DTC Ensemble Testbed
Design Plan

Created February 2011

Modified March 2011

The DTC Ensemble Testbed Objective

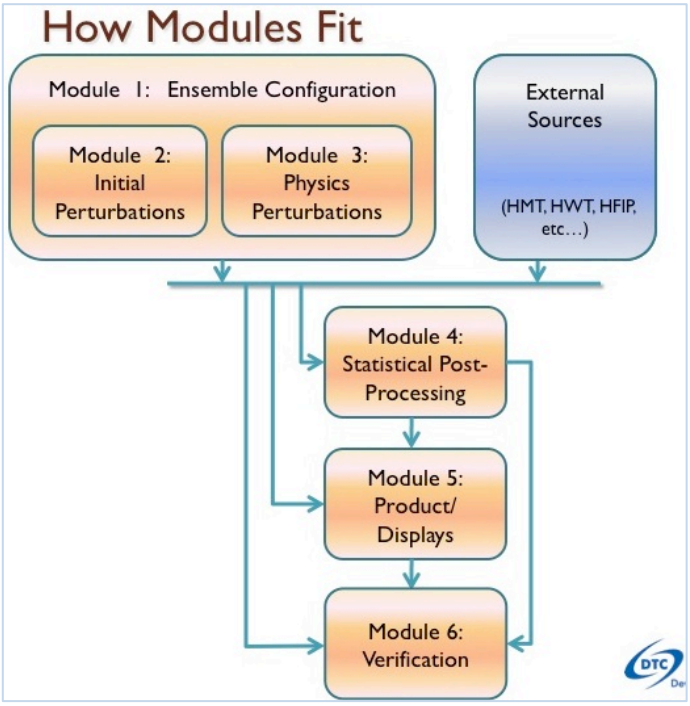
In response to the US NWS operational centers' move toward ensemble-based probabilistic forecasting, in March 2010 the DTC Ensemble Testbed (DET) was established by the DTC to expand its contribution in this important area. The goal of the Ensemble Testbed is to:

> Provide an environment in which extensive testing and evaluation of ensemble-related techniques developed by the National Weather Prediction (NWP) community can be conducted such that the results are immediately relevant to the national operational forecast centers.

When fully developed, Testbed activities will involve maintaining and supporting community ensemble software codes, as well as conducting extensive testing and evaluation of promising new capabilities and techniques that have been incorporated into these community codes. The community codes the DTC already supports (WRF, etc.) will serve as building blocks for the end-to-end ensemble testing and evaluation system assembled by the DET.

To facilitate testing and evaluation of competing techniques and capabilities for specific components of the ensemble system, the DET infrastructure is a modular design in that individual software modules represent separate concerns. To keep the testing and evaluation results of new ensemble capabilities developed by the research community relevant to operational upgrade decisions, the DET modules will need to be configured with the ability to replicate the algorithms used operational centers, namely, at NCEP and AFWA.

This plan outlines each of the eight elements of the Ensemble Testbed currently in development.



DET Overall Design

The DET design activities include establishing the software environment and tools to accomplish the task of running an end-to-end Ensemble Prediction System (EPS). Due to the complexity of the whole system, the Ensemble Testbed modular component design allows for flexibility, understanding, and management of the EPS infrastructure. By using a web-based graphical user interface tool, the Ensemble Testbed can interact with all the software modules in a unified and efficient way.

Basic Infrastructure

The six named modules to be included in the Ensemble Testbed infrastructure are 1) ensemble configuration, 2) initial perturbations, 3) model perturbations, 4) statistical post-processing, 5) product generation, and 6) verification. The figure here, called How Modules Fit, shows how the modular software components fit together. There is a distinction between the *ensemble generation modules,* 1-3 and the *use of ensemble information modules*, 4-6 to be described further.

System Considerations

The three *ensemble generation modules*, namely the initial conditions and physics perturbations, will require massively parallel processing systems on which to run. These three modules are to be designed to be portable across available systems as a subgroup. The system considerations for this type of computations include petascale calculations—a computer system capable of reaching one petaflop, i.e. one quadrillion floating-point operations per second, or greater. Some national compute resources that allow for this type of computing are the TeraGrid, DOE Incite, NSF Centers, Texas Advanced Computing Center, and Oak Ridge National Lab's Jaguar. The Ensemble Testbed needs to apply for the use of CPU cycles on these systems. Accessible through the Ensemble Testbed, we will use Jet and Bluefire as well as the NOAA HPC Consortium computational resources for the DTC.

The three *modules using information from the EPS*, statistical post-processing, product generation and verification modules, will take advantage of fast computer processing but will have less intensive CPU demands than the task of running the EPS. These latter three modules are likely to run together at one location, the DTC Center, and can be applied to model grids from ensembles run external to the DTC. This means that another organization could use the DTC software to post-process ensemble grids generated, for example at the University of Oklahoma, and take advantage of the suite of tools available to the DTC. This meets the community need for collaboration to carry out testing and evaluation of community methods and fosters involvement with the community. A second look at the figure, How Modules Fit, shows how External Sources fit in the Ensemble Testbed schema.

The Ensemble Testbed basic infrastructure will be defined in further detail on a DET website section to be developed that specifies the EPS and how to use it. Discussions are being conducted now on the software requirements, procedural instructions, file format specifications, data and protocols needed. This information is to be available on the website, as well. We consider the DET reference webpages to be a living document as we are using ever-changing technology. A living document is not one that is constantly being written but one that networks documents—a website to organize, share, interact, and communicate with the DET community. We will consider frameworks used on developer hubs, like addons.mozilla.org. This particular website has available a *how-to library, software interface and language references, developer's forums, and software validators* – all subject matters that the Testbed could use in its own developers' hub website subsection.

The Ensemble Testbed design goals include:

- Definition of software modularity to facilitate testing and, ultimately, the transition of software to operations, called R2O short for research to operations.
- Definition of what it takes for code to be portable, for ease of execution where computational resources are available.
- Requirement of flexibility to allow testing of new ideas with, hopefully, a plug-and-play approach.

Performance benchmarks will be established and used when testing and evaluating the EPS a) with various configurations and b) with community contributions to the EPS. Considerations for real-time

and retrospective use are important to the DET and we have started the EPS with retrospective use. Near real-time use would increase demands (staff, disk space, scripting, tolerance for failures of data arrival/ of computing systems) and will be considered secondarily.

System Requirements

The EPS model output formats that are acceptable to the DET are: GRIB1, GRIB2, and NetCDF. In the case of NCEP's Global Forecast Model, native (or binary) model output will be considered as a format within the DET. It is important to note that creating binary output is a trivial effort; the ability to reuse this data for input for model initialization, visualization, statistical post-processing and verification is not. A software development effort would be required to utilize binary formatted output within the DET.

Output data formats for products of the DET require either AWIPS-II NetCDF or CF (Climate and Forecasting)-Compliant NetCDF.

Proposed languages for the NWP models of the EPS are Fortran and C. Languages for DET products generation are Python, Fortran90 and C++. Languages to be used for the display of EPS output and post-processed output are Python, JavaScript (the language of dynamic web applications), and Java.

