Aspects of Effective Mesoscale, Short-Range Ensemble Forecasting

F. Anthony Eckel

Air Force Weather Agency, Offutt AFB, Omaha, NE, 68113

And

Clifford F. Mass

Department of Atmospheric Sciences
University of Washington
Seattle, WA, 98195

Submitted to Weather and Forecasting on 23 February 2004

Revised on 21 June 2004

Corresponding author address: Major F. Anthony Eckel

106 Peacekeeper Dr., STE 2N3

Offutt AFB, NE 68113-4039

e-mail: tony.eckel@afwa.af.mil
ABSTRACT

This study developed and evaluated a short-range ensemble forecasting (SREF) system with the goal of producing useful, mesoscale forecast probability ($FP$). Real-time, 0 to 48-h SREF predictions were produced and analyzed for 129 cases over the Pacific Northwest. Eight analyses from different operational forecast centers were used as initial conditions for running the Fifth-Generation Pennsylvania State University–National Center of Atmospheric Research Mesoscale Model (MM5).

Model error is a large source of forecast uncertainty and must be accounted for to maximize SREF utility, particularly for mesoscale, sensible weather phenomena. Although inclusion of model diversity improved $FP$ skill (both reliability and resolution) and increased dispersion toward statistical consistency, dispersion remained inadequate. Conversely, systematic model errors (i.e., biases) must be removed from a SREF since they contribute to forecast error but not to forecast uncertainty. A grid-based, two-week, running-mean bias correction was shown to improve $FP$ skill through: 1) better reliability by adjusting the ensemble mean toward the mean of the verification values, and 2) better resolution by removing unrepresentative ensemble variance.

Comparison of the multimodel (each member uses a unique model) and varied-model (each member uses a unique version of MM5) approaches indicated that the multimodel SREF exhibited greater dispersion and superior performance. It was also found that an ensemble of unequally likely members can be skillful as long as each member occasionally performs well. Lastly, smaller grid spacing led to greater ensemble spread as smaller scales of motion were modeled. This study indicates substantial utility in current SREF systems and suggests several avenues for further improvement.
1. Introduction

Operational use of mesoscale, short-range (0 – 48 h) ensemble forecasting (SREF) has lagged far behind the successful implementation of medium-range (2 – 10 d) ensemble forecasting systems at NCEP and ECMWF (Toth and Kalnay 1993; Tracton and Kalnay 1993; Molteni et al. 1996). While research into mesoscale SREF application has been generally positive (Du et al., 1997; Stensrud et al. 1999; Stensrud et al. 2000; Hou et al. 2001; Wandishin et al. 2001; Grimit and Mass 2002; Stensrud and Yussouf 2003), many questions remain concerning optimal design and potential benefits. At the University of Washington, a real-time, mesoscale SREF system was implemented to evaluate the value of SREF over the mountainous Pacific Northwest and to explore methods for advancing its effectiveness (i.e., the utility to the customer).

Compared to a medium-range ensemble system, some additional challenges posed for SREF include:

- The small-scale, surface variables of interest to SREF are less predictable and their errors may saturate too quickly for an ensemble to be of use.
- Model error, which is poorly understood and difficult to account for, has a larger impact on the surface variables in the short range (Stensrud et al. 2000).
- The best method for defining the initial conditions (ICs) is unclear since large-scale error growth is initially linear (Gilmour et al. 2001). Most ensemble IC methodologies (e.g., breeding) were developed for the medium range in which nonlinear error growth generates a large spread of solutions.
- Use of a limited area model may inhibit ensemble dispersion even when perturbed lateral boundary conditions (LBCs) are applied (Nutter 2003).
- It may be important to capture variability at small scales using very high resolution.
The research primarily deals with further study of the impact of model error on the quality of SREF-based forecast probability \( (FP) \), but some of the other issues noted above are also addressed.

The primary goal of ensemble is probabilistic prediction of weather events (Epstein 1969; Leith 1974), the utility of which is maximized by producing a sharp probability density function (PDF), subject to calibration (T. Gnetting 2003, personal communication). A sharp PDF is narrower and generates \( FP \) with higher \textit{resolution}—the ability to distinguish events from nonevents. Calibration is an adjustment to an ensemble’s PDF to improve \textit{reliability}—the ability of \( FP \) to match the observed relative frequency of an event over many cases. Since calibration of an underdispersive ensemble can negatively impact resolution, adding “good spread” to such an ensemble is highly desirable. The term good spread is defined as an increase in ensemble variance that simultaneously improves statistical consistency (i.e., ensemble variance matches the mean-square error, \( \text{mse} \), of ensemble mean), reliability, and resolution. For instance, adding noise does not create good spread since a decreased in resolution would result.

The importance of accounting for model error to improve ensemble dispersion is well established (Houtekamer et al. 1996; Stensrud et al. 2000; Mylne et al. 2002). When using an imperfect model, to have a chance at drawing members from the same forecast PDF as the verification, the members should ideally have various model attractors that bound the true attractor (Hansen 2002). However, there is still a question of how much of the forecast uncertainty arises from analysis error versus model error. Stensrud et al. (2000) found that the relative influence of model uncertainty becomes quite large when considering mesoscale, sensible weather phenomena in the warm season. Our study demonstrates that the impact of model error is still significant even in the cool season over complex terrain where forecast
uncertainty is largely driven by synoptic-scale errors originating from analysis uncertainty.

