



DRAFT



**Plan for the  
Joint Ensemble Forecast System (JEFS)**

**1<sup>st</sup> Draft: 1 Jun 2004**

**2<sup>nd</sup> Draft: 15 Jan 2005 (*in coordination*)**

Maj F. Anthony Eckel, PhD  
Chief, Air and Space Models Branch  
HQ Air Force Weather Agency

## Table of Contents

|  |    |
|--|----|
| Introduction.....                      | 1  |
| JEFS Design Considerations.....        | 1  |
| System Design .....                    | 4  |
| Joint Global Ensemble (JGE).....       | 5  |
| Joint Mesoscale Ensemble (JME).....    | 8  |
| Calibration.....                       | 11 |
| Products.....                          | 13 |
| Phase I Products.....                  | 14 |
| Phase II Products.....                 | 19 |
| Project Tasks & Responsibilities ..... | 23 |
| Timeline .....                         | 23 |
| Communications and Data Storage.....   | 23 |
| Training and Education.....            | 25 |
| Verification and Validation (V&V)..... | 26 |
| Summary .....                          | 27 |
| References.....                        | 28 |

## Introduction

This Joint Ensemble Forecast System (JEFS) is a multi-year pilot project directed by the Fleet Numerical Meteorology and Oceanography Center (FNMOC) and the Air Force Weather Agency (AFWA) to **prove the value, utility, and operational feasibility of ensemble forecasting (EF) to enhance DoD operations**. Comprehensive use of EF promises substantial benefits but entails a challenging transition from deterministic to stochastic processes. JEFS will explore how EF can best be exploited to improve DoD forecast processes and warfighter decision making. The need for an EF capability has been identified in the following documents:

- 1) AFW Strategic Plan and Vision, FY2008-2032
- 2) Operational Requirements Document, USAF 003-94-I/II/III-D, Centralized Aerospace Weather Capability (CAWC ORD)
- 3) Air Force Weather, FY 06-30, Mission Area Plan (AFW MAP)

This plan describes the many facets of JEFS and breaks out development into two phases. *Phase I* will produce a basic, initial capability in a fairly short period of time. This takes advantage of the available computer hardware associated with the Weather Research and Forecasting (WRF) model Dedicated Distributed Center (DDC) project, a joint venture of the FNMOC and AFWA supported by the High Performance Computing Modernization Program. Phase I will produce one of the DDC project deliverables--the capability to generate short-range, mesoscale EF products. *Phase II* will produce a more robust ensemble prediction system and involve more extensive analysis and application of the EF data and products to DoD weather forecasting and decision making.

### JEFS Design Considerations

EF is a revolutionary method in operational meteorology that enables a radically different, stochastic approach to weather forecasting. Instead of the rather limited (and often misleading) traditional use of a single model run in which the error (or uncertainty) is mostly unknown, ensemble forecasting uses multiple model runs (called ensemble members) to incorporate uncertainty into the modeling process and reveal a spectrum of forecast possibilities. The general goal in EF is to produce a probability density function (PDF) for the future state of the atmosphere that is reliable (consistently encompass the truth) and sharp (small degree of

variance). The JEFs project does not seek to eliminate deterministic forecasting since there will always be such a need as well as applications where a single forecast is more appropriate. The key point is that whenever the uncertainty of the wx phenomena exceeds operational sensitivity, either a reliable probabilistic or a range-of-variability prediction is required. In that case, which occurs for most weather forecasts, providing a deterministic forecast unnecessarily cripples the warfighter.

There are several key issues that must be considered in designing an ensemble system:

**Issue #1:** *Incorporation of the two different aspects of uncertainty, analysis uncertainty and model uncertainty.* This is the heart of what EF is all about. Incorporating analysis uncertainty (i.e., possible errors in the model initial condition, IC) is generally the primary concern since the non-linear complexities of the atmosphere cause errors from the analysis to grow to dominate the error in the forecast. Analysis errors are captured in an EF by using a perturbed (slightly different) but equally likely IC for each ensemble member. Incorporating model uncertainty (i.e., possible errors and deficiencies in the model's representation of the atmosphere) is also important, especially for surface variables in mesoscale models where parameterizations greatly contribute to forecast error. Model errors may be captured in an EF by either adding tiny perturbations to the members' solutions during forecast integration, called stochastic physics, or by using a different model (or different version of the same model) for each member. Both analysis and model uncertainty can therefore be simultaneously incorporated into an ensemble system by having each member start from a unique IC that is applied to a unique model. In a well designed EF system, the various resulting forecast solutions are all valid possibilities to be considered all together (i.e., as an ensemble).

**Issue #2:** *Ensemble size, resolution and forecast length vs. processing power.* To sample all possible forecasts for a given model and IC distribution, an infinite number of ensemble members would be required. Additionally, it is necessary to use very fine resolution to generate ensemble information on the small-scale features of interest to the forecaster or customer (e.g., probability of surface winds). Obviously, those needs must be balanced against available processing power in the context of an operational system (i.e., timeliness of output and products). A worthwhile approximation of forecast uncertainty can be accomplished with ~10 members. For highly skilled probability forecasts, at least ~25 members are necessary. To consistently

capture low probability events (which may or may not be extreme events), on the order of 100 members is needed.

**Issue #3:** *Calibration of ensemble data to make up for the approximations and compromises in issues #1 and #2.* It is very difficult to thoroughly incorporate all sources of uncertainty into an ensemble, and limited ensemble size creates under sampling problems. In an ideal EF, all members are equally likely, but that is normally not realized in EF due to model biases, other systematic errors, and the challenges of designing proper IC perturbations. A postprocessing calibration can correct for much of the resulting degradation in the ensemble data and is therefore critical for maximizing EF utility.

**Issue #4:** *Generation of products that effectively convey the EF data.* The final challenging step in an EF system is boiling down the potentially overwhelming amount of information from raw EF output into usable products. Expecting a forecaster to simultaneously use standard prognostic plots from many model runs, or to attempt to pick out the “model of the day” upon which to base a traditional forecast is not how ensembles are meant to be used. To properly apply ensemble data to forecasting and decision making, a stochastic mindset is required in which all the data are considered at once and forecast uncertainty is conveyed to the customer. EF products must be designed to enable that process.

**Issue #5:** *Education and Training.* To make best use of EF output and products, forecasters and customers need to have some understanding of the basic theory (i.e., the reason and need for EF) and methodology (i.e., assumptions, limitations, etc.) and must be trained on proper application. This issue is not directly related to the designing an EF system, as the issues above, but is brought up since it is so critical for the ultimate successful use of EF within the DoD. Transitioning to stochastic weather forecasting is a significant challenge since deterministic forecasting is strongly ingrained in DoD weather operations from end to end. Part of the JEFs project will therefore be to examine how to educate and train forecasters/customers and make recommendations on how to proceed with the necessary transition.

## System Design

Following the general predictability of the atmosphere, the JEFs will consist of two separate ensembles, the Joint Global Ensemble (JGE) for medium-range, large-scale stochastic forecasting and the Joint Mesoscale Ensemble (JME) for short-range, small-scale stochastic forecasting (Figure 1). In general, for the short range (0-72 h), skillful forecasting is possible on the mesoscale (~10 km) and above. Once into the in the medium range (3-14 days), skillful forecasting is only possible on the synoptic scale (> ~100 km) and above. The JGE will be an ensemble of lower resolution, global model runs, ideal for providing meteorological guidance for planning purposes. The JME will be an ensemble of limited area model runs that focus on the mesoscale weather phenomena of interest to operations.