Computer scripts, text files that can contain commands that are executed in a particular order in an automated process, are critical to as complex a project as running an EPS. This plan recommends and allows XML, ksh, csh, etc. scripts to run and manage the Testbed EPS. Note that XML is the language of a tool helpful in running NWP models called Workflow Manager (http://wrfportal.org/ExternalWorkflowManager.html); it will be highly recommended to use this tool and it is discussed further in the Plan under ensemble configuration.

Software compilers to be used in this Testbed are Portland Group's PGI and Intel compilers with the likely use/support of other compilers as needed. As the DET evolves, a software revision control system will be used for the management of changes to programs, scripts, documents, and other information that will be stored. For this the Subversion revision control, SVN, will be used for the DET community's access.

Computer operating systems (OS) used by the DTC will vary. The DTC has access to Linux systems that run a variety of acceptable operating systems. Functional similar OS, like Red Hat, Suse, or Debian, etc, have differences that are so small they are not a focus of concern. The DTC is aware that the NCEP operational center runs computer processes on an IBM architecture running the AIX OS. Code developed on Linux OS will be ported and tested for AIX OS.

Areas of involvement and communication with the DET community have been considered. We propose a DET reference website portal with a "developer's hub" mentioned earlier with access to the software through a password-controlled revision control system. This web resource will instruct the community on how to acquire, run, and contribute to the development of the DET EPS. On the DET web portal (an information and services page), we will also provide links to a DET user forum, DET presentations, and feedback from the August 2010 DET Workshop and future workshops. The portal will also contain testing and evaluation plans, reports, results, collaborative tools, and tutorials.

Workshop Questions, Working group recommendations, DET Response

At the DET Workshop held in Boulder in late August 2010, the presenters asked questions of the audience as part of each presentation: (http://www.dtcenter.org/events/workshops10/det_workshop/index.php). To respond to the questions and foster discussions, we created three working groups for the purpose of the workshop, comprised of the attendees and presenters at the workshop. Each group, called a Working Group (WG), met in early afternoon and late afternoon breakout sessions for discussions. The same question was posed to each group in order to be answered from a variety of perspectives. WG one, discussed ensemble configuration and initial perturbation, WG two, discussed model perturbation and statistical post-processing, and WG three, discussed issues related to product generation, display, and verification. The recommendations from the Working Groups were recorded and DET responded to each recommendation in a separate, more complete document called *DET Working Group Recommendations DET Responses*. The, questions as you will see, have an identifier like (Mod1.2) followed by a response from the DET. The following apply to this section of the Plan.

(DET1.1) *Do you see a need for the Testbed to have a real-time capability? If so, where?*
Based on the WG recommendations, the Testbed will develop the EPS system for both real-time and retrospective capability. For retrospective runs, it was suggested to evaluate the performance level of the EPS as it relates to the CPU choices. As time allows for this sensitivity, testing will be conducted.

(DET1.2) *How do we make sure technology transfer to operations is possible?*
Based on WG recommendations, we will continue to plan for compiler continuity, standard coding practices, standard data formats, etc. We will also plan to continue coordinating with EMC for the transition to operations. There are also plans to have staff exchanges between DTC and NCEP, including an operationally useful testing platform.

(DET1.3) *How do we identify a list of candidate methods to consider for testing?*
The Testbed will evaluate what is in the field and what the members of the WG suggested.

(DET1.4) *How do we insure plug-and-play and portability?*
The WG suggested: keep dependencies to a minimum, be involved in other initiatives working to make common software structures (i.e. ESMF, the Earth System Modeling Framework), ensure all input data formats are accounted for, and have well documented test data and methodologies for creating new products, both from DET and the community. DET likes these ideas.

Ensemble Configuration, Module 1

The Testbed ensemble configuration module is to accomplish the task of configuring and running an end-to-end ensemble prediction system. The definition, development and utility of a web-based graphical user interface tool are the means for this to be achieved.

Motivation

The motivation for an ensemble system that is configurable came from the community, culminating in the National Mesoscale Workshop Probabilistic Prediction (NMPP) Workshop, which was held in Boulder in September 2009. The workshop and white paper asked questions of the community (http://www.weather.gov/ost/NMPP_white_paper_28May10_with%20sig%20page.pdf). There were science issues posed at the workshop to be answered about running ensembles. Which ensemble configuration offers better forecasting: An EPS with a small number of members run on a high-

resolution grid or an EPS with a large number members run on a course-resolution grid. A single model with multiple perturbations or an ensemble comprised of multiple models. There also needs to be choices of analysis with model perturbations.

Goals of Ensemble Configuration

With the above motivation in mind, the goal of this module's development effort is to define the NWP model configurations, which is ability to edit files, parameters and scripts in the process to setup and run ensembles. The goal of running the Testbed EPS system that you can configure is to facilitate testing and evaluation to obtain optimal results.

The ensemble design factors that dominate the computer resource demands for an operational EPS include are: model resolution horizontally and vertically, number of ensemble members, methods of initializing the ensemble, methods for simulating model uncertainty during the forecast integration, forecast length, domain size, forecast update frequency and forecast timeliness. To quote the September 2009 NMPP report, "Computational resources will always be a strong constraint in ensemble design. Balancing the effort devoted to each factor should be accomplished in a way that serves needs."

The User Interface Solution

Running an EPS, even a simple one, can be a complex undertaking. A user interface eases this challenge. Graphical User Interfaces (GUIs) have windows that enable users to easily view, control and manipulate multiple things at once and see relationships and dependencies between the data and software components. With use of a GUI, tasks are commonly completed much faster than can be done manually and monitoring of jobs visually becomes possible. DET has joined forces with the developers of the DTC WRF Portal GUI, a graphical user interface for defining, running and monitoring an NWP model. The WRF Portal GUI is a web-based tool written in Java. It is actively being funded and supported. The ability to extend the WRF Portal for use as a graphical tool to run an EPS is being developed. DET had a series of meetings and has made a list of requirements that can allow the DET GUI to meet the needs of and ease the complexity of running the EPS for the DET community. Note that the use of a GUI does not preclude the ability to change namelists, edits scripts or control job submission from the command line and through Linux-type file editors.

The key to the GUI that is being developed is a mature interface to DTC's Workflow Manager, a script-generating tool that *is the communication interface* to the computational resource of choice.

First Year Deliverables

The basic infrastructure has been defined and a basic capability of the system will be demonstrated this year. With the use of the GUI we will run ensembles and begin to consider answers to some scientific questions. We will also make status reports available on the DET website for the ensemble community.

Workshop Questions, Working group recommendations, DET Response

(Mod1.1) *What other system should we consider in our initial investigation?*
Who are the contacts for the systems? The DET will add the system recommended by the Working Groups to the list presented at the DET Workshop.

(Mod1.2) *What are the similarities and differences between how the current ensemble systems are engineered?*

DET will keep these issues in mind while developing the system. EMC tends to prefer many executables to one, NEMS is an instance of ESMF, our real-time system will require fault tolerance; our scripts will be written in commonly used languages

Initial and Lateral Boundary Condition perturbations, Module 2

The Testbed initial perturbations module allows for decisions to be made about the initial conditions (IC) and lateral boundary condition (LBC) perturbations—an important part of a regional EPS configuration. Early in the development of ensemble forecasting technique, ensembles were designed based on perturbed initial conditions only, and the ensemble mean values, ultimate products of the EPS, were found to estimate the verifying state (usually large-scale circulations) better than the forecast from a single ensemble member.

