AFWA's Joint Ensemble Forecast System (JEFS) Experiment

Evan Kuchera, Timothy Nobis, Steven Rugg, Scott Rentschler, Jeff Cunningham, Jim Hughes and Matthew Sittel,
Air Force Weather Agency

Introduction

As technology has progressed in recent decades, environmental information for mission planning and execution has been increasingly generated using numerical methods. These methods have been almost exclusively deterministic (e.g. single answer with unknown certainty), with human forecasters primarily creating deterministic products using these models.

It is of course commonly accepted that deterministic weather forecasting can be imperfect. The cause of this is summed up in the National Research Council's 2006 white paper "Completing the Forecast:"

"The chaotic character of the atmosphere, coupled with inevitable inadequacies in observations and computer models, results in forecasts that always contain uncertainties. These uncertainties generally increase with forecast lead time, and vary with weather situation and location."

The question may then arise—why are so many current environmental forecast products created deterministically when these uncertainties are known to exist? One primary reason has been the success of deterministic forecasting for many weather phenomena. Large scale weather patterns are often well predicted, and the uncertainties on benign days are generally not important to most users of the information and therefore not noticed. Another reason is the need of the end user to make a deterministic decision with the information. You cannot half close school or half fly a mission; so, these users want a yes or no answer as to if the weather will require a change of plans. Finally, those forecasters who present uncertainty can give the impression of "hedging" and therefore can be seen as incompetent or untrustworthy.

Corresponding Author:
Evan Kuchera: kucherae@offutt.af.mil

The inevitable conclusion to deterministic forecasting with uncertain information is that some percentage of these forecasts will end up being a "bust." This is an undesirable situation for everyone involved, as it makes the forecasters look even more unintelligent than if they had hedged, and with the incorrect information costly or even potentially dangerous decisions can result.

Therefore, the thesis of this project was the following: **How can the DoD create and communicate environmental information that includes relevant information about certainty?**

To be relevant to a decision maker, information needs to have the following qualities:

Timely (available before the decision is made)
Communicative (information is easy to understand)
Focused (directly impacts the decision)
Useful (not something that is already known)
Reliable (information is generally correct)

It is this last quality where certainty becomes crucial, as a probability can be reliable but still retain a measure of the certainty, whereas a deterministic forecast cannot. If when a forecaster says there will be a 20% chance of thunderstorms they are observed 20% of the time, this information is reliable and correct. However, one cannot forget that the information needs to be useful—a forecast of a 20% chance of thunderstorms every day in the summer (e.g. "climatology") may be reliable but it can be of limited usefulness to a decision maker.

So the goal of this project was not only to create reliable forecast information that was also useful, but that also retained the qualities of timeliness, communicativeness, and focus.

Economic Value

Using probabilistic weather forecasts for military operations is not a new notion. Scruggs (1967) identified the value of probabilistic forecasts in a military context, highlighting the additional information operators can glean from a probabilistic forecast. Eckel et al. (2008) further demonstrate this potential by objectively evaluating the use of

probabilistic forecasts in idealized defensive and offensive military scenarios. The stochastic operator yields a 30% cost savings over a deterministic operator for the defensive scenario (decision to move aircraft with an approaching Typhoon). On the offensive side, a stochastic operator can lead to the destruction of enemy air defenses faster than a using a deterministic operator.

The energy industry uses temperature forecasts for their load predictions, accounting for 40% to 90% of load forecast error (Altalo and Smith, 2004). Poor load forecasts can result in large financial costs, especially if the load is under-forecast. Clearly this is a situation where an improved forecast can have a significant financial impact. In fact their study showed "…that weather forecast accuracy can be significantly improved using a multi-model ensemble; and that use of probabilistic information represents reduction of costs of over 50% to the load forecaster." These kinds of benefits could be translated to a variety of military applications.

Methodology

Certainty diagnosis can be accomplished through many techniques. Climatology is the most straightforward, as the historical probability of an event can be used to estimate its probability in the future. Statistical techniques can also be used to correlate deterministic numerical model output to the probability of observing certain phenomena (Glahn and Lowry 1972), thereby combining information about the current state of the weather and the past performance of the model.

A third method of determining certainty is by running multiple dynamic numerical models. Each model is designed to be an equally likely possibility of the future state of the atmosphere by perturbing the initial conditions and model equations within their margin of error. The "flow dependent" impact of these uncertainties on the resulting model output can then be assessed. This differs from the purely statistical technique in that the uncertainties on a given day are allowed to interact physically as the model integrates, producing non-linear changes that cannot be discovered statistically.