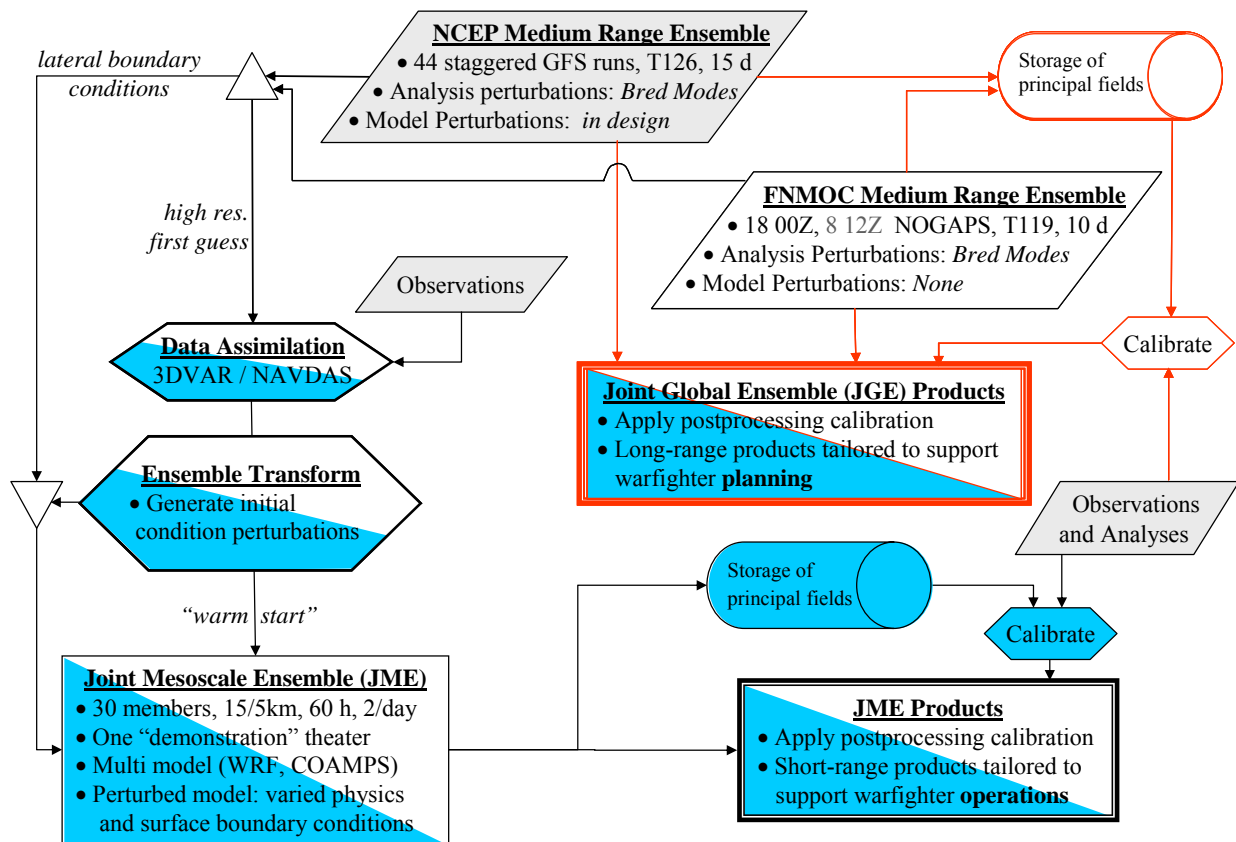


Figure 1. Flow chart of the objective configuration of JEFs' components. Lightly shaded icons represent external data, unshaded icons depict FNMOC process, and darkly shaded icons depict AFWA process. Icons with split shading are processes that occur at both FNMOC and AFWA

## Joint Global Ensemble (JGE)

The JGE will be formed by combining the Global Forecast System (GFS) medium-range ensemble from the National Centers for Environmental Prediction (NCEP) and the Navy Operational Global Atmospheric Prediction System (NOGAPS) medium-range ensemble from FNMOC. Since both the GFS and NOGAPS ensembles will likely be upgraded in 2005, it is difficult to describe the final JGE configuration. Basically, JGE will consist of 2 daily cycles in which all available members are combined together. For the current ensembles, this would be 39 ensemble members at the 00Z cycles and 31 members at 12Z. (Note: the odd split of ensemble size is explained below.) While combining different ensembles into a larger ensemble may seem unscientific at best, it has clearly been shown that this procedure produces superior results to either original system because of greater diversity and sampling. Since neither the GFS nor the NOGAPS ensembles currently incorporate model uncertainty, JGE will be somewhat deficient in that requirement except for the fact that JGE will contain solutions from two different models.

The GFS ensemble currently consists of 45 members per day that are initialized and run over four cycles per day with various resolutions as shown in Figure 2. JGE will use 44 of these

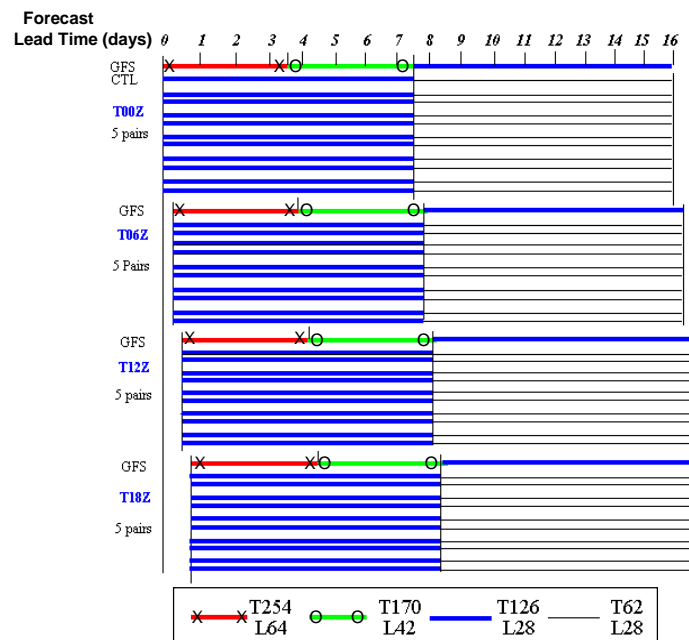


Figure 2. The four daily cycles of the GFS ensemble. One line represents a unique model run. In the key, T stands for triangular spectral truncation wave number and L for number of levels. The 5 pairs at each cycle are the bred mode ensemble members in which one pair consists of a positive and a negative perturbation (generated by the breeding method) to the high resolution control forecast.

members and omit the 00Z control (CTL) member, which is a lower resolution repeat (i.e., same IC) of the high resolution control forecast. The 18Z and 00Z members will contribute 22 members to the 00Z JGE cycle, and the 06Z and 12Z members will contribute 22 members to the 12Z JGE cycle. This 6-h time-lagged approach for the GFS members is acceptable for medium-range forecasting in which forecast skill normally diminishes only slightly for forecasts initialized within 6 h of each other. Each of the 4 GFS ensemble cycles consists of one high resolution control forecast and 10 lower resolution forecasts from five pairs of initial conditions created with the breeding technique (Toth and Kalnay, 1997). Upgrades to the GFS ensemble planned for the spring 2005 include increased resolution of each member, increase in number of members per cycle to 20, introduction of the Ensemble Transform (ET) technique for generating ICs, and possibly model perturbations.

The NOGAPS ensemble currently consists of 26 members per day that are initialized and run over two cycles per day. At 00Z, 17 members are run (using resources at both FNMOC and NAVO MSRC) including the full resolution control (T239/L30), and 16 reduced resolution (T119/L24) forecasts with initial conditions perturbed using the breeding technique. At 12Z, fewer resources are available so there are only 9 members including the full resolution control and 8 perturbed runs. Upgrades to the NOGAPS ensemble planned for 2005 include adding more members and introduction of the ET technique for generating ICs.

To generate gridded ensemble products, GFS and NOGAPS forecasts will be transformed (using a simple bilinear transformation) to a common grid, chosen to approximate the dominant resolution present among JGE members for maximum efficiency in data handling and storage. Of the 6 resolutions present, T126 (~110 km at 45°N) represents the biggest fraction followed by T119 (~120 km), T62 (~230 km), T254 (~55 km), and T239 (~60 km). JGE will use a 1.0°×1.0° (~80 km × 111 km) grid in anticipation of the planned GFS and NOGAPS ensemble upgrades. The 10-d forecast of the NOGAPS ensemble will limit the JGE forecasts to 10 d so the additional GFS forecast lead times will not be used. Each 00Z JGE cycle will consist of the forecasts shown in Table 1. Figure 3 shows that JGE products will be ready for dissemination by 8 h after initialization time.

Table 1. Break out of JGE members.

| JGE Cycle | Model  | # of Members | Cycle Init. Time | Forecast Lead Times (h) |
|-----------|--------|--------------|------------------|-------------------------|
| 00Z       | GFS    | 11           | 18Z              | 06, 12, ... 246         |
|           | GFS    | 11           | 00Z              | 00, 06, ... 240         |
|           | NOGAPS | 17           | 00Z              | 00, 06, ... 240         |
|           |        |              |                  |                         |
| 12Z       | GFS    | 11           | 06Z              | 06, 12, ... 246         |
|           | GFS    | 11           | 12Z              | 00, 12, ... 240         |
|           | NOGAPS | 9            | 00Z              | 00, 06, ... 240         |