Motivation

Today different centers use different approaches to the challenge of initial perturbations. Knowing how other forecast centers set up EPS is important for the DTC Ensemble Testbed to learn, consider, and evaluate. Usually the initial and LBC question for regional EPS is addressed by using the corresponding global EPS.

One such example to consider is the UK Met Office routine running of their Global and Regional Ensemble Prediction System (MOGREPS). The purpose of the regional EPS was to provide uncertainties in short-range forecasts, while the main purpose of the global EPS was to provide LBCs for the regional EPS. Both the global and the regional parts of the system utilized the same initial condition perturbations generated using the Ensemble Transform Kalman Filter (ETKF) approach. In the original implementation of MOGREPS, the regional component provided 'dynamical downscaling'. The idea was to have high-resolution initial condition perturbations specifically for the regional EPS. In 2007 a version of MOGREPS was implemented, in which the ETKF produced the initial perturbations for the regional component. This system showed that regional perturbations contain more detail on the small scale and that these small-scale details last for less than eighteen hours. Results from studying this influenced the decision to reverse the regional component of EPS back to its original version—using a global component as a driver.

Similarly, at Environment Canada, the lateral boundary conditions are provided by the Canadian operational global EPS. Each of the 20-member EPS drives one of the 20-member regional EPS, via dynamical downscaling. The initial condition perturbation approach used for the global EPS is the Ensemble Kalman Filter (EnKF). Recently, an effort to produce regional EnKF initial condition perturbations has begun.

NOAA's National Center for Environmental Prediction / Environmental Modeling Center (EMC) creates the Global Ensemble Forecasting System (GEFS) and the Short Range Ensemble Forecast System (SREF). In 1992 a breeding technique was employed for GEFS. In 2000 the same technique was implemented for the regional EPS, SREF. In 2006 a change was made to GEFS, the Ensemble Transform with Rescaling (ETR) technique was implemented as a way of handling initial condition perturbations. The same technique has, unfortunately, not yet been implemented for the regional EPS. Consequently, there is an inconsistency between the global and regional components of the EPS at this time.

Further, different EPS groups in the community use a variety of approaches to the challenge of addressing initial perturbations. Some of these community systems evaluated are from: CAPS

Norman-Oklahoma, RENCI North Carolina, HMT Boulder-Colorado, University of Washington Seattle-Washington, NCAR Boulder-Colorado's JME and others. The initial condition approach among these groups includes variation in numerical model output used to initialize the next forecast cycle, use of 3D VAR techniques, and use of various ensemble data assimilation systems including EnKF and ETKF.

Objectives

Addressing initial and lateral boundary conditions perturbations, the main objectives for the DTC Ensemble Testbed are:

- Become aware of approaches to address initial and lateral boundary conditions currently in use
- Create infrastructure that allows for testing of different methods to capture analysis uncertainty and uncertainty associated with LBCs.
- Port NCEP SREF system to the DTC as benchmark for testing.
- Test various community-developed methods for IC and LBC perturbations.
- Work with NCEP and other agencies on transition of selected methods to operations.

As a first step for the Testbed, we will set up a regional ensemble forecasting system that replicates the SREF operated by NCEP. The SREF-like system will be used as a benchmark. To provide initial and LBCs for the regional component, the NCEP global ensemble prediction system will be used. This agrees well with the Working Group recommendations (Mod2.1) and (Mod2.2).

As a part of the Ensemble Testbed effort in regional EPS and for initial and boundary perturbations we will test the following options:

1. Test cycling approach when using GEFS as a step beyond simple interpolation; The Global and Regional EPS simple interpolation approach will be compared with the cycling approach.
2. Evaluate impact of using various analyses (e.g. RUC, NAM, Rapid Refresh…).
3. Evaluate regional EnKF.
4. Evaluate regional ETKF.
5. Use quasi-orthogonal perturbations for both global and regional components of the current NCEP/EMC system.
6. Create regional ETR initial perturbations independently of the global component.

A large part of the infrastructure for testing various options for initial and lateral boundary condition perturbations will be addressed by the development and use of a DTC Ensemble Testbed graphical user interface. Issues related to portability and modularity will be taken care of by using this tool and the overall design plan. Details related to GUI tool and its early features are provided in the first sections of this plan.

Timeline and Milestones

Year 1 – Development the infrastructure and preliminary tests of the cycling approach.
Year 2 – Implement additional perturbation methods and testing of user's community requests.
Year 3 – Continue to focus on testing community-developed techniques and on transitioning of promising techniques to NCEP for operational consideration.

Model perturbations, Module 3

Numerical models used for the generation of ensembles are continuously developed to be more realistic. Nevertheless, even if the state of the atmosphere were exactly known, they cannot reproduce reality due to various truncations involved, including the use of finite time steps, a particular spatial resolution, and the choice of physics used. Processes not resolved in the models

due to truncations and possibly other approximations and errors give rise to a further divergence of numerical forecasts from reality, beyond that caused by unavoidable errors in the initial conditions. A well designed ensemble will attempt to capture forecast uncertainty associated not only with the initial conditions but also with the use of imperfect models.

In constructing a forecast ensemble, it is imperative to account for model uncertainty. The inability to thoroughly account for model error in an ensemble is likely a major contributor to the commonly observed problem of under dispersion among ensemble members. Though there is a growing literature associated with the study of model errors, the theoretical foundation for this line of work is much less developed than that for initial condition related errors. Therefore no commonly accepted theory exist to guide efforts aimed at representing model related errors in ensemble forecasting.

A number of techniques have been proposed for representing model error-related uncertainty in an ensemble, including, but not limited to:

- **Multi-model** – Different models and/or different physics schemes among the members (e.g., Stensrud et al. 2000).
- **Stochastic Physics** – Perturbations (which may be formulated with spatial/temporal structures or other dependencies) to state variables' tendency during model integration (e.g., Buizza et al. 1999).
- **Stochastic Backscatter** – Return dissipated energy via scale-dependent perturbations to wind field (e.g., Berner et al. 2009).
- **Random Parameters** – Random perturbations to physics parameters (e.g., entrainment rate), which may be fixed prior to model integration or varied during model integration (e.g., Bowler et al. 2008).
- **Perturbed Surface Parameters** – Perturbations to surface temperature, albedo, roughness length, etc., which may be fixed before model integration or varied during model integration (e.g., Eckel and Mass 2005).
- **Stochastic Parameterizations** – Explicit modeling of the stochastic nature of subgrid-scale processes (e.g., Teixeira and Reynolds 2008).

Historically, the multi-model approach is more common in regional ensemble systems. The NCEP SREF, for example, uses different forecast models and different physics (most notably, convective parameterization). The multi-model ensemble is a pragmatic approach in which different models or model versions are being developed by the community. Often, it is diversity of opportunity, taking present models or physics packages 'off the shelf', trying them in ensembles, and assessing the impacts. In general, this approach is not as scientifically well founded as methodologies for developing model initial conditions.

