRE MEANINGFUL METHODOLOGIES

by Scott Sandgathe, Barbara Brown, Brian Etherton, and Edward Tollerud

Ensemble prediction systems have become ubiquitous as a means to capture forecast uncertainty in climate, global, regional, and mesoscale prediction systems. Ensembles of a single numerical prediction model have been proposed and developed as a solution to capturing initial condition uncertainty in weather forecasts, but model uncertainty is more difficult to capture and incorporate into ensemble systems. Consequently, the United States and other countries are formulating and running many different multimodel ensembles at climate, global, regional, and mesoscales in order to better capture model uncertainty. Ready examples are the North American Ensemble Forecast System (NAEFS), the National Multimodel Ensemble (NME), the National Unified Operational Prediction Capability (NUOPC), and the National Weather Service’s Short-Range Ensemble Forecast System (SREF). These ensembles are formulated from available members from different organizations, from different models run at a single site, or from variations of a single model. In some cases, studies are undertaken to formulate these ensembles; however, they are frequently constrained by available computational resources, available models, and the proposed operational environment. Numerous questions remain unanswered about optimally designing ensembles in order to effectively quantify initial condition and model uncertainty. This workshop was convened to review science issues governing the optimal use and configuration of single and multimodel ensembles (MMEs) at all scales and the simulation, quantification, and presentation of forecast uncertainty. One of the principal issues concerns the design of the core of a multimodel ensemble prediction system (EPS), which includes simulation of forecast uncertainty (from both the analysis and the model), postprocessing of raw ensemble output, and verification of probabilistic forecast skill.
The quantification and characterization of uncertainty was the main focus of the workshop. The following specific questions were addressed during the presentations and discussion:

1) What are the relative scientific merits and drawbacks of the multimodel approach versus an approach based on a single-model framework that incorporates stochastic forcing?
2) Given the merits and drawbacks of the two approaches, how can current ensemble systems advance to enable appropriate representation of model uncertainty?
3) Given the answers to questions 1 and 2, how should current multimodel ensemble systems appropriately quantify model uncertainty?
4) How should design aspects such as number of members, forecast length, model resolution, and update frequency be objectively tuned to optimally use existing computing resources to meet the needs of end users?
5) What metrics and approaches are most useful for verification and calibration to a) provide meaningful evaluations of the skill and value of probabilistic predictions that can guide further advancements, and b) provide meaningful postprocessed forecasts (e.g., without artificial clustering)?

Presentations from an international group of experts focused on this important set of issues and led to wide-ranging discussions regarding these topics as well as approaches to resolving the questions posed (see www.dtcenter.org/events/workshops12/noupc_2012/). In particular, discussions and presentations centered on the broad topic of uncertainty, with a specific discussion of the benefits and limitations of stochastic physics in representing uncertainty, the dependence of verification results on the selection of metrics for evaluation, and the use of information theory approaches for evaluation of the entire forecast system. A surprising concept proposed was that ensembles do not provide enough information to make rational decisions and that it might be better to use odds (for and against) as opposed to probabilities for decision making in many cases. Specific methods of addressing model uncertainty as well as approaches currently applied by the operational centers (the National Centers for Environmental Prediction, the Fleet Numerical Meteorology and Oceanography Center, the Air Force Weather Agency, and Environment Canada) and in experimentation (the Center for Analysis and Prediction of Storms at the University of Oklahoma and NOAA/Earth System Research Laboratory) were described, along with issues associated with capturing the appropriate sources of uncertainty associated with the ensemble systems. Methods for improving the ability of the ensemble systems to represent model uncertainty, such as the application of stochastic kinetic energy backscatter methods and the estimation of uncertainty in all model physical parameterizations, were also considered. Analog approaches to ensemble prediction and new postprocessing approaches, including new applications of Bayesian model averaging methods, were shown to have some promise in representing forecast system uncertainty. Finally, attributes to be considered in evaluating ensemble systems, as well as questions regarding the focus on spread–skill comparisons, were addressed.

An important result of the workshop was the lack of definitive answers to the questions posed. While some partial answers were provided through the presentations and subsequent interactions at the workshop, the discussions brought forward the fact that no clear and consistent evaluations have been undertaken that can provide unambiguous and meaningful answers to the questions that were initially posed. In addition, a number of additional questions were raised, such as these: How do we appropriately sample in a multimodel space? Is there more benefit in many forecast situations to assigning probabilities to objects (front, thunderstorm cluster, hurricane) than to grid point values? Is mean value and sensitivity of the mean a better decision tool than a probability distribution? Are we able to construct a tractable set of metrics that track forecast value for the large domain of users and applications and also provide diagnostic value? This need for methodologies to make objective comparisons of approaches for ensemble design is a paramount result of the workshop.

Thus, the main conclusion of the workshop participants was that a more scientific approach is needed to answer the ensemble design questions. Previous experiments have generally been driven by legacy systems, resource constraints, often inappropriate verification, and single forecast objectives, and have not been formulated to effectively answer important design questions. The research and operational communities need a more unified approach to dealing with key ensemble deployment questions, allocating resources for ensembles and also in evaluating the multitude of proposed methods to deal with model uncertainty. However, they also felt that the questions are sufficiently important and tractable that they
should be answered for current and future mesoscale, regional, global, and climate systems.

The participants discussed a collaborative framework where datasets and metrics/targets are standardized for careful intercomparison, such as in the Coupled Model Intercomparison Project (CMIP), the Hurricane Forecast Improvement Project (HFIP), and the Spatial Forecast Verification Methods Intercomparison Project. The framework or “playground” developed in such a project would allow researchers to answer critical design questions, test results against standard test cases and metrics, and allow for intercomparison of models, ensemble formulations, and methods of quantifying model uncertainty on an even playing field.

They proposed a follow-on meeting to define an experimental infrastructure and program to address optimal ensemble formulation. Specifically, the meeting would discuss ways to perform the following:

1) Establish a standard set of metrics for presentation of ensemble testing that will allow useful and valid intercomparison of ensemble formulations and methods for dealing with uncertainty.

2) Establish a small set of target parameters or scorecards to assess value and allow meaningful operational resource allocation.

3) Establish a global ensemble data archive for research, evaluation, and testing that includes the following:
   a) Standard test series (i.e., specific time periods)
   b) Ensemble initial conditions and perturbations
   c) Ensemble forecasts from major centers
   d) Easy access by the research community
   e) Metrics meeting recommendations 1 and 2 (above).

4) Establish a clear experimental program to address the workshop questions.

ACKNOWLEDGMENTS. The workshop was jointly sponsored by the National Unified Operational Prediction Capability (NUOPC) and the Developmental Testbed Center (DTC) Ensemble Task (DET).