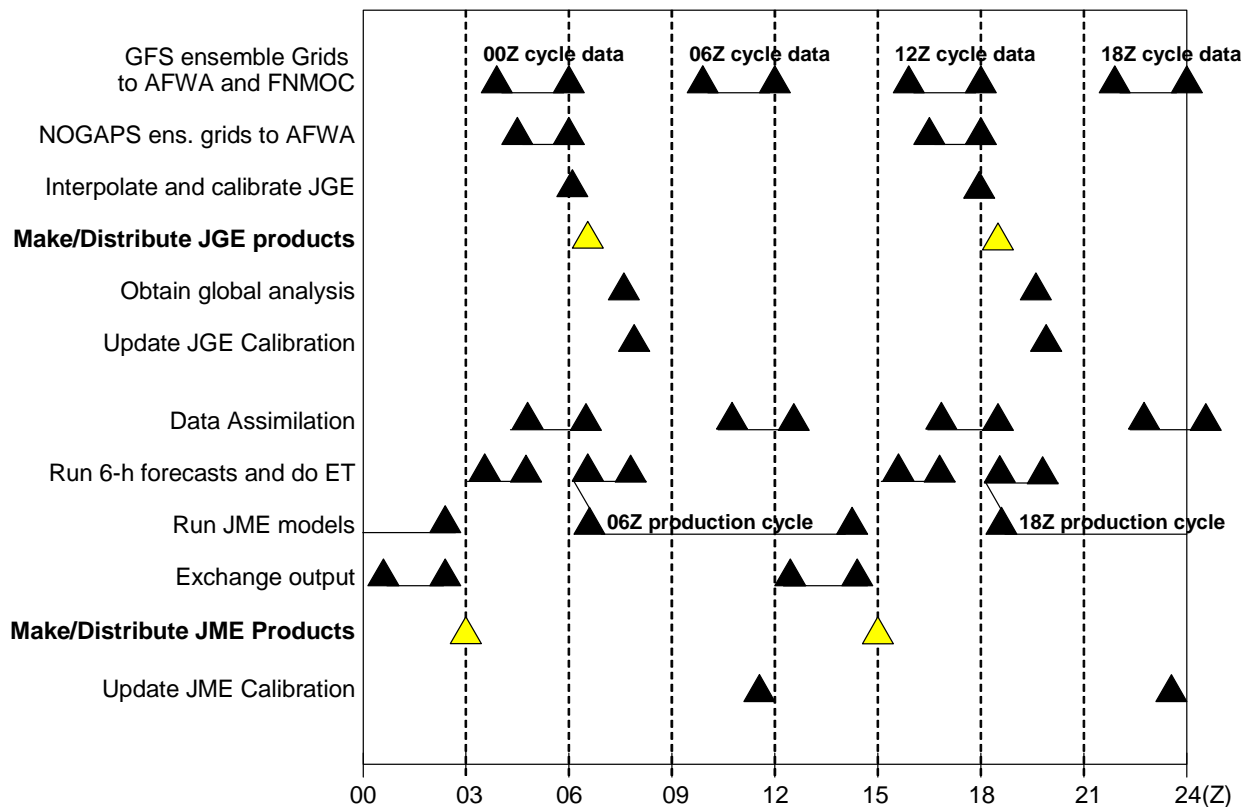


Figure 3. JEFs daily production schedule.

## **Joint Mesoscale Ensemble (JME)**

The JME will consist of two daily cycles with an objective (threshold) of 30 (20) individual mesoscale model runs at each cycle. (Note: 30 members is a target based on preliminary capability estimates of the DDC hardware--IBM 1600 p655, 192 processors.) The Ensemble Transform (ET) technique (Bishop and Toth 1999) will be used to generate ICs and lateral boundary conditions (LBCs) for each member. To maximize efficient use of the DDC hardware, 15 members will run on the FNMOC hardware and 15 on AFWA hardware. Output of the principal fields (Table 2) will be exchanged to allow each center to generate JME-based products tailored to their customers' missions.

A key question for JME concerns use of multiple models. The objective is to use more than one mesoscale model (i.e., WRF and Coupled Ocean Atmosphere Mesoscale Prediction System, COAMPS) to help generate ensemble dispersion that represents model error. However, as described below, this would complicate pre- and post-processing due to the incompatibility between COAMPS and WRF output. For both ET calculations and for product generation, the data need to match (i.e., format, grid, levels, and variables). If it is determined that matching the data involves a considerable effort, only WRF will be used for Phase I (i.e., FNMOC will run 15 WRF members in lieu of COAMPS) and the expansion to a multimodel JME will be included in Phase II. The following material assumes multimodel in Phase I.

Table 2. Principal model output fields used in JEPS product generation and archiving for use in calibration and verification. Parameters noted with an asterisks (\*) may not be available in the JGE data. With model output frequency of 12 h and 3 h for JGE and JME respectively, maximum and minimum (max/min) data of highly variable surface parameters is desirable for probabilistic forecasting of mission-critical weather thresholds (e.g., surface winds > 50 kt).

| <u>Surface (or near sfc) 2D Data Fields</u>    | <u>Pressure Level, 3D Data Fields<br/>(at 1000, 850, 700, 500, and 300mb)</u> |
|--|---|
| mean sea level pressure                        | geopotential height   |
| 2-m temperature, and *max/min                  | temperature   |
| 2-m dew point                                  | dew point   |
| 10-m wind components ( <i>u</i> and <i>v</i> ) | wind components ( <i>u</i> , <i>v</i> , and <i>w</i> )                        |
| *maximum wind speed and direction              |   |
| cumulative precipitation                       |   |
| *ceiling(?), and max/min                       |   |
| *visibility(?), and max/min                    |   |
| *severe wx parameters(?)                       |   |

<Craig: This is the section that needs your input. Please edit, expand, and clarify however you see fit.>

The ET technique uses differences among short-range forecasts to estimate analysis uncertainty and generate IC perturbations. Figure 4 shows the basic preprocessing cycles for JME. The very first perturbations can be random but after a few cycles, the process becomes self sustaining when run continually. FNMOC and AFWA will run independent, 6-h data assimilation cycles but mutually dependant ET cycles. To simplify development, data assimilation will mirror the operational design at each center: FNMOC will use the high resolution NOGAPS with the Navy Atmospheric Variational Data Assimilation System (NAVDAS) on COAMPS while AFWA will use the high resolution GFS with the three-dimensional variational (3DVar) data assimilation system on WRF. Perturbations from the previous ET calculation will then be applied to the two analyses (i.e., 15 WRF perturbations for the 3DVar analysis and 15 COAMPS perturbations for the NAVDAS analysis). The resulting ICs will then be run through the appropriate model to produce 6-h forecasts for the next ET calculations. The ET calculation will be performed on the entire set of 30 forecasts rather than separately on each 15-member subset. This necessitates an exchange of 6-h forecasts between the centers every six hours. (Note: This data exchange is separate from the exchange of the full 60-h forecasts for product generation every twelve hours.)

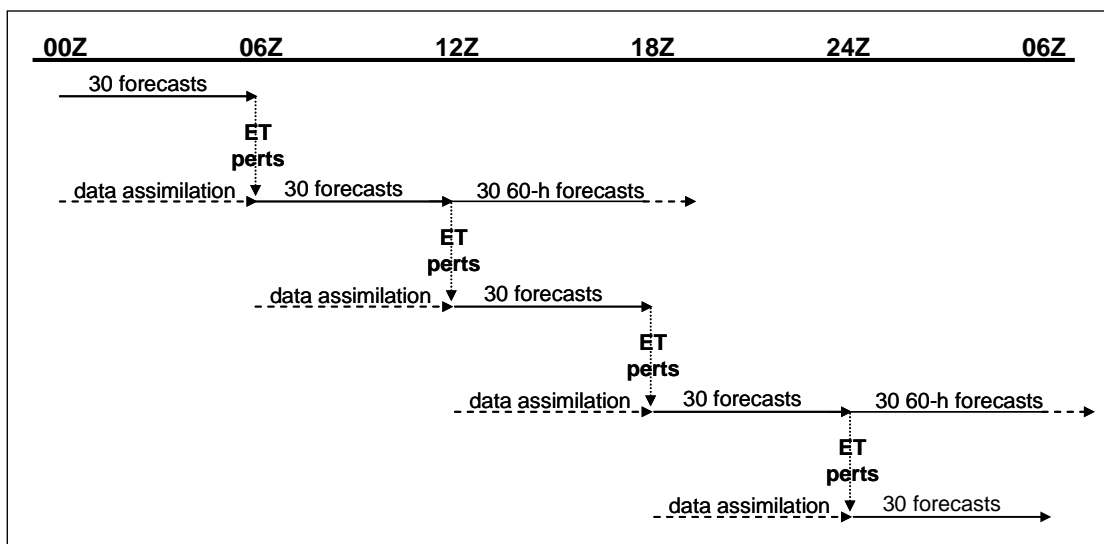


Figure 4. JME preprocessing cycle

With the current NOGAPS and GFS ensembles use of bred mode ICs, it will be non-optimal and difficult to apply ET for JME. The LBCs provided from the global ensembles would have to be merged with the ET ICs early in the forecast period, likely resulting in inconsistencies and poor dispersion. If/when, both the NOGAPS and GFS ensembles switch to the ET for generating ICs (planned for 2005), running ET for JME will become much simpler and robust. In Phase II, the IC generation may be further improved by employing the Ensemble Transform Kalman Filter (ETKF, Wang and Bishop 2003).

The objective (threshold) JME forecast period is 60 h (48 h) with an output frequency of 3 h (6 h). The two daily cycles will be initialized at 06Z (using the 18Z and 00Z GFS, and 00Z NOGAPS) and 18Z (using the 06Z and 12Z GFS, and 12Z NOGAPS). This choice is based on timing the availability of JME products to coincide with 00Z and 12Z cycle standard deterministic products from AFWA. Figure 3 shows that JME products will be disseminated ~9 h after cycle initialization time, which may seem like less desirable information compared to the more current, ~3 h old deterministic forecast products. However, it can be shown that less timely ensemble products still contain more valuable information than newer deterministic products.