Stochastic approaches involve more development than multi-model approaches, but this development has been shown to have a positive impact on model performance in the references above. Stochastic approaches have a more solid scientific foundation than multi-model approaches, with regards to the proper amount and direction of ensemble spread. To develop stochastic approaches, one needs a single model (or model physics package) that can represent all model-related uncertainty by design. This uncertainty includes the effects of sub-grid scale and other unresolved processes. This approach will require the community to embrace a new paradigm to both focus on model development for ensemble use and to allow different groups work on various aspects of same model to be considered.

As part of the ensemble testbed, we will test what is in operations while allowing for the opportunity to explore new ideas. The model perturbations development effort objectives are to: 1) create infrastructure that allows for testing different methods to capture model related uncertainty, 2) to test various community-developed methods for creation of model perturbations, and 3) to work with community and agencies on the suggested configuration.

We plan to learn more about the choices of different models or model physics packages. Deciding, for example, which design guidelines to use. The options for differences in the models are the NMM-B and WRF-ARW (when NEMS compliant) models, and for physical parameterizations (microphysics, PBL, convective schemes, etc., different options in) the options within those models. For stochastic methods, such options will have to be developed and/or imported from the community.

Benchmark

Define next-generation benchmark using operational model configurations at NCEP/EMC (e.g., the NAEFS/SREF ensemble system) as the benchmark to conduct parallel testing of new methods. Current NCEP SREF Operational Systems include 10 WRF members (Oct. 2009) and 11 Eta and RSM members, noting that Eta/RSM members will be removed in 2011.

Major tasks

The major tasks of this module development start with creating an environment in which we can test model related uncertainty within the Ensemble Testbed framework. Then, we will design the tests for multiple dynamic cores, physical packages, and stochastic methods by providing the test platform for the community-developed portable techniques and conducting inter-comparison of different methods.

Timeline and Milestones

Year 1 – Develop a plan for three selected options (dynamic core, physics and stochastic approach)
Year 2 – Work on developing the benchmark
Year 3 and 4 – Test the three options and provide recommendations to NCEP and /or other agencies for operational consideration

Workshop Questions, Working group recommendations, DET Response

(Mod3.1) *How can we have a stochastic forcing that is more generalized for use with several model cores?*

DET will consider defining tendency forcing and internal physics forcing per recommendations of the Working Group. Tendency forcing:

- General tool set/modules that could be plugged into different models to draw spatially and temporally correlated structures, subject to amplitude determination and specific model variables.
- Module to apply multi-scale (forced backscatter) perturbations.
- Going much more general is difficult (if not impossible)
- DET/DTC has a role to make tuning of these schemes more efficient.

Internal physics forcing:

- Pre-tuned/tested parameter distributions could be allowed for models that already have the code built in (e.g. Grell Cu scheme).
- Scripting layer could support namelist options for those models that do support e.g. parameter values or stochasticity.

(Mod3.2) *What methods and metrics should be used to separate the uncertainty generated by initial perturbations versus model core uncertainty?*
A suggestion for consideration is that methods are not known well enough. DET/DTC supports this research question by facilitating easy ensemble runs using both model and IC uncertainty, and DA.

(Mod3.3) *Are there suggestions for how to execute core interchangability (i.e. ARW, NMM-e, NMM-b initially but others in the future)?*
Recommendations:

- Specify requirements for input and output.
- DET provides regridding tools for input and output (NEMS could help).
- Use small amount of DET/DTC resources to collaborate on working out interface details (consulting, etc). Maybe visitor program to facilitate.
- DET should encourage common map projections being adopted by any participating models.

DET response….

(Mod3.4) *Is there clear guidance on which to test first: Single model with stochastic forcing or multi-model?*
Recommendations: SREF 2011 (multi-model) is the benchmark; No clear guidance on testing, except National Mesoscale Probabilistic Prediction White Paper recommends research thereof; Research improving/optimizing multi-model ensemble, then compare to performance of single model ensemble with thorough stochastic forcing; Test combining both approaches
DET response….

(Mod3.5) *Other comments or questions?*
This generated the suggestion that DET use 1-D model to cheaply investigate some research questions e.g. investigate sensitivities of stochastic forcing techniques. DET acknowledges this idea.

Statistical Post-Processing, Module 4 (to be edited further)

Ensemble post-processing is essential to producing high-quality probabilistic forecasts. A probabilistic forecast is a probability distribution of a future weather quantity or event. A calibrated forecast is one in which intervals or events that we declare to have probability P happen a proportion P of the time. A sharp forecast is one in which prediction intervals are narrower on average than those obtained from climatology (i.e. the long run marginal distribution); the narrower the better. The goal of post processing is to maximize sharpness subject to calibration (Gneiting et al 2003).

In the recent past, there has been significant progress in four types of ensemble post-processing methods, those being (1) Bayesian Processing of Forecasts algorithm (Krzysztofowicz and Evans 2008); (2) Bayesian Model Averaging (BMA, Raftery et al. 2005; Wilson et al. 2007; Berrocal et al. 2007; Sloughter et al. 2007, 2009; Fraley et al. 2009); (3) applications of and technique development with reforecast training data sets (Hamill and Whitaker, 2006, 2007; Wilks and Hamill 2007; Hagedorn et al. 2008; Hamill et al. 2008; Fundel et al. 2009a, 2009b); and (4) ensemble-MOS techniques (Gneiting et al 2005; Glahn et al. 2009; Unger et al. 2009).

The goal of DET post-processing work is to examine state-of-the-art post-processing techniques to integrate information in ensemble and high-resolution control forecasts with the latest observations to produce bias-corrected and calibrated ensemble data, and to facilitate future operational

implementations of those techniques. This work will be a collaborative effort among ESRL/GSD, NCAR, and NCEP. Simple algorithms are preferable to complex ones, given the cost of code maintenance and for similar reasons, maintaining only one algorithm that works well for any variable (both raw model output and derived quantities) is preferable to maintaining multiple, specialized algorithms.

Definition of functionalities

- Bias correct individual ensemble members in ensemble forecasts and reduce lead-time dependent biases. The reference is always analysis data on model grids.
- Fuse all available ensemble model information (NAEFS, SREF, HMT etc.) to produce better ensemble distribution with better spread on model grids
- Downscaling to high-resolution grids
- Fine-resolution analysis data (e.g., RTMA, CCPA, LAPS)
- New derived user relevant variables
- Transfer promising techniques to NCEP/EMC for further testing and possible operational implementation

Modularity/portability

- Develop software and other necessary infrastructure supporting modular DET
- Build the modules for bias correction and downscaling procedures
- Post-processing module/model/code
- Input data (NetCdf, GRIB, binary) and data conversion
- Output data (ASCII, binary, NetCdf, GRIB) and data conversion
- Running job and submission scripts
- Graphical package for statistics (NCL, Grads, IDL)
- Instruction documentation to build the system
- Subroutines, scripts, modules are portable at different platforms and each component easily be plugged in by users
- Language (Fortran 90/95, C, Unix shell scripts) and common compilers (e.g., PGI, GCC)

Operational user requirements

With regards to NCEP operational requirements: generate statistically bias free, calibrated, and dynamically consistent ensemble members on high-resolution user grids such as the NDFD grid, including all parameters of interest. Achieve this with stability across platforms, reliability, computational efficiency, and common data/language format. Use a design that facilitates collaborative development and minimizes redundancy in development and operations. With regards to research developer and user requirements: design a flexible system (common data/ language/ compiler/ format) with ample documentation (user's guide) and clear rules of engagement. Ensure that the flexible design leads to easy transitioning of the techniques. With regards to end user requirements: build a user-friendly interface to provide forecast information in a choice of display formats for various environmental forecast applications. Provide a verification package to give end users tools for assessment.