A combination of statistical and dynamical techniques can be used to take advantage of uncertainties due to the flow of the day, in addition to uncertainties in formulation of the dynamically generated ensemble (Gleeson 1970). In this way non-linear uncertainties can be simulated dynamically; moreover model biases and deficiencies in ensemble formulation can be mitigated statistically, resulting in both useful and reliable probabilities. This statistical correction of a dynamically generated ensemble is referred to as "calibration."

Real-Time Data Flows

.

Both model and observation data (~30 GB/day) were ingested for JEFS. Four global models were ingested, including three ensembles and one deterministic used for gridded verification and calibration. No mesoscale models were ingested due to communication limitations. Navy sea surface temperatures and output from AFWA's AGRMET land surface model were also ingested.

Numerous difficulties in reliability, availability, and format were encountered with ingesting data for the JEFS experiment. However, information assurance issues were the most substantial, largely because the data flows were not operational so problems were not considered a priority and dealt with in a timely manner. These problems prevented a real-time exchange and evaluation of an AFWA-FNMOC combined mesoscale ensemble.

Ensembles

A Joint Global Ensemble (JGE) of 79 members was created using one degree global ensemble data from the National Centers for Environmental Prediction (NCEP), Fleet Numerical Meteorology and Oceanography Center (FNMOC), and the Canadian Meteorological Centre (CMC) ensemble. A Joint Mesoscale Ensemble (JME) was created by running 10 independent models within the WRF framework with varied physics (Table 1) and initial/boundary conditions from the GFS global ensemble. AFWA was able to create a real-time multi-center mesoscale ensemble over CONUS using NCEP's mesoscale ensemble (SREF). AFWA domains currently running in real-time can be seen in Figure 1.

Table 1. WRF physics packages for each member of the AFWA portion of JME.

Member	Physics Packages						Stamp Placement
	Surface	PBL	Cumulus	Micro-physics	Longwave Radiation	Shortwave Radiation	
1 (3)	Thermal	MRF	Grell	WSM3	CAM	Dudhia	Top Center
2 (4)	Thermal	YSU	Grell	Ferrier	CAM	CAM	Bottom Right
3 (5)	Thermal	MYJ	KF	WSM6	RRTM	CAM	Bottom Left
4 (9)	Noah	MRF	KF	Lin	RRTM	CAM	Middle Right
5 (10)	Noah	YSU	KF	WSM5	RRTM	Dudhia	Large Control
6 (11)	Noah	MYJ	Grell	WSM5	RRTM	Dudhia	Top Right
7 (15)	RUC	YSU	BM	Lin	CAM	Dudhia	Bottom Center
8 (16)	RUC	MYJ	KF	Ferrier	RRTM	Dudhia	Middle Left
9 (17)	RUC	YSU	BM	Ferrier	RRTM	CAM	Top Right
10 (18)	RUC	YSU	Grell	WSM6	CAM	CAM	Middle Center

Product Suites

Recall focused data "directly impacts the decision." For the most part, the raw data from the ensemble were not very focused, and required some sort of additional post-processing to directly forecast the variable of interest. This introduced another level of uncertainty that needed to be accounted for: algorithm uncertainty. JEFS was unable to explore probabilistic algorithms for all needed forecast variables but progress was made on a few. Specifically, lightning and visibility algorithms were created by researching physical causes of the phenomena, and then choosing potential predictors of the phenomena from the model output. Once predictors were selected, a probabilistic forecast was made using logistic regression. Therefore, the uncertainty in the predictors was accounted for in the regression, leading to a more reliable probability. A relatively simple algorithm was also developed to forecast the probability of rain, freezing rain, snow, and mixed precipitation.

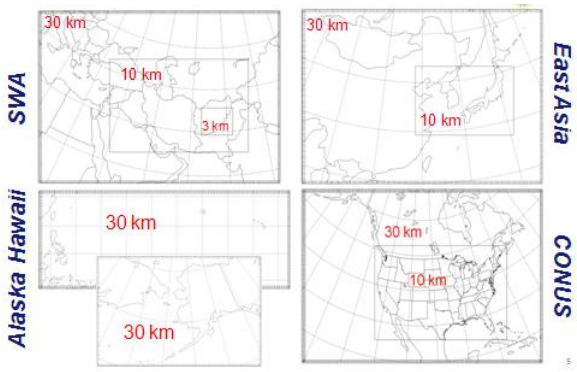

Figure 1. AFWA real-time mesoscale ensemble domains as of April 2009.