JME will use a 15-km model domain over East Asia (Figure 5) in Phase I. The objective in Phase II, dependant upon additional hardware acquisition, is to add a 5-km inner nest over the Korean Peninsula to capture the uncertainty in smaller-scale motions. East Asia was chosen by the Committee for Operational Processing Centers (COPC) since it is a region of high geopolitical interest that contains challenging weather and a wide assortment of DoD assets. It is an excellent region to prove the value of ensembles to DoD operations.

The JME will account for model uncertainty by using multiple models (WRF and COAMPS), various model versions (i.e., different combinations of physics packages), and perturbations to surface boundary parameters (sea surface temperature, soil moisture, soil temperature, roughness length, and albedo). The 15 members run at FNMOC will use various versions of COAMPS while the 15 members run at AFWA will use various versions of WRF (Advanced Research WRF, ARW, with Eulerian Mass core and possibly NCEP's WRF Non-hydrostatic Mesoscale Model, NMM, core as well). The model versions and perturbations will be held static during Phase I to allow for effective calibration and system evaluation.

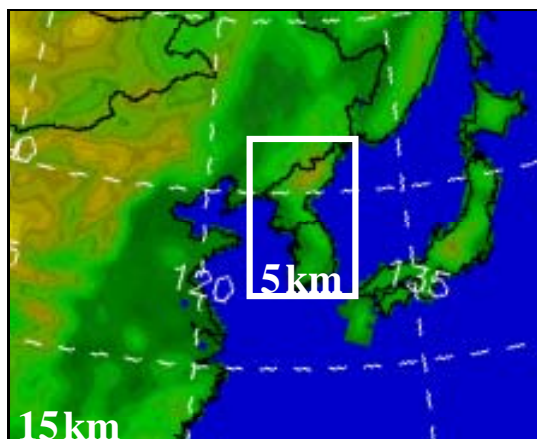


Figure 5. Approximate nested domains of the JME: 15-km over East Asia for Phase I and an additional 5-km inner nest over the Korean Peninsula for Phase II.

### **Calibration**

The goal of calibration is to improve skill by correcting for systematic (i.e., recurring) errors. The two basic components of skill that may be improved for probability forecasting are *resolution* (the ability to distinguish events from nonevents) and *reliability* (the ability of the forecast to match the observed relative frequency of an event over many cases). As mentioned above, proper calibration is a critical aspect of any EF system given the limitations of incomplete representation of uncertainty, limited EF size, model bias, etc. While calibration cannot completely make up for these deficiencies, it can greatly enhance and thus maximize the utility of EF products.

For Phase I, JEFs will employ the method of weighted ranks (Eckel and Walters 1998, Hamill and Colluci 1997) will be used because of its adaptability to different parameters and straight forward application. Basically, the method uses the information from verification rank histograms (VRHs) to adjust the EF. A VRH is a history of where the observation occurred among rank ordered ensemble members for a given parameter. It can be produced using all past forecast/observation data for a particular parameter, or broken up spatially and temporarily for greater specification (with respect to having a large enough sample per VRH to be meaningful). This method accounts for gross model bias but not specific biases of each member. When using multiple or varied models, the diversity created by varied biases should be removed since it does not represent uncertainty. Correcting each member's bias separately reduces unrealistic

ensemble spread, thus drawing the ensemble back together towards truth and improving resolution.

For Phase II, a more advanced postprocessing calibration algorithm will be employed in which an estimation of the forecast PDF is performed. This involves a separate bias correction on each individual member followed by construction of the PDF. Bias correction may be done using a recent (~2-4 weeks) training period of gridded analysis or observational data spread over the grid, to calculate and subtract out the mean error at every grid point and every lead time. Estimating the PDF may be done with direct estimation of PDF moments (when PDF type is known and ensemble is large), a kernel dressing technique, or Bayesian model averaging (which is particularly well suited for a multimodel ensemble).

The choice of what to call truth greatly influences the quality of a calibration. Training an ensemble toward a truth that contains significant error may result in a degraded product rather than an improvement. Therefore, only the highest quality observations/analyses should be used. Additionally, systematic error can be highly spatially and temporally dependent, so a truth is needed that covers the domain with frequent updates. To capture the spatial variability of systematic error, calibration will be broken up by location or by weather regime (areas with similar meteorological patterns/conditions).

An independent, accurate, model-based analysis at the same (or better) resolution than the forecast is ideal for calibration (as well as for verification). For JGE, the UKMO global model analysis (fit to the JGE grid) will be used as truth for calibration. For JME, there may be no independent mesoscale analysis over the domain. The availability of analysis data from the Korean Meteorological Agency (KMA) and the Japanese Meteorological Agency (JMA) needs to be investigated. (Note: Use of a control model analysis from the JME data assimilation is not an option since it would contain the same systematic errors.) These data do not need to be retrieved in realtime, which may make them easier to obtain.

Standard observation data can also be used to calibrate. Observations are very useful for point forecasting calibration and can potentially be used for calibration over the entire model domain. If enough representative observations are available, site-specific calibration results could be applied across the domain to grid points with a similar weather regime that presumably share similar systematic error. Successful research into this method is ongoing at the University

of Washington. JME Phase II will apply this observation-based bias correction if a mesoscale analysis is not available.

Choosing the training period (i.e., number of past forecast cycles to include) is another sensitive aspect of calibration since systematic error is dependent upon various time scales (seasonal, synoptic, and diurnal). Since years of data will not be available, a short-term, running calibration will be used in which the training period is the most recent  $x$  number of forecasts/observations. This training period must be long enough to generate a meaningful calibration (i.e., from a large dataset) but not so long as to be degraded by seasonal variations. (E.g., using a 6-month training period would make the summer model bias influence the correction in early winter, which is not appropriate in most cases.) The ideal training period length may vary by parameter, but is likely between 10-30 days. To capture the diurnal fluctuations, calibration will be broken up by forecast valid time relative to the local time of day. One flaw with a running mean calibration is that sudden shifts in the weather pattern, typically in the transition seasons, are not handled well. A possible solution under development at NCEP is to correlate the current forecast cycle's conditions with the recent past cases and train with only those cases with high correlation.

Consistent and extensive data archiving is required for calibration. To maximize use of available storage and avoid duplication, the archiving responsibilities (and therefore the calibration calculation as well) will be divided as shown in Figure 1; FNMOC will be responsible for long term storage of JGE (and UKMO) data and calculation of JGE calibration, while AFWA will do the same for JME. Since calibration results will be updated daily and will be required for generating products at each center, calibration results will be exchanged daily between the centers. Archival will be done only for the principal fields (Table 2) and will be continual once data flow begins. With the JGE and JME held static, a larger dataset (at least 1 full year) can be accumulated for a seasonal-based calibration instead of a running one.

## **Products**

Coordinating product development with end users is an important part of JEPS for both maximizing the positive impact of EF and for exploring training and education issues. Because of their mission responsibilities within the JME model domain, the primary product test beds identified for JEPS are the 20 OWS, Yokota AFB Japan, the 607<sup>th</sup> WS in Korea, and the Naval

Pacific Meteorological and Oceanographic Center, Yokosuka NB Japan. The level and details of participation in JEFS by these units is TBD. The idea is to work closely with the forecasters on designing and evaluating products that help them do their job better. Additionally, leadership at these centers may aid in the effort of incorporating stochastic data into the warfighters decision making processes.

All JEFS data will be available at both FNMOC and AFWA to allow each center to generate and disseminate products tailored to meet their mission. To give warfighters superior information for decision making, products will be designed to fully exploit the stochastic nature of EF data while focusing on operational weather sensitivities. The JGE products will be mainly for long-range planning and the JME products will be for short-range, mesoscale concerns for operations. Products will be disseminated via JAAWIN.

### **Phase I Products**

In Phase I, prototype examples of the four basic types of EF products described below (*model confidence, consensus, data range, and probability*) will be generated from both JGE and JME. The focus will be on probability products since they have the greatest potential for improving decision making. It is unfortunate that for most situations in today's operations environment, a probability forecast would not be accepted. Forecasters have been trained to provide a deterministic, yes/no or single-value forecast. Operators make decisions using deterministic forecasts that they know are imperfect. This style of weather support can harm the decision making process since uncertainty is ignored or left up to the customer to estimate.

**Model confidence products** provide the first-order information from EF, forecast uncertainty. In general, when the ensemble members are in closer agreement (i.e., lower spread among members), there is higher confidence, or less uncertainty, since less error can be expected in any of the ensemble of solutions. When there is high spread, there is low confidence and a greater chance of large error. Figure 6 is an example model confidence product that shows how ensemble spread often varies dramatically over a model domain. A forecaster concerned about weather for Western British Columbia could use the model guidance with confidence since the area around the storm moving onshore has relatively low ensemble spread. In contrast, forecasting for the Aleutians would be very difficult since the ensemble spread is so high that the forecaster could not trust the model. JEFS will generate similar products to aid forecasters along these same lines.