Benchmark

Define current or preferably next-generation benchmark, such as using operational bias correction techniques operated at NCEP/EMC (e.g., the NAEFS ensemble system), as the benchmark to conduct parallel testing of new methods.

Major tasks

- Set up the prototype of bias correction and downscaling procedures in the DET framework
- Establish workflow of the implementation steps (compile the code, input, running code, and output) at several standard platforms (e.g., Jet system at ESRL/GSD)
- Collect ensemble data and provide DET subset testing data for the community-developed portable techniques
- Set up procedures to transfer the techniques to operational centers (NCEP)

Timeline and Milestones

Pending on computing requirements and model set up.

Year 1 – Develop detail plan for selected functionalities
Year 2 – Initial software development and preliminary testing on sample data and other ensemble systems (e.g., NAEFS)
Year 3 – Development of other functionalities and testing of prototype for the DET ensemble system
Year 4 – Testing community-developed new techniques for various functionalities and transitioning of the promising techniques to NCEP for operational consideration

Workshop Questions, Working group recommendations, DET Response

(Mod4.1) *What test data should we consider for Statistical Post-Processing?*
The recommendations: Need lots/ years of data and observations at many scales

- Training (dependent) dataset – 2008?
- Validation (independent) dataset – 2009?
- "Ground truth" data – gridded analysis and observations. Need information on error in the ground truth

Products and Displays, Module 5

There is a growing demand for easy-to-interpret meteorological products created from ensemble model output within the meteorological community. Well-thought out algorithm calculations can process model output to create derived parameters, probabilistic guidance and visual output, etc. that allow the end-user to make more informed decisions of the probability that a weather-sensitive event may occur. Research activities in the use of forecast uncertainty throughout the universities, NOAA testbeds, and operational forecast centers are identifying promising processing techniques that should be evaluated for inclusion within this setting.

The general approaches to creating ensemble products include: one, pre-process the ensemble into desired information in intermediary binary files such as GRIB or NetCDF and two, use a visualization tool with the dataset, selecting options of interest from a graphical user interface. Both methods utilize algorithm calculations to interact with the data.

There are many operational forecast organizations producing and visualizing ensemble products. And, there are operational workstation applications from which to draw comparative benchmarks for the Products and Display module. Those evaluated are NAWIPS, AWIPS, AWIPS II, ALPS, and web-based technologies like used at UKMET and AFWA. Keeping in mind the goal of R2O, from the products perspective, the first comparative benchmark will include calculation of single-value and probabilistic products similar to those generated at NCEP/EMC. For displays, there will be the ensemble imagery and interrogation tools available on the NAWIPS system.

Objectives

The primary goal of this module is to provide the community with a means to explore and test product generation methods and display capability. This work will provide the ability to specify techniques for deriving single-value (deterministic-style) forecast products and probabilistic products from the ensemble. This module will read in output from the end-to-end ensemble prediction system for the DTC Ensemble Testbed program as well as outside organizations (see Figure). This will be a collaborative effort among ESRL/GSD, NCAR, NCEP, and the research community.

Initial development will focus on tackling some of the most basic ensemble product needs including:

1. Evaluating methods for determining probability of occurrence at grid point (i.e. straight calculation, weighted calculation, applying a smoothing filter before or after).
2. Exploring ways to make ensemble spaghetti plots more meaningful.
3. Evaluating single-value (deterministic-style) ensemble products for scientific integrity and human impact.
4. Development interface to allow model output to plug into operational platforms including NAWIPS and AWIPS II.
5. Development interface to allow model output to plug into decision support products.

To achieve testing and evaluation of meteorological products for transition to operational centers, such as NCEP and AFWA, the module will need to utilize reliable, computationally efficient algorithms. System flexibility, ease-of-use, and documentation of system interface and constraints is important to engage the research community. Common language and data formats are mentioned in the overall design section.

Points of Reference and Discussion

AWIPS is not considered as a benchmark because there are no ensemble products available at this time. Phase I of AWIPS II development will be a functional equivalent of AWIPS and hence no ensemble products are yet included, until Phase II development. While AWIPS II is under development, a substitute benchmark could be the ensemble capability available in the Advanced Linux Prototype System (ALPS) an AWIPS-like system developed at NOAA/GSD. Some distinctions in the forecast workstation application referenced above are discussed in more detail here:

1) Advanced Weather Interactive Processing System (AWIPS) is an interactive computer system that integrates meteorological and hydrological data, enabling forecasters to prepare forecasts and issue warnings. It includes a full suite of satellite imagery, radar data, surface observations, and numerical model guidance. The primary tool of AWIPS is a graphical software application that incorporates most available weather information into an easy-to-use interface. The system is used at National Weather Service (NWS) Weather Forecast Offices (WFO'), River Forecast Centers, and National Centers for Environmental Prediction to support weather and hydrologic forecast and warning operations. Ensemble products are not available in AWIPS, the future for the NWS is the operational AWIPS II discussed in more detail below.

2) NCEP Advanced Weather Interactive Processing System (N-AWIPS) is a solely separate tool from AWIPS that targets the needs of a different and diverse user base: NCEP Operational Centers, namely AWC, CPC, HPC, HNHC, OPC, SPC, and +SWPC; NWS Alaska Region and Pacific Region WFO and River Forecast Centers; Unidata (supporting about 300 universities), and other government labs. The with broader customer base, the tool operates on the full spectrum of geographic and temporal scales, local to global and seconds to months. N-AWIPS has a diverse set of forecast products; it provides access to operational and experimental numerical model graphics and grids, geostationary satellite

data, surface and upper-air observations, text products, and ensemble imagery and interrogation tools.

This system consists of the GEMPAK processing software and the NAWIPS graphical user interface programs. Perhaps the most significant feature of N-AWIPS is its GEMPAK-based (desJardins and Petersen, 1985) grid access module, which allows forecasters to perform virtually any mathematical operation on model grids. N-AWIPS has an X-based display system; it is a distributed computing and communications system software package.

3) AWIPS-II is the AWIPS Software Re-Architecture Product Improvement Plan, authored by Raytheon, and written in Java and built on a service-oriented architecture. This unifies tools like N-AWISP and AWIPS and others. A new hardware definition is required and defined by Raytheon. This allows the use of new tools and technologies like GeoTools, etc, requiring adopting and adapting and moving forward. The AWIPS II structure takes advantage of open source practices and software where key attributes include non-proprietary software, high performance data services using advanced data serialization techniques, situational awareness and decision-making visualization, visualization is customizable through XML files and scripts, warnings and reports through GIS interactions and automated text generation through template engine, and it adaptable to a variety of data types and scalable from laptop to servers. The first release of AWIPS II is anticipated mid-2011.