Some forecast variables are more challenging to predict than others, specifically those that occur at scales much smaller than the model resolution. Unfortunately, variables that most impact the warfighter often fall into this category. Predicting these variables reliably and usefully requires not only a well designed ensemble to account for flow uncertainties, but also statistical and physical algorithms to account for the sub-grid scale uncertainties due to model deficiencies. For JEFS, algorithms with uncertainty were used to predict visibility, lightning, electromagnetic ducting, and precipitation type, and results, while subjective, were mostly favorable.

Customer Interaction

There is a wide spectrum of potential users of ensemble information. At one end of the spectrum, sophisticated users with a high degree of weather knowledge are likely to want as much information as possible, while still receiving it in a clear and concise form. These users are more interested in information about the physical processes in the atmosphere so they can create a conceptual model in their minds and therefore fill in gaps where the model does not or cannot provide information. For these users more complex products like meteograms (Figure 2) and stamp charts are more appropriate. At the other end of the spectrum are users who want to know the specific chance of their problematic atmospheric condition occurring in a given time frame. For these users probabilities describing the certainty of a mission-impacting event are most appropriate (Figures 3-5). In between are products describing the range of possible values of a variable, such as a mean/range product.

Potential users of the ensemble data were engaged during JEFS to evaluate utility and provide feedback for improvement. Imagery was placed on a website in real-time and forecasters were quite receptive. Feedback on utility was very positive, and numerous additional products were created as a result of customer requests. Some examples of feedback:

45 WS (Cape Canaveral): *"Our launch weather officers have looked at the JME products, particularly the probability plots (winds, precip, and lightning). These products are definitely value-added, as they provide information in the format we need….we would like to see JEFS products made permanent over our AOR."*

51 OSW (Korea): *"The <1 mile Vis probability was instrumental in deciding if the first goes of the day should be scrubbed which included the 7th AF CC flying."*

28 OWS (Southwest Asia): *"Preliminary findings for the "New" dust lofting product are quite promising."*

17 OWS (Pacific): *"Integrating JEFS analysis in place of "regime-based forecast processes" mitigates the risk of incorrectly eliminating consideration of possible weather threats"*

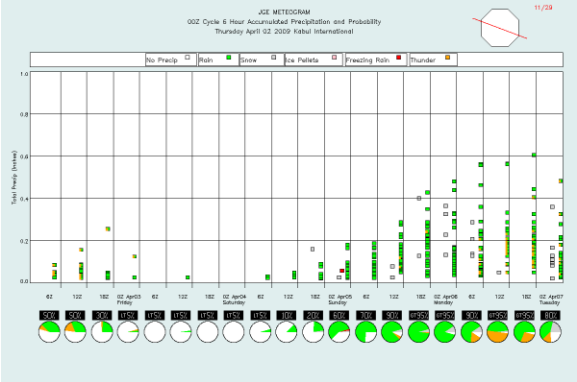

Figure 2. Example of an ensemble precipitation meteogram. Precipitation from each member of the ensemble is plotted as a square, and the total ensemble probabilities are below for each time.

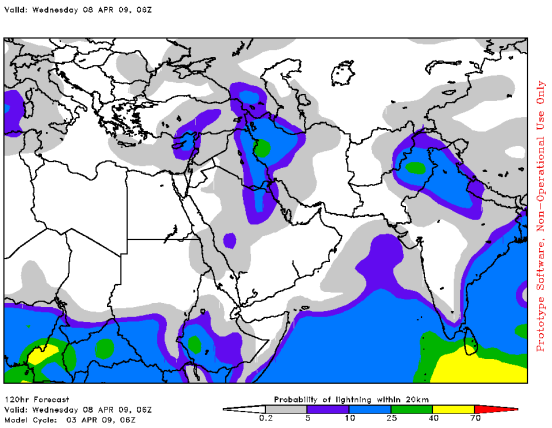

Figure 3. Example of an ensemble lightning probability forecast.