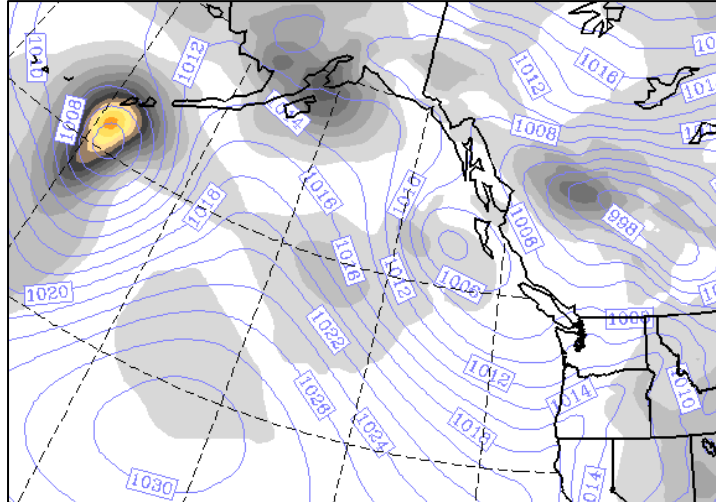


Figure 6. Ensemble spread (shaded) and mslp of the ensemble mean from the University of WA Mesoscale Ensemble, valid 12Z, 6 Jun 2004. This is a combined model confidence and consensus product.

**Consensus products** are regular weather forecast charts (e.g., 500 mb height and vorticity) or other model output (e.g., meteograms) that convey data from a single, best-guess forecast from the ensemble. Typically the ensemble mean (as in Figure 6), simple average of all the members, is used as the consensus forecast since over many cases, it has the lowest average error compared to all EF members. However, on any given day the consensus forecast is not necessarily the best forecast among all the members. One or more of the EF members may frequently beat the consensus forecast. The consensus forecast should be thought of as the safest bet in the long run and has been shown to out perform the control deterministic forecast in the long run.

Use of consensus products is a very limited way to employ EF since it is a deterministic application of stochastic data. However, consensus products will be included in the JEFS product line to support the traditional forecast process in which standard weather charts are used to analyze the forecast situation. In a deterministic or stochastic approach to forecasting, understanding the flow of the day is of course essential. Instead of using the ensemble mean that can smooth out key features, JEFS will use the ensemble median (i.e., the member closest to the mean) as the consensus forecast. This will preserve the average error minimization realized by the ensemble mean and allow atmospheric structures to remain intact in the consensus products. How to determine the ensemble median is TBD since the answer may vary depending upon such

factors as which parameter(s) is used, over which area it is calculated, and at what lead time it is calculated. One question to consider is whether to maintain the same member as the ensemble median throughout the forecast period for the sake of consistency or allow it to switch from one member to another where appropriate?

**Data range products** give a span of likely possibilities for a particular variable of interest, which gives information that begins to tap the full power of EF. Currently, a popular tool for point forecasting is the meteogram in which a single forecast for variables such as mslp, surface winds, surface temperature, and cumulative precipitation, is plotted over the forecast period for a location of interest. The data range version of today’s meteogram would be a “multimeteogram” in which the trace from all the ensemble members is displayed for any variable. Figure 7 is an example for just a single variable showing the range of possible forecast mslp at Offutt AFB, NE. Notice that this product also reveals forecast uncertainty as the width of the data range, which increases rapidly for 19-20 June indicating low model confidence, followed by a notable increase in confidence.

For an ensemble with a large number of members, a plot like Figure 7 becomes cluttered. Figure 8 provides the same type of information by showing the extremes of the range, a 90%

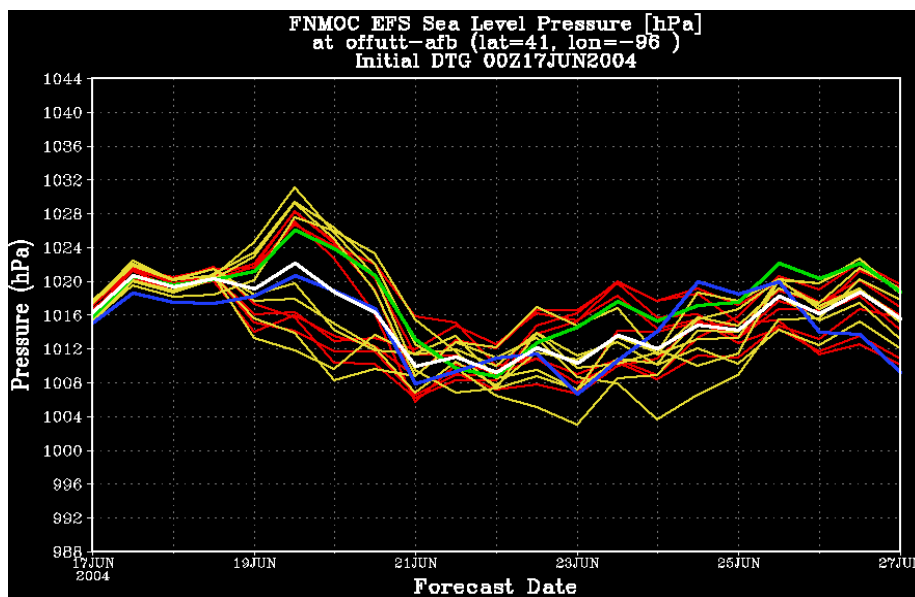


Figure 7. Five-day forecast mslp at Offutt AFB, NE, showing a range of possibilities in which each line is a unique forecast from the NOGAPS ensemble.

confidence interval (the range in which the verification will occur 90% of the time), and the mean. Consider how the information in Figure 8 would influence a forecaster’s decision on issuing a warning for winds  $\geq 35$  kt for the period 12/06Z – 12/21Z. A single model forecast over the period may only show a max wind of 30 kt, but with no idea on how much error is likely, the forecast is unaware of the significant potential to observe winds  $> 35$ kt. The data range information in Figure 8 clearly shows that a warning is warranted.

**Probability products** have the potential to unleash the full power of EF and enable optimal decision making for the warfighter. Ironically however, such a product is the most difficult to use because of the psychology of the human mind. Since only one reality is ever observed, the mind is patterned to think deterministically. It is quite challenging to make a decision when presented with multiple possible outcomes. Decisions based on a probability forecast are often made rather arbitrarily with little understanding of what the forecast actually means. For example, given a 20% chance of rain, a family goes ahead with its picnic plans since to them that basically means no rain and definitely doesn’t merit canceling their plans. They are then shocked when they get soaked and blame the weather forecaster for such a horrible forecast. They fail to realize that there was nothing wrong with the forecast and that in 1 out of 5 cases like this, they should expect such a soaking.

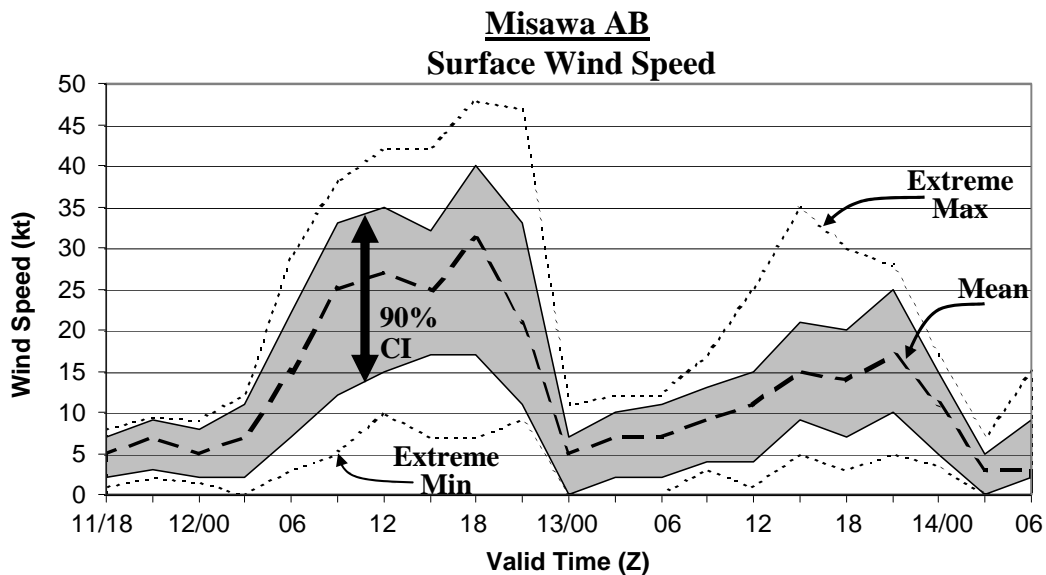


Figure 8. Hypothetical 60-h forecast for surface winds at Misawa AB. The range of possible wind speed is given with a 90% confidence interval (shaded) and the extreme maximum and minimum (dotted).