4) ALPS' objective started as an effort to accelerate the transition of NWS to an all-Linux AWIPS system architecture and to address some of the anticipated near-term AWIPS system challenges prior to release of AWIPS II. ALPS software architecture decomposes functionality into components, that can reside within or beyond process boundaries, across host/platform and site boundaries, which communicate with each other via well defined interfaces. ALPS has a suite of ensemble product displays. With the system you can show a plume diagram of surface temperature for an N-number ensemble, the mean of the plumes overlaid on the raw data plume plot, screen shot of a spaghetti plot of 500 heights from the 21-member SREF ensemble, the previous data description displayed with a standard deviation of the ensemble members overlay, a suite of probability of precipitation threshold graphics for several accumulation thresholds, and imagery with the probability of exceeding flash flood guidance converted to an image, etc.

5) Web-based display method solutions will be researched and evaluated. The UK Met Office's web site, to quote, "is a comprehensive source of weather information. It offers news, a customizable forecast and a bookmarks feature for quick access to sections. The Weather tab links to information for the UK, Europe and the World, as well as aviation and marine data. Select UK to view weather warnings, regularly updated and fully customizable five-day forecasts, weather statistics and more".

The Air Force Weather Agency (AFWA) acquires weather satellite imagery from several orbiting spacecraft, other observation, runs forecast models and ensemble models. The products are updated frequently, every one to three hours and are available for a variety of geographic regions or theaters. Recently AFWA has moved in the direction of SOA, service oriented architecture, in the form of a web-based workstation for forecaster and staff in the field needing weather information. The code is written in Java but converted to JavaScript for use with any browser. There are AF weather products available to the public, but the operational system has restricted access. The meteorological imagery is geo-referenced and plotted on Google Maps. We will evaluate the ensemble products available from AFWA for consideration for broader use.

Timelines and Milestones

Year 1 – Develop detail plan for selected functionalities
Year 2 – Determine framework and optimal languages for modularity and plug-compatibility with N-AWIPS and AWIPS-II. Develop generalized framework for both Product Generation and Display.
Year 3 – Test system using methods already developed by collaborating organizations. Solicit and evaluate additional community contributed methods for testing.
Year 4 – Test additional method. Set up procedures tech transfer to operational centers (i.e. NCEP, AFWA, NWS Offices and Centers, etc.).

Workshop Questions, Working group recommendations, DET Response
(Mod5.1) *How do we interface with AWIPS II development*?
Recommendations included a need more information on when AWIPS II will be deployed and what the follow-on development will be. DET should not expend resources, but should only play an advocacy/awareness role. Look beyond to NextGen. This seems reasonable.

(Mod5.2) *Should we think beyond AWIPS II to web-based displays for the research community?*
The recommendation is that web-based displays should be pursued. Web-based is primary for AFWA, also useful for testbeds. DET agrees.

(Mod5.3) *How should we include social scientists in this portion of the testbed?*
Recommendations: Engage social scientists to participate as possible in testbeds and to also pursue external funding sources for more extensive participation.
DET response….

(Mod5.4) *What is a prioritized list of ensemble-relevant forecast products?*
Recommendations:

- Ensembles absolutely need to be researched and developed so that they become a useful tool for alerting users to high impact events
- If ensembles can provide useful info on impact high events, how do you extract it and display it? Will need to meet various levels of complexity
- What kind of forecast info is required?
- Garden variety weather impacts many users (e.g. energy use, financial market, agriculture)
- Tools for developers, researchers

DET will explore these products.

(Mod5.5) *What is a prioritized list of ensemble-relevant display algorithms?*
Recommendations:

- How people are going to display data
- Interactive tools and displays that user controls
- Spaghetti-grams, make more useful
- Sounding diagnostics for ensembles
- Effective visualizations that “grab” attention
- Web II to address interrogation needs
- Mobile technologies
- ENSEMBLE MEAN: Not on model or atmosphere’s attractor

DET will explore these display algorithms.

(Mod5.6) *How do we develop and solicit new display possibilities?*
Recommendations:

- Ensemble products workshop
- Strong representation of appropriate end users and decision makers at workshop
- AMS presentation
- Search for "entities" that have *avant-garde* tools and ideas that address user needs
- Start a forum to search info

DET will explore these avenues. Some consideration for an Ensemble Products Development Workshop II similar to the 20-21 April 2010 Workshop held in Boulder, CO at NCAR. (http://www.ral.ucar.edu/jnt/tcmt/events/2010/hfip_ensemble_workshop/)

Verification, Module 6

Assessment activities including quantitative verification will be a critical role for the Ensemble Testbed. The utilities that will facilitate that role will be both retrospective and near-real-time in nature, the former to accommodate both ambitious re-analysis and sensitivity studies similar to several ongoing tasks in the DTC, and the latter to complement real-time forecasting demonstration projects such as those currently operating in HMT and HWT collaborations with the DTC.

The tools to achieve verification tasks for the DET will of necessity focus primarily on probabilistic verification techniques. Within the broad scope of these techniques, three roughly separate areas of development are necessary: 1) ensemble preparation and processing; 2) score definition and computation; and 3) results display. This section focuses on needs specific for verification; other post-processing tasks are described in Modules 4 and 5.

As will be emphasized in a later section, there are many tasks and techniques that are common to the objectives of the DET as well as to those of several other projects of the DTC. It is important that these common areas are carefully considered in the course of development of DET plans, both to leverage as much effort as possible and to keep the DTC components as inter-connected as possible. With respect to the verification plans of the DET, the HMT and HWT accomplishments and infrastructure are of particular interest. Both have designed and implemented demonstrations of real-time verification systems for forecast exercises in these testbeds that have heavily leveraged off each other and that can be considered as prototypes for the DET verification module. Since both have employed Model Evaluation Tools (MET) verification utilities—a DTC-sponsored and funded verification package. It is proposed that MET will continue to be the principal verification software for DET. However, DET will require additional utilities not yet available in MET, so that it is also important to investigate other interesting packages and techniques that are available for either adaptation or application within MET or for stand-alone use by themselves. The next section describes MET and a other leading product candidates and their potential usefulness for the DET.

Existing Verification Packages

Existing utilities in MET include a large set of scoring and data-ingest options, both for probabilistic and deterministic forecasts. For probabilistic uses, it now incorporates scores including the Brier score and its decomposition. It also includes the computation of the Receiver Operator Characteristic (ROC) Curve, the area under the ROC, the points for the reliability diagram, calibration, refinement, likelihood, base rate and ranked probability scores.

A Brier Skill Score can be calculated assuming the sample climatology and using the decomposition of the Brier Score. A method to define or ingest a standard climatology field from which skill scores may be produced is not included at this time but is feasible. MET routines require that user probabilities must be produced off-line and ingested into the MET workflow. However, basic capability within MET includes a module to process ensemble model data and thereby produce the simple arithmetically defined probabilities and other parameters used in probabilistic forecast verification.

On the evaluation post-processing side, a prototype version of 'METViewer', database and display package based on 'R' statistical routines and graphics, is available for use by DET. This package currently allows for the calculation of median values across user-defined stratifications (i.e. time, region of interest, thresholds, etc…) for all statistics currently available in MET. METViewer will also display the accumulated rank histograms across user-defined stratifications. Displays of aggregated statistics as well as Reliability, Attribute, and ROC diagrams are still to be added.