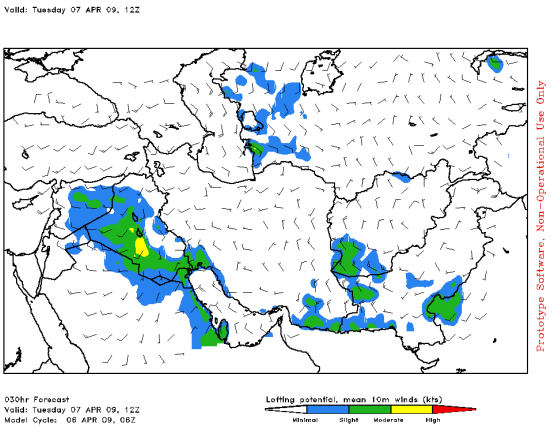

Figure 4. Example of an ensemble dust lofting potential forecast. Several predictors favorable for lofting dust are combined into one index for the entire ensemble.

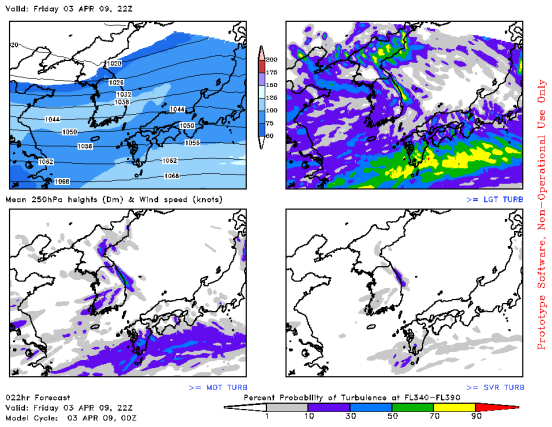

Figure 5. Example of an ensemble 4-panel turbulence forecast.

Validation

Objective verification showed that the global ensemble was both useful and quite reliable. This was attributed in large part to the fact that ensembles from multiple centers were included in its design. Subjective verification of high-impact events in the global ensemble showed that while JGE was able to predict the large scale features (e.g., upper waves and deep layer moisture) with usefulness, it was largely unable to simulate the finer scale details owing to its coarse resolution. This resulted in forecast output that remained reliable, but was not as useful. The single NCEP global ensemble was compared to the combined JGE ensemble, as well as bias correction of both using a 30 day history at the grid point compared to the UKMET office global analysis. The combined ensemble was almost exclusively superior in skill and reliability to all other methods (Figure 6).

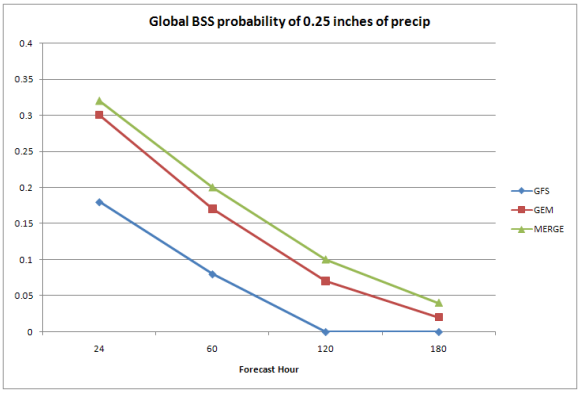

Figure 6. Brier skill score (compared to 30-day model climatology) for the GFS ensemble, GEM ensemble, and the two combined for the probability

of 0.25 inches of precipitation in 6 hours. Note that even though the GFS scores worse than the GEM, when it is combined with the GEM the skill increases—a non-intuitive result.

Objective verification (Figure 7) showed that the mesoscale ensemble was more useful than the global ensemble, but not as reliable as the global ensemble. The lack of reliability was attributed to a design limitation of only using NCEP ensemble members for initial conditions, a lack of mesoscale data assimilation, and a limited number of ensemble members. Increased usefulness was likely due to the improved resolution of smaller scale phenomena. Subjective verification of high impact events in the mesoscale ensemble showed that it could forecast useful probabilities of "rare" weather events the global ensemble was unable to detect. Reliability was often less than desired more than 24 hours before an event. Improvement was often noted when increasing horizontal resolution, but the value of increased horizontal resolution for any reason other than to better simulate terrain when resolutions are less than 10-15 km is uncertain.