When used properly, a skilled probability forecast has tremendous value for a customer because it concisely conveys forecast uncertainty focused on the customer sensitivities to weather. Simplistically, uncalibrated probability can be produced from an ensemble by calculating the fraction of members that exceed a customer's critical weather threshold. For example, a critical threshold for the B-2 is a 30 kt runway crosswind, beyond which it cannot land or takeoff. If 23 of the 30 JME members forecast crosswinds  $> 30$  kt, the probability of crosswinds out of limits would be 77%. (Note: A more sophisticated statistical routine to calculate calibrated forecast probability will be used in JEFS.) This process can be performed for any weather parameter and any threshold, and can be plotted over the model domain, as in Figure 9. For point forecasting, Figure 10 displays the concept of a "probagram" (probabilistic meteogram) in which probability for exceeding weather warning criteria is plotted over the forecast period. This invaluable forecast tool will be generated for test locations in Phase I.

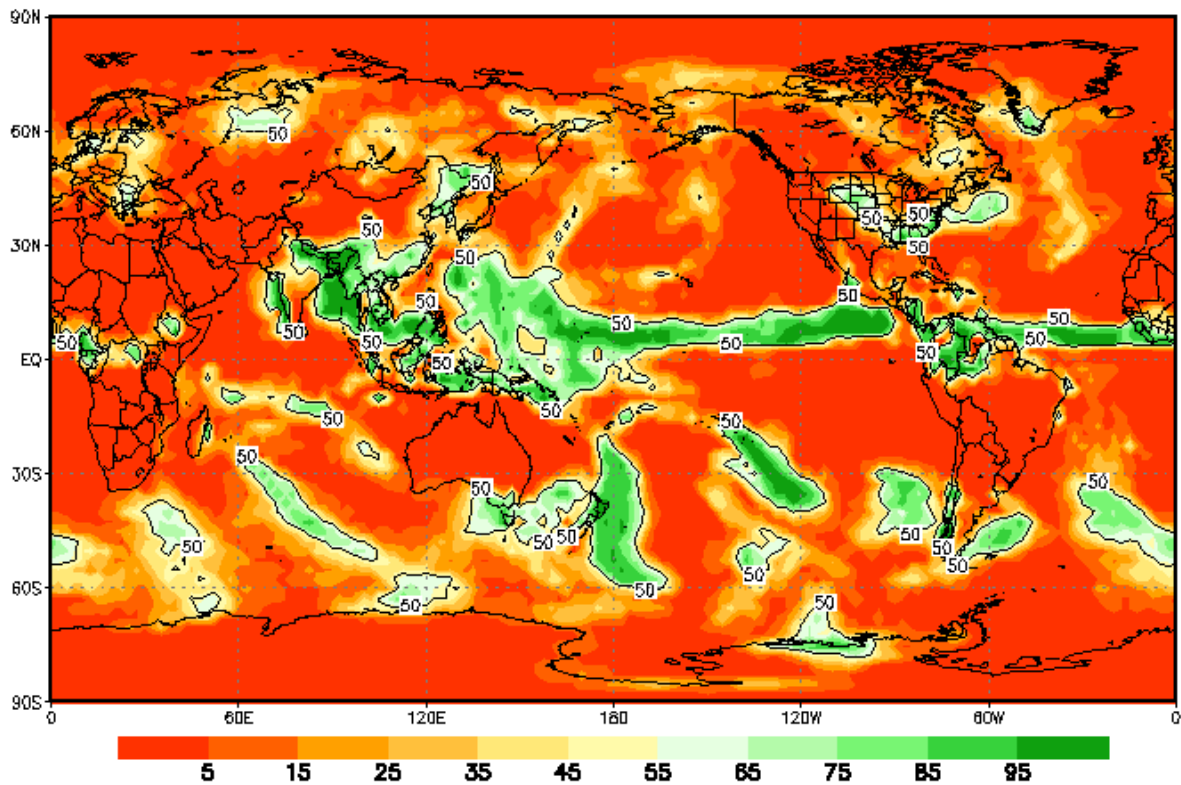


Figure 9. Probability of 24 h cumulative precipitation  $> 5$  mm from NCEP GFS ensemble, valid 12Z, 19 June 2004.

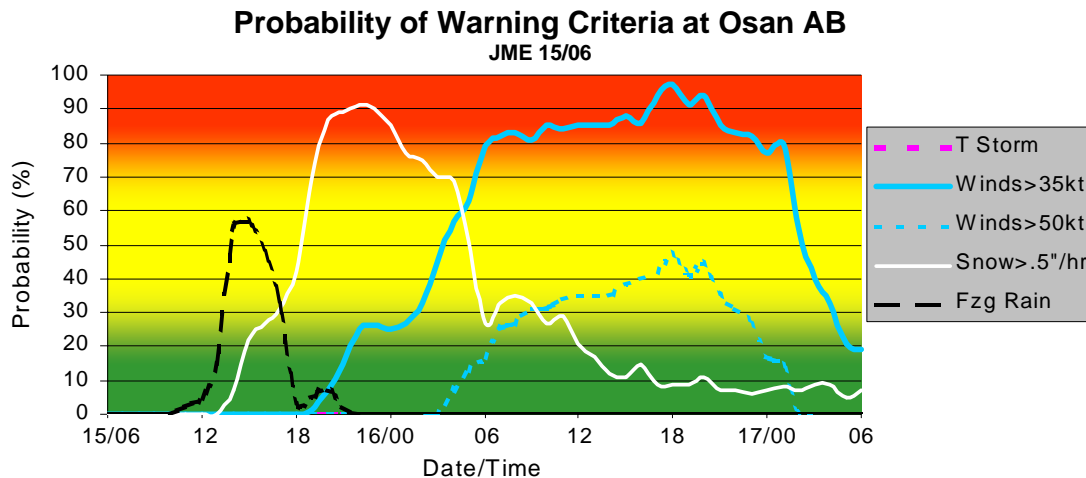


Figure 10. Probogram for forecast probability of weather warning criteria at a specific location.

### **Phase II Products**

In Phase II, there will be two separate vectors for expanding the utility of JEFs output. First, an interactive product generator will be designed. Rather than created an extensive set of fixed products that attempt to cover everything of concern to a forecaster, a more flexible interactive product generator would allow a forecaster to rapidly generate products tailored to support a specific mission or focus in on problem weather. This would be similar to the IGRADS system in use today on JAAWIN, but provide choices such as parameter and threshold of interest (e.g., low level wind shear > 20 kt) for generation probability and other products.

The second product vector of Phase II is exploring ways to inject stochastic weather information into the warfighters' decision processes. This may include anything from adding confidence intervals to weather data on a 175-1 weather briefing to fusing stochastic data directly into warfighter support tools. The latter would entail adaptation of existing weather-dependent, warfighter support algorithms (that are currently designed to ingest and produce deterministic data) to use EF data and produce stochastic output. For example, the utility of the Dust Transport Application (DTA) output is fundamentally limited by its designed reliance on a single atmospheric model solution. The DTA could be redesigned to ingest ensemble output data then generate probabilistic dust plume plots based on critical dust concentration or visibility thresholds.

Just getting stochastic forecasts to the warfighter is only half the battle since acceptance and proper use of the information are significant challenges that must be met in order to fully exploit

the data. The key aspect of the necessary transition is learning how to base a binary (deterministic) decision on stochastic information. One means to do that and is from the point of view of economics, using *decision theory*, or *cost-loss analysis* (Zhu et al. 2002). The goal of decision theory is to minimize the cost of operating **in the long run** by choosing an optimal threshold of probability for taking protective action against some event that can negatively impact the operation. In a simple scenario, the optimal threshold is the ratio of the cost of protecting to the loss if there is no protection for an event that occurs (i.e., the cost loss ratio), but gets more complex when various levels of protective action are involved (Murphy 1985).