The DET will have an excellent opportunity to steer development of both MET and METViewer for ensemble related verification needs. As part of that process, it will probably become incumbent on DET personnel to help with development of scoring algorithms and with programming tasks related to display of results. A quite complete set of probabilistic verification software has already been designed and installed in 'R' at RAL for DTC; an immediate question is how to best integrate these routines for use with MET. It is quite likely that between existing and planned DTC development, most of the direct needs of the DET can be satisfied.

In addition to previous use in the DTC, other large advantages of MET are the pool of expertise available locally, its close relationship to development of the WRF model, its highly conversant relationship to innovative techniques (e.g., object-based verification), and the growing use of MET at other agencies.

There are a variety of other tools for consideration. They are 1) the Network-Enabled Verification Service (NEVS) from the Aviation, Computing and Evaluation Branch at GSD; 2) NWS AWIPS and ALPS Capabilities development at GSD related to ensembles, it has emphasized interactive displays through the Advanced Linux Prototype System (ALPS); 3) the National Precipitation Verification Unit (NPVU), this web-based verification utility has a several-year history at Office of Hydrologic Development (OHD), but it is not designed for true probabilistic verification or display; 4) Verification for HRRR convective probability forecasts, in support of HRRR ensemble convective forecasts, display products have been developed to visualize the set of multiple lead times and initialization times for convective probabilities.

5) NCEP/EMC operational and developmental probabilistic forecast verification, the set of probabilistic verification utilities now installed in the real-time product stream at NCEP will be very valuable for DET verification activities. 6) Office of Hydrologic Development (OHD) has developed ensemble hydrologic forecasts, including verification. And, 7) the 'R' package is an open-source set of statistical utilities that serve both computational and display functions. These are evaluated in the more detailed verification plans.

Verification Datasets

Increasingly, the impact of verification dataset choices on verification results has become a topic of interest. In many practical cases, there are no choices to be made. To facilitate comparison with other centers, the DET will initially make available the principal verification data streams available at

NCEP including the RTMA, operational radiosondes, radar products, precipitation gages, and eventually satellite products. Where choices are available (e.g., non-operational rain gage networks), options to individually select verification data will be offered. Data quality evaluation for these verification sets will be primarily a user responsibility, but could eventually become a DET activity, if warranted.

Scoring Algorithms and Display

Eventually most of the probabilistic scores that have been considered valuable at other centers should become available as part of the DET verification module. Initially, however, a set of highest priority scores and display products will be developed with input from the potential user community. These will include, for instance, Brier skill scores, rank probability and cumulative rank probability scores, and decompositional products such as reliability and resolution. The first set of visual displays will include ROC curves, Talagrand diagrams, rank histograms, and reliability curves. Binned spread-skill, ROC skill score, and economic value diagrams will be added when available. Additional ensemble displays not specifically related to verification needs but useful nonetheless for qualitative evaluation are described in the display products module of this plan.

User Group Verification Packages

While numerical weather prediction from the EMC and other operational and research forecasting groups have been the principal active proponents for the DTC and the DET, there are several other distinct user groups for which the DET could have potential value. Four such users are the aviation, hurricane, severe weather, and hydrology communities. For each, verification needs could be significantly different from those of the other, and from those of the weather forecasting agencies. For instance, verification for hydrology will be particularly directed toward QPF and will often take the form of time series validation instead of gridded fields. Aviation users will likely need to provide verification for derived aviation-relevant variables (icing and turbulence; visibility) that may not be routinely produced by numerical weather models. The probability of track and intensity are of most interest to the hurricane community. Likewise, the probability of convective initiation leading to hail, wind, and tornadic outbreaks is important to the severe weather community. An additional consideration is data formats between user groups are not identical. To accommodate these users as much as possible, fact-finding visits and meetings will need to be arranged to identify the possible areas of interest and to specify sets of metrics that the DET could provide.

Timeline and Milestones

August 2010 – Verification module plan presented and reviewed by WRF ensemble working group
January 2011 – Initial probabilistic scoring utilities demonstrated as part of HMT winter exercise
March 2011 – Written module 6 plan for verification is completed as part of overall DET planning process
April 2011 – Prototype partial DET verification workflow and display utility that closely emulates the HMT parallel structure is assembled and applied to extended-CONUS test runs of HMT-based ensemble
June 2011 – Prototype real-time web-based verification site with basic capabilities completed
August 2011 – User group meetings in DC are arranged and held to integrate verification subsets for major potential DET users
September 2011 – Verification utility installed as full module in end-to-end DET workflow

Working Group Recommendations

(Mod6.1) *Do we need to be discussing the inclusion of MET (or MET-like capability) in AWIPS II?*
Based on this recommendation, we will consider NextGen requirements more closely.

(Mod6.2) *What is a prioritized list of ensemble-relevant verification products?*
DET will explore verification ideas recommended by the Working Groups including basic tools., time-space scales, decomposition tools, account for various needs of users, etc.

(Mod6.3) *How should we filter through ideas provided by other workshops (i.e., RAL Verification Workshop 11/2010; WMO Verification tutorials; others?)*
DET accept the recommendations to focus on sessions for ensemble verification at workshops and conduct "Literature review" by aware individual

(Mod6.4) *What analysis fields and/or observation data should we consider for verification?*
The Working Group recommended both, to use ensemble analyses (e.g. Torn and Hakim), disclaimer for uncertainty, etc. DET will explore the list of recommended ideas.

Testing and Evaluation Plans

Testing and evaluation is a core activity of the DTC, and will be a central effort of the DET. Through numerous model evaluation studies, the DTC has established criteria for the testing and evaluation process, to make results of the activity credible and meaningful. Similar criteria will be utilized for DET testing and evaluation activities. The DET infrastructure described in previous sections will facilitate the testing and evaluation activities and intercomparisons of the performance of ensemble systems and system components.

Several types of testing will be undertaken. The first type will focus on comparisons of the performance of different system configurations. For example, it may be of interest to compare the performance of an ensemble system with different model initialization methods, or with two or more approaches for statistical post-processing. Another type of testing and evaluation activity will concern a "reference configuration" (RC) of a modeling system, in which a particular formulation of a modeling system is re-tested, and performance is tracked, as model changes are implemented. RCs may be defined by operational centers, or may be contributed by members of the community; more information about RCs can be found at http://www.dtcenter.org/config/. Finally, the DET may undertake evaluations of ensemble forecasts that are contributed by an outside group, for example, through a special forecasting program. While DET testing and evaluation activities will typically be based on forecasts produced within the DET infrastructure, in some circumstances it will be beneficial to take advantage of the availability of forecasts from these types of special forecasting efforts, and will allow more testing than can be done solely with the DET's own resources.