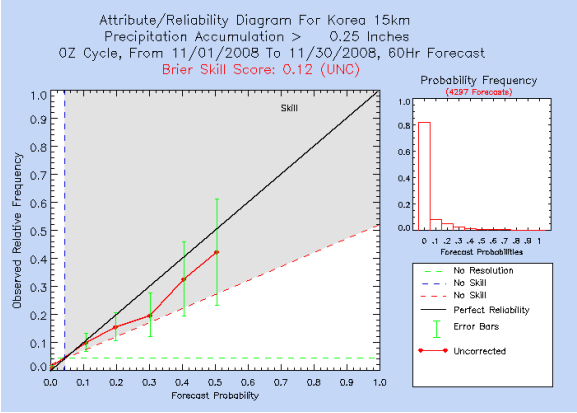

Figure 7. Attributes diagram for the JME over the Korean domain for 60 hour forecasts of probability of 0.25 inches of precipitation. Note that the JME never forecasts more than a 50% chance, but remains reliable (e.g. 0.25 inches is observed roughly 50% of the time it is forecast).

Discussion and Future Work

As a result of the successes in JEFS, plans are being made to make ensembles an official part of day-to-day weather forecasting operations in the Air Force. Much work remains however to ensure the information is of the highest quality, and is used appropriately.

It was clear from JEFS that initial conditions are the most important aspect of a good numerical forecast, and correctly sampling the uncertainties in the initial conditions is the most important aspect of a good ensemble forecast. Model diversity is important for specific applications where the physics are directly relevant to the variable of interest (e.g. boundary layer, precipitation). Model resolution is also important to simulate scales of interest—in JEFS, resolutions around 10 km simulated most high-impact events (even parameterized convection) well. Only when dictated by complex terrain is an ensemble of higher resolution necessary based on what was found in JEFS, although improvement is certainly possible with higher resolutions. The question is the benefit of that increased resolution versus its cost.

The JME only utilized 10 members and the result was an under-dispersive forecast. However, the lack of initial condition spread from the NCEP GFS ensemble is more likely responsible than the lack of members. A small ensemble that is well designed is preferable to a large ensemble that is poorly designed, so more effort should be given toward design than toward increasing membership. It was also clear at times that the lack of mesoscale information at the initial time hurt the ensemble. When the synoptic situation was quiet before a mesoscale event, the ensemble often simulated it quite well. When the synoptic situation was complex, with a lot of mesoscale detail that the one degree global ensemble could not simulate, the mesoscale ensemble was very under-dispersed. This argues for some method to get mesoscale information into the initial conditions. The Ensemble Transform Kalman Filter (ETKF) was attempted in this experiment, but since perturbations were added to a coarse global analysis mean, any significant errors in the analysis rendered the perturbations ineffective. The question of how best to create mesoscale initial conditions remains an open one.

A multi-center (e.g., entire ensembles from NCEP, CMC, & FNMOC) global ensemble showed great promise in this experience, and has also shown promise in the experience of other users. It is unknown if this promise will transfer to a mesoscale ensemble, but the idea of using a multi-center global analysis to initialize a deterministic model could result in substantial forecast improvement. Ideally, a number of analyses from several global ensembles could be used to initialize a mesoscale ensemble. Additionally, time lagging the global ensemble to increase membership may be an acceptable practice

for longer term forecasts, although a definitive answer was not found in JEFS.

Systematic model bias was observed during JEFS for some variables, particularly for JME. Individual members sometimes showed substantial biases (e.g. surface dewpoint) while the ensemble as a whole appeared to over-forecast precipitation. These biases can be removed by improving the physics formulations so that the improved information can interact within the model run. Until model improvement occurs, statistical techniques can also mitigate the problem.

Smart software design for the mesoscale ensemble created by NCAR allowed for a high level of efficiency and flexibility. JEFS was able to configure physics options and processing allocations for different domains and ensemble members easily. This allowed for rapid progress learning about the ensemble and its behavior. Some inefficiencies remain—developers need to keep timeliness in mind when creating new techniques if they are to be used in an operational environment. With this in mind, the JME software design resulting from the JEFS project is well postured for transition to operational implementation.

Ensemble post-processing requires even more flexibility than the ensemble model to maintain a timely and reliable flow of information. A risk always exists that failures in models can occur; this risk rises when more models are being run in ensembles. Additionally, given the increasing evidence of the benefits of multi-center ensembles, a robust system needs to be able to operate on a variety of different data sets from different centers with potentially missing information.