As a hypothetical, simple example application of decision theory, consider a case where it costs \$150K to evacuate and deploy an aircraft to avoid damaging winds, defined as surface wind speed > 50 kt. The cost to repair the aircraft if damaged by 50 kt winds is \$1M so the aircraft should be protected if the probability of damaging winds is > 15%. Figure 11 compares

**(a)**

|           |     | Observed?     |                             |
|-----------|-----|---------------|-----------------------------|
|           |     | YES           | NO                          |
| Forecast? | YES | Hit<br>\$150K | False<br>Alarm<br>\$150K    |
|           | NO  | Miss<br>\$1M  | Correct<br>Rejection<br>\$0 |

**(b)**

| Case              | Deterministic Forecast (kt) | Observation (kt) | Cost (\$K)  | Probabilistic Forecast | Cost (\$K) by Threshold for Protective Action |     |             |             |             |             |             |
|-------------------|-----------------------------|------------------|-------------|------------------------|---|-----|-------------|-------------|-------------|-------------|-------------|
|                   |                             |                  |             |                        | 0%  | 15% | 30%         | 60%         | 75%         | 90%         | 100%        |
| 1                 | 65                          | <b>54</b>        | 150         | 42%                    | 150   | 150 | 150         | <b>1000</b> | <b>1000</b> | <b>1000</b> | <b>1000</b> |
| 2                 | 58                          | <b>63</b>        | 150         | 71%                    | 150   | 150 | 150         | 150         | <b>1000</b> | <b>1000</b> | <b>1000</b> |
| 3                 | 73                          | <b>57</b>        | 150         | 95%                    | 150   | 150 | 150         | 150         | 150         | 150         | <b>1000</b> |
| 4                 | 55                          | 37               | 150         | 13%                    | 150   | 0   | 0           | 0           | 0           | 0           | 0           |
| 5                 | 39                          | 31               | 0           | 3%                     | 150   | 0   | 0           | 0           | 0           | 0           | 0           |
| 6                 | 31                          | <b>55</b>        | <b>1000</b> | 28%                    | 150   | 150 | <b>1000</b> | <b>1000</b> | <b>1000</b> | <b>1000</b> | <b>1000</b> |
| 7                 | 62                          | <b>71</b>        | 150         | 85%                    | 150   | 150 | 150         | 150         | 150         | <b>1000</b> | <b>1000</b> |
| 8                 | 53                          | 42               | 150         | 11%                    | 150   | 0   | 0           | 0           | 0           | 0           | 0           |
| 9                 | 21                          | 27               | 0           | 51%                    | 150   | 150 | 150         | 0           | 0           | 0           | 0           |
| 10                | 52                          | 39               | 150         | 77%                    | 150   | 150 | 150         | 150         | 150         | 0           | 0           |
| Total Cost (\$M): |                             |                  | 2.1         |                        | 1.5   | 1.1 | 1.9         | 2.6         | 3.5         | 4.2         | 5.0         |

Figure 11. Cost analysis example. (a) Contingency table. (b) Cost analysis comparison of decisions based on 10 deterministic forecasts versus probabilistic forecasts for the same hypothetical cases.

the cost of operating the aircraft using deterministic or probability forecasts given 10 instances when there was a possibility of getting damaging winds. On the left, protective action is taken when the deterministic forecast exceeds the wind threshold. On the right, many choices of probability threshold are shown for taking protective action when the probability forecast exceeds the level of choice. While 15% is clearly the best choice, the more costly choices are included to show that this method can easily get very costly if not done correctly.

In this example, since the cost of repair is so high in this case, the operators should be willing to spend money on many false alarms to avoid costly repairs. With reliable and sharp probability forecasts, the operating cost will be minimized cost in the long run by only taking action when the probability threshold is exceeded. Making decisions using uncertain deterministic forecasts of the parameter will normally cost more (and potentially much more), unless of course they are perfect. But this method is not a magic bullet to solve all weather related problems since the uncertainty still exists. In the above example, a costly miss will occasionally happen even when the decision maker follows the 15% rule, but that gets absorbed into the long-term costs.

Decision theory fits in very well with the current Operational Risk Management (ORM) movement and the future Machine to Machine (M2M) operating environment, but will be challenging to apply. Defining the optimal probability decision threshold is complicated by such things as considering loss of life, rare or unique events, and complex situations with intertwined decisions and many possible branches. Furthermore, many weather-influenced decisions are not about protecting an asset. However, decision theory can also be put in other terms applicable to DoD operations, such as maximizing weapon effectiveness.

One of the key aspects of applying decision theory is knowing the critical weather sensitivities of the various weapon systems—an area in which the DoD excels. There are numerous sources of such data. An example is the Integrated Weather Effects Decision Aid (IWEDA) that currently links deterministic forecasts to weather sensitivities to provide a stoplight-type chart to the decision maker. For example, in static line parachute operations, the stoplight categories are: green for 0-9 kt surface winds, yellow for 10-13 kt, and red for > 13 kt. A deterministic forecast would fall into one of those categories and give one color, thus disregarding the chance of occurrence in the other categories. Providing reliable probabilities

across all categories (e.g., 65% green, 25% yellow, 10% red) will allow the warfighter to make an informed decision. But again, this would only aid in the decision if the user knew the optimal probability decision threshold.

One avenue that JEFS Phase II may pursue for getting uncertainty information to the warfighter is through the Weather Risk Analysis and Portrayal (WRAP) decision aid, an initiative of the Army Research Lab (ARL). WRAP is ARL's effort to create an interface for forecasters and other users that visually conveys the weather forecast and its uncertainty for any parameter of interest. This project is under development and has made significant progress. JEFS could provide robust EF data to feed WRAP algorithms.

## Project Tasks & Responsibilities

This chapter outlines the various tasks involved with JEFS. AFWA and FNMOC are of course primarily responsible for accomplishment of the tasks, but may delegate as necessary. *< Since manpower resources are currently (as of Jan 05) being established for this work, we purposely avoided assigning tasks to specific offices at this time. That will be part of the final plan (Apr 05). >*

### Timeline

Figure 12 is an approximate timeline of JEFS' milestones, showing the target for periods for development and implementation dates. Table 3 details the deliverables for the two phases. While there is some flexibility in the timeline, tasks should to be assigned to meet these deliverables within the timeline as much as practical since many components are interdependent.

The JGE will likely become operational first since developing it involves only getting existing data together, fitting to a common grid, and building basic products. Data archiving of principal fields and generation of uncalibrated products will begin once all data (GFS and NOGAPS ensemble members) are flowing. Calibration testing will begin after several months of data have been accumulated. Calibrated JGE will be generated after the calibration becomes stable.

JME development contains much more risk and uncertainty and must stay flexible. The goal is to have the initial design of the data assimilation, ICs, model configurations, and surface boundary perturbations ready by the end of 2005. This will allow a several month test and optimization period for JME before realtime implementation in the Spring of 2006.

### Communications and Data Storage

There are many scientific and technical challenges to deal with in this ground breaking effort as the project proceeds, but two critical infrastructure issues that must be resolved ASAP are communications and data storage.

Establishing the bandwidth necessary for data flow between the centers may be a major issue considering the large volumes of data involved. The amount of data and timing of the data sets needs to be estimated and compared to current capabilities. Any shortfalls need to be established and programmed for ASAP. Preliminary estimation indicates that the current DoD's Research and Education Network (DREN) connection between FNMOC and AFWA will be sufficient to meet the data flow requirements between AFWA and FNMOC (for JME, NOGAPS, and

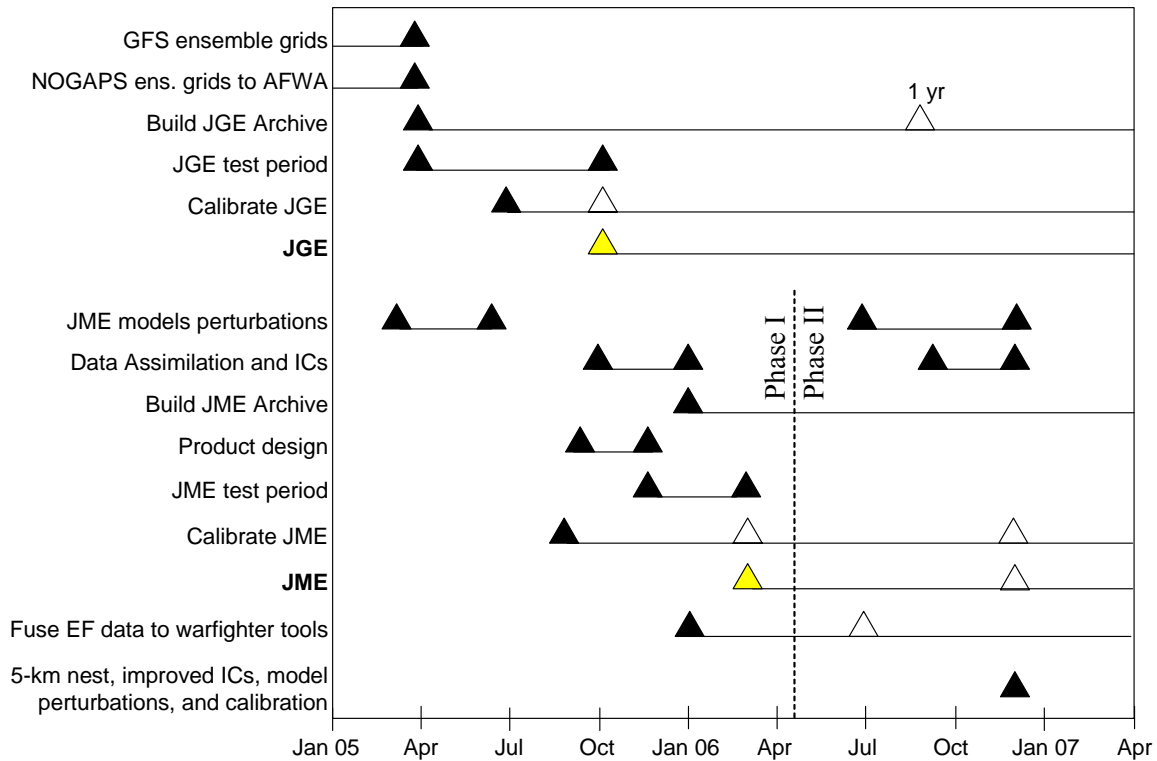


Figure 12. Timeline for JEFS milestones.