DET testing and evaluation will follow a specific set of principles. In particular, for each testing and evaluation activity, the following criteria and steps will be taken into account:

- A formal test plan will be developed before the test is initiated, which will define all important aspects of the testing and evaluation study;
- The model or component developer may have a role in helping to create the test plan but will not control the process;
- Execution of the test will be independent of the developer;
- Focus of the test will be on questions and variables that are relevant for the specific application intended for the forecasting capability, and may be dependent on the module being considered;
- Many cases (not just a few case studies) will be evaluated to allow measurement of statistical significance, and the sample will include multiple seasons and times of day;

- Results of the evaluation will be stratified using meaningful criteria, such as location/region, season, and other user-based criteria to create homogeneous subsets; the stratification criteria may vary for different testing exercises;
- Statistical significance of comparisons will be evaluated and reported;
- The evaluation will measure and document the strengths and weaknesses of the modeling system;
- Complete results of the testing and evaluation exercise will be freely and easily available to the operational and research communities.

Development of the test plan is a critical first step in the testing and evaluation process. The test plan essentially represents an experimental design, which details all aspects of the evaluation. The plan specifies the codes to be evaluated and the output to be produced. In addition, the plan identifies the forecast periods to be included (e.g., dates, hours of the day, lead times) and completely specifies the type of post-processing to be applied to the model output, as well as the verification measures and results that will be examined. The plan also considers archival and dissemination of the data and results. Computer resource requirements and project deliverables are also specified. An example of a DTC test plan, which will be a model for future DET test plans, can be viewed at http://verif.rap.ucar.edu/eval/gfs_nam_pcp/.

Examples of questions that should be considered in formulating a test plan include the following:

- Which aspect(s) (or modules) of the ensemble system will be evaluated?
- What performance aspects are we trying to evaluate or compare?
- Who are the "users"?
- What are the variables of interest?

Answers to these questions will help determine many components of the plan, including the types of verification statistics that will be computed.

Many aspects of testing and evaluation by the DET will be the same as for other DTC model testing situations. However, some characteristics of ensemble systems will require special consideration when developing plans for testing and evaluation of ensemble systems. For example, the number of cases required to evaluate ensemble systems will typically be larger than the number required for non-ensemble evaluations. In particular, many probabilistic and ensemble verification scores (e.g., measures of reliability) require relatively large subsamples. In addition, the subsamples must be large enough to allow assessment of statistical significance, while still being representative of meaningful subsets of forecasts. In addition, the verification methods for ensembles are somewhat unique and (especially for some weather phenomena, such as tropical cyclones) less mature than standard verification approaches for non-ensemble forecasts. Thus, care must be taken in selecting and applying appropriate verification measures for ensemble forecast evaluations, and in interpreting the results of the studies. Finally, computer resource requirements will be very large, to provide the large number of cases needed to provide meaningful tests; computer resources may be a limiting factor for these activities and will need to be taken into account early when planning an ensemble evaluation.

Collaborations with Testbeds

NOAA Testbeds and topic-specific programs, like the Hurricane Forecast Improvement Project (HFIP), provide an excellent source of promising ensemble approaches and techniques for testing and evaluation by the Ensemble Testbed. Additionally, the DET should be able to help address issues such

as optimal ensemble configuration, suggested post-processing techniques (both statistical and general products), and best metrics for evaluating the forecasts. DET also plans to provide real-time mesoscale ensemble (the SREF benchmark from Module 2) for use as lateral boundary conditions into the finer scale ensembles used in the forecast exercises. Specifics on the interaction with three ensemble testbed collaborations are discussed below, in detail.

DET and the HWT

The focus of the Hazardous Weather Testbed (HWT) Experimental Forecast Program (EFP) is to evaluate the utility of research models in the operational forecasting arena. This evaluation is done during a 4-6 week period, called the Spring Experiment, during the peak of the severe weather (hail, wind and tornadoes) season. The past few years has focused on convection-allowing models and ensembles. The collaboration between the HWT and DTC has grown from DTC providing an off-line evaluation of one product and two models in 2008 to near real-time evaluation of six products and thirty models in 2010. The focus over the next few years will shift to less evaluation and more collaboration on the design and running of convection allowing ensembles.

Evaluation of Products The Center for Analysis and Prediction of Storms (CAPS) will provide a 40-50 member ensemble during the 2011 Spring Experiment. The goal of the HWT/DTC collaboration will be to evaluate all ensemble products for two or three forecast variables and determine which algorithms appear to have the most utility. DET will then acquire the most promising algorithms and use these to provide a foundation for Module 5.

Ensemble Configuration Once DET has the facility to generate products (Module 5) and evaluate the spread and skill of products (Module 6), a natural extension of the DET would be to explore the ultimate configuration of convection resolving ensembles. It is anticipated that in coming years (2012 and 2013), the HWT/DTC collaboration will move toward DET guiding HWT on ensemble configuration and possibly eventually producing an ensemble for the HWT Spring Experiment.

Statistical Post-Processing Another area for fruitful collaboration is statistical post-processing. The EMC. National Severe Storms Lab (NSSL) and CAPS all have very talented staff of individuals that are working on this problem. The role of DET would be to identify the most promising statistical post-processing algorithms for inclusion in Module 4 and more robust testing and evaluation.

DET and the HMT

Ensemble Configuration As a first step toward a prototype ensemble model setup, the DET will leverage existing development of a WRF-based ensemble that has been established over several seasons of the HMT winter exercises in California. While these configurations have changed over the years, the 2010-2011 experiment will again include WRF-ARW and WRF-NMM model ensemble members but with different initial fields, boundary conditions, and physics packages. As detailed in a previous section, initial prototype testing will in fact be performed using this HMT ensemble in near-identical configurations over a CONUS domain extended to capture regions of hurricane occurrence to the south and east of the US.

Verification During 2009-2010, a collaborative DTC-HMT USWRP-funded project has systematically developed a web-based real-time verification facility built around tools developed in the DTC (c.f. the website at http://verif.rap.ucar.edu/eval/hmt/2010/). These tools include both standard verification

scores and object-based algorithms that exist now as part of the MET and the MET/MODE packages. To date, the goals of this project have been primarily focused on QPF verification. During 2010-2011, one task within the HMT/DTC project will include development of additional probabilistic utilities for the MET, especially those applicable to ensemble spatial and object-based applications. Since many of these capabilities are central to the DET mission, strong interaction between the DET and the HMT is anticipated, and it is likely that the HMT/DCT verification workflow management scripts and website will serve as a prototype real-time system for initial DET verification activities.

DET and HFIP

Ensemble Configuration Another natural collaboration is with the Hurricane Forecast Improvement Project (HFIP). DET is currently working with HFIP on helping coordinate the regional scale ensemble modeling efforts to obtain the optimal ensemble configuration for the hurricane forecast intensity and track problem. Several members of the HFIP regional ensemble team introduced a collaboration plan in January. It is aimed at enhancing collaboration among interested groups via (1) development of web-based tools that allow participants to share their ensemble output data / images, with web services; (2) organization of periodic telecon meetings; (3) construction of a prototype ensemble system based on consensus of participants. This approach is similar to the NOAA THORPEX-funded collaborative ensemble-based data assimilation work (2004-2008) that led to the successful development of the global EnKF system at ESRL that is now used in HFIP and being transitioned to NCEP operations. The approach will involve the careful monitoring of ensemble performance with the goal of identifying the weak points in the system for attaining the stated goals (high quality probabilistic intensity and structure forecasts) and strengthened collaboration among the HFIP teams (ensemble, data assimilation, modeling, products, verification). It is expected that DET will play a major role in ensemble set-up and evaluation.