One of the most significant hurdles that must be overcome is the processing and I/O burden. The ensemble must be parallelized and efficiently address the I/O requirements that will inevitably arise as the number of datasets grows. Further, the software design of the system must be modular and flexible; this enables the administrator of such a system to make additions and modifications to the system as new needs and requirements arise. Writing code that is well documented, easy to read, and robust is obligatory.

In JEFS, ensemble member output was post-processed for basic variables using the community WRF post-processor, while most derived variables necessary for the focused ensemble products like lightning or turbulence were created within the ensemble post-processor. This is an unresolved area of design for an ensemble system—should all the post-processing be done locally (which requires more resources), or should some or all of it be done by a single community group (which leads to less control). In JEFS using a community post-processor to get the raw model state variables into a more digestible format worked well, with the more DoD specific post-processing needs saved for local production. Tension remains here however—if more complex variables are needed from the raw model output (e.g. hydrometeors, aerosols) a community approach may be less tenable.

An area that had not been addressed much outside of JEFS is the usage of algorithms that contain uncertainty with ensembles. This accounts for both the uncertainty in the flow and the in the diagnosis of the variable of interest. Much more work could be completed in this area, with potentially substantial benefits.

Statistical techniques exist to calibrate ensemble output based on observations, resulting in more reliable ensemble output. While these techniques are somewhat established, the reliance on observations leads to problematic application for the DoD, which often needs weather information in data-sparse/data-denied areas, and for variables that are not always readily observable (e.g. turbulence, clouds, dust). Gridded calibration can be used to mitigate the lack of direct observations, but the availability of a quality independent gridded data set (e.g. model climatology, model from another center) is not assured, especially on short notice. Finally, purely statistical calibration techniques may not capture rare events if they are designed to improve the majority of forecasts. Arguably, these rare events are more important to DoD.

Therefore a focus on algorithm development that utilizes sound physical principles and broad applicability to different regions and scales for high-impact variables of interest to the DoD should be the primary focus for post-processed ensemble forecast improvements, with statistical calibration from observations used when relevant and worthwhile.

When DoD implements a multi-center mesoscale and/or global ensemble in the future, preparations must be made in advance to reliably retrieve member ensembles in a timely manner in accordance with existing information assurance rules and policies. As this project has experienced, failure to plan in advance for this can result in a show-stopping implementation event. As a result of this experience,

the groundwork has already been laid to establish alternative methods to exchange data, and these are also running into significant problems and delays.

Having a flexible development environment where users could see and comment on products was extremely helpful in JEFS. Although there is risk that the users will inappropriately use a test product, the benefits of feedback and education that come from this collaborative development effort outweigh the costs.

Experiences with users in JEFS showed that local units are willing to develop training on their own and can readily utilize stochastic information if it is presented in a clear, concise format and if it is relevant for their forecasting challenges. Full exploitation will require not only training on stochastic concepts and philosophy, but also on how to appropriately use whatever software tools get developed to interrogate ensemble data. The most success was achieved in JEFS when new users of ensemble data had a knowledgeable ensemble person imbedded to explain the concepts and products. When users were asked to evaluate the information without much personal interaction, it did not take hold. Therefore, training may be best accomplished in the short term by training a few personnel in each forecasting unit and allowing them to propagate their knowledge. These personnel could then work closely with DoD ensemble subject matter experts and training leads to ensure maximum product and training relevance.

Acknowledgments

JEFS was a multi-center and multi-institution experiment involving many very hard working and intelligent individuals, to whom the authors wish to express their gratitude.

References

Altalo, M. G., and L. A. Smith, 2004: Using ensemble weather forecasts to manage utilities risk. *Environmental Finance*., **Oct. 2004**, 48-49.

Eckel, F. A., J. G. Cunningham, and D. E. Hetke, 2008: Weather and the calculated risk. *Air & Space Power Journal*., **Spring 2008**, 71-82. [Available online at http://www.airpower.maxwell.af.mil/airchronicles/apj/apj08/spr08/Eckel.html.]

Glahn H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.

Gleeson T. A., 1970: Statistical-dynamical predictions. *J. Appl. Meteor.*, **9**, 333–344.

Scruggs, F. P., 1967: Decision theory and weather forecasts: A union with promise. *Air University Review*., **Jul-Aug 1967**, 53-57. [Available online at http://www.airpower.maxwell.af.mil/airchronicles/aureview/1967/jul-aug/scruggs.html.]