Table 3. JEFS Phase I and II deliverables.

| Phase/Period                 | System  | Calibration   | Products  | V&V  |
|------------------------------|---|---|---|--|
| <b>I</b><br>Present – May 06 | <u>JGE</u><br>- 30+ 1°×1° members<br>- 2 cycles/day<br><br><u>JME</u><br>- 30 15-km members<br>- 2 cycles/day<br>- ICs by ET<br>- Multi-model (?)<br>- Model perturbations<br>- SBP perturbations | Weighted Ranks  | <u>Test Samples</u><br>- Confidence<br>- Consensus<br>- Data Range<br>- Probability | <u>Cursory</u><br>- Cons. vs. Cntl.<br>- Spread/Skill<br>- VRH<br>- Reliability diag.<br>- BSS                     |
| <b>II</b><br>May06 – Apr08   | <u>JGE</u> (no change?)<br><br><u>JME</u><br>- 30 15/5-km members<br>- 2 cycles/day<br>- ICs by ETKF ?<br>- Multi-model<br>- Model perturbations<br>- SBP perturbations<br>- Stochastic physics   | <u>PDF estimation</u><br>? Bayesian Model<br>Averaging (BMA)<br><br><u>Bias correction</u><br>- Grid-based<br>- Obs-based | - Interactive Product<br>Generator<br>- WRAP Decision<br>Aid                        | <u>Comprehensive</u><br>- ROCSS<br>- CRPS<br>- Ignorance Score<br>- Econ. val. diag.<br>- Cost-benefit<br>analyses |

calibration data). For data flow from NCEP (GFS ensemble data) to AFWA, a recently upgraded DATUMS-U connection may be sufficient, but needs to be tested. For data flow from NCEP (GFS ensemble data) to FNMOC, currently no direct communication line exists. The options are for FNMOC to download the GFS ensemble files from NCEP ftp server, which may be too slow, or for AFWA to forward the files via DREN.

Meeting data storage requirements is also a concern since extensive data archives on easily accessible media are necessary for both calibration and analysis of results. While this can be limited to principal fields, it will still be many terabytes of data that must be stored. Preliminary estimation shows that necessary storage space is available for Phase I but additional space is likely required for Phase II. The amount of space needs to be estimated and storage devices need to be acquired if current capacity is insufficient.

### **Training and Education**

To successfully demonstrate the value of ensembles in DoD operations, forecasters and decision makers must effectively apply and comprehend ensemble-type products. A basic EF training program should be completed by all forecasters who test and evaluate JEFS products. Education on application of stochastic forecasts in decision making is necessary for warfighters to experiment with using output from decision aids that are linked to JEFS output.

The JEFS project will enlist the aid of the training divisions at FNMOC (????) and AFWA (AFWA/DNT) to help find appropriate existing EF training materials (such as the new COMET module “Ensemble Forecasting Explained”) and develop new DoD specific materials. These materials will be used to train/educate forecasters and leadership at the JEFS test bed locations (28<sup>th</sup> OWS, 607<sup>th</sup> WS, and NPMOC). The final JEFS report will include recommendations for more wide spread education based on experiences and observations during the JEFS program.

One of the most difficult obstacles in training forecasters to use stochastic methods is that the current training and operations environment is primarily deterministic. A good strategy for JEFS to try is to use EF products in a manner that helps forecasters smoothly transition from deterministic to stochastic methods. For example, consider a staged approach for employing the proposed program (Figure 10). At first, it could be used simply to help forecasters quickly identify what problem to focus on in their forecast process. It may even alert a forecaster to a potential event that did not show up in the deterministic model run. Later on at a higher level, a

probogram could be used to determine when to issue a warning using the ideas of decision theory. For a high impact event where a miss is very costly, such as freezing rain, a warning should perhaps be issued at probability as low as 10%. For a lower impact event where many false alarms can disrupt operations, like sfc winds > 35 kt, a warning may only need to be issued when forecast probability reaches 60%.

### **Verification and Validation (V&V)**

Rigorous V&V of forecasts will be necessary to prove the worth of ensembles for military applications. This presents a significant challenge because V&V of a stochastic forecast carries all the problems of deterministic forecast V&V (e.g., what to call truth, issues of scale, etc.) plus the basic problem of comparing a stochastic information to a deterministic truth since stochastic truth is unavailable. The best way to deal with that latter problem is to perform V&V on a very large dataset of results. It is meaningless to say that a probability forecast of 80% chance of winds > 35 kt performed well on a day that the observed wind was 39 kt. It is only after many such forecasts that the skill can be determined. So as discussed above, a large archive of results (minimum of ~6 months) is necessary to thoroughly evaluate JEFs.

The standard suite of EF metrics will be used to analyze the skill and value of the JGE and JME forecasts. Phase I V&V will be a somewhat cursory evaluation using standard deterministic metrics (rmse comparison of JME consensus forecast vs a deterministic control), spread/skill evaluation, verification rank histogram, reliability diagram and Brier skill score (Wilks 1995) for select sensible weather parameters. Phase II V&V will be much more comprehensive, using much larger data sets and additional metrics such as the relative operating characteristic (ROC) diagram (Jolliffe and Stephenson 2003), ROC skill score (ROCSS), economic value diagram, continuous rank probability score (CRPS), ignorance score (Roulston and Smith 2002), as well as complete cost benefit analysis studies.

## Summary

FNMOC and AFWA already provide superb support to the warfighter, but there is now an opportunity to make a quantum leap forward. Transitioning operations to a stochastic approach would provide radically higher value, comprehensive weather information. Such a transition is now possible because of the emergence of the ensemble forecasting (EF) technology. The Joint Ensemble Forecast System (JEFS) will prove the operational feasibility of EF and demonstrate the value of EF to the warfighter. This will be done with innovative EF products designed to incorporate uncertainty into the forecast process and to provide stochastic information, focused on operational weather sensitivities, to the decision makers. These forecasts will be generated from two complementary ensembles: the Joint Global Ensemble (JGE) for medium-range, low-resolution stochastic forecasting and the Joint Mesoscale Ensemble (JME) for short-range, mesoscale stochastic forecasting.

Full utilization of the power of EF will require a revolution in almost all aspects of DoD weather support (education, data processing, analysis, modeling, forecasting, verification, etc.). The sweeping changes necessary are beyond the scope of the JEFS project since it is only designed to build a prototype system. However, JEFS will provide the impetus for FNMOC and AFWA to head towards the future of weather support in which forecast uncertainty is exploited instead of ignored. A warfighter basing a decision on a deterministic forecast, with little or no knowledge of the uncertainty, is at an unnecessary disadvantage. A telling analogy is the blackjack player who cannot count cards but would be greatly advantaged if he could. EF is like being able to count cards—you still don't know exactly what is coming next, but you have a good idea of the possibilities, which helps you place a smart bet.

## References

- Bishop, Craig H., Z. Toth, 1999: Ensemble Transformation and Adaptive Observations. *Journal of the Atmospheric Sciences*, **56**, 1748–1765.
- Eckel, F. Anthony, and M. K. Walters, 1998: Calibrated Probabilistic Quantitative Precipitation Forecasts Based on the MRF Ensemble. *Weather and Forecasting*, **13**, 1132–1147.
- Gneiting, Tilmann, A. Westveld, A. E. Raftery, and T. Goldman (2004). Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. Technical Report no. 449, Department of Statistics, University of Washington. Available at: <http://www.stat.washington.edu/MURI/>
- Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification; A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons, Ltd., West Sussex, England, 240 pp.
- Murphy, Allan H., 1985: Decision Making and the Value of Forecasts in a Generalized Model of the Cost-Loss Ratio Situation. *Monthly Weather Review*, **113**, 362–369.
- Roulston, Mark S. and L. A. Smith. 2002: Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review*. **130**, 1653–1660.
- Toth, Zoltan, and E. Kalnay, 1997: Ensemble Forecasting at NCEP and the Breeding Method. *Monthly Weather Review*, **125**, 3297–3319.
- Wang, Xuguang, and C. H. Bishop, 2003: A Comparison of Breeding and Ensemble Transform Kalman Filter Ensemble Forecast Schemes. *Journal of the Atmospheric Sciences*, **60**, 1140–1158.
- Yuejian Zhu, Z. Toth, R. Wobus, D. Richardson, and K. Mylne. 2002: The Economic Value Of Ensemble-Based Weather Forecasts. *Bulletin of the American Meteorological Society*, **83**, 73–83.
- Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.