Two general methods have emerged to account for model error. One method, called model diversity in this paper, is for each member to have a different model attractor. Model diversity can be achieved by either a multimodel technique, in which completely different models are employed, or by a varied-model technique, in which only one model (e.g., the MM5) is used but each member uses a unique version of that model (i.e., varied combinations of model physics options and/or perturbed parameterizations). A varied-model ensemble likely generates less dispersion compared to a multimodel ensemble in which model differences may be much larger. Both model diversity techniques were applied in this research to explore the question of whether a varied-model ensemble can compete (in dispersion and skill) with multimodel ensemble.

A second method to account for model error (not applied in this research) is called stochastic physics, in which random error is added to the evolving solution during model integration. Introduced by Buizza et al. (1999), stochastic physics attempts to account for the effects of model error by randomly perturbing the tendency of state variables with an appropriate scaling and spatio-temporal autocorrelation. While stochastic physics does increase spread and improve performance, it likely fails to represent the full spectrum of model uncertainty, particularly if there are major deficiencies in the parameterizations (Evans et al. 2000; Ziehmann 2000; Richardson 2001). The chief limitation of stochastic physics may be the use of the same model attractor in each member, making the solutions too similar. As noted in Mylne et al., 2002, the differences among the ECMWF EPS members, perturbed by stochastic physics, fall well short of those found in a multimodel ensemble. Ideas for improving upon the stochastic physics technique have been proposed (Palmer 2001) and research is ongoing.

Adding model diversity to an ensemble does not necessarily provide good spread because of
the negative influence of model bias. While dispersion is increased, reliability and resolution may not be improved and statistical consistency can be degraded due to an increase in the mse of the ensemble mean. Only a few papers have addressed the need for bias correction in ensemble forecasting (Richardson 2001; Atger 2003; Stensrud and Yussouf 2003). This study examines the benefits of bias correction and shows that it is critical for mesoscale SREF since: 1) mesoscale models often exhibit significant bias, and 2) model diversity amplifies the negative impact of bias. It is only after bias correction that model diversity can dependably add good spread.

Section 2 of this paper describes how the data and results of this research were generated. Section 3 details the bias correction design and its impact to SREF. Section 4 discusses the impact of model error on mesoscale SREF as well as possible ways to improve SREF. Section 5 presents conclusions and recommendations.

2. Data and methodology

a. Single-Model Multianalysis Ensemble

Grimit and Mass (2002) implemented a five-member multianalysis SREF by using analyses/forecasts from five operational forecast systems to provide a spread of initial conditions (ICs) and LBCs to drive the Fifth-Generation Pennsylvania State University–National Center of Atmospheric Research Mesoscale Model (MM5). Encouraged by their successful prediction of mesoscale forecast skill, their design was expanded in this research to the eight-member University of Washington Mesoscale Ensemble (UWME), using an outer domain with 36-km grid spacing and a 12-km inner nest (Figure 1) with 32 sigma levels. UWME has only IC and LBC perturbations and does not account for model uncertainty. Data were collected for 129 forecast cases during the 2002-03 cool season, all initialized at 0000 UTC (Figure 2).
The eight IC/LBC sources of UWME are displayed in Figure 3 and the MM5 configuration (optimized for the Pacific Northwest) used for all UWME members is shown in Figure 4. A key assumption of the multianalysis approach is that the analyses can be considered random samples of the analysis PDF, which is challenged by: 1) similarities (i.e., correlation) among analyses with shared observational data and objective analysis techniques, and 2) unequal quality of the analyses due to use of different observation networks (e.g., satellite data) and the varied skill of the different model first guesses and objective analysis schemes. The success of Grimit and Mass (2002) and this research suggest that the multianalysis approach does provide a valuable approximation of the analysis PDF. An important reason for the success of the multianalysis approach is that many mesoscale weather phenomena are driven by the synoptic-scale flow, particularly in areas of complex terrain such as the Pacific Northwest (Mass et al. 2002). Indeed, Errico and Baumhefner (1987) found that the largest component of the mesoscale forecast error originates from synoptic-scale errors in the analysis. Furthermore, Tribbia and Baumhefner (2004) demonstrated that the up-scale cascade of error growth from the smaller to the larger scales is of lesser importance compared to the exponential growth originating within the synoptic scales. Thus, using different synoptic analyses as ICs is a useful technique for mesoscale SREF.

b. Varied-Model Multianalysis Ensemble

Houtekamer et al. (1996) was the first to apply the varied-model approach by varying four model options (horizontal diffusion, convection/radiation, gravity wave drag, and orography) and three surface parameter perturbations (SST, roughness length, and albedo) in a medium-range ensemble system that also included IC perturbations. They concluded that while using model variations did make an improvement, “more dramatic perturbations to the model” would be required to produce more realistic error growth. Stensrud et al. (2000) applied the varied-model
approach to an MM5 ensemble for two forecast cases by varying two model options (convection and PBL) and moisture availability. They found the spread grew two to six times faster in the first 12 h compared to an ensemble with no model diversity, but that the system was still underdispersive.

To test the varied-model approach for SREF over the Northwest, the UWME+ system added model variations to the UWME members. The model diversity of UWME+ is meant to enhance ensemble dispersion while staying within the bounds of suspected uncertainty. Figure 4 shows that each UWME+ member uses a fixed combination of seven MM5 model options (PBL, soil, vertical diffusion, cloud microphysics, convection, shallow cumulus, and radiation) and four surface parameter perturbations (SST, moisture availability, albedo, and roughness length), each tied to one of the eight analyses. The combinations of model options were chosen arbitrarily without regard to the appropriateness of how various options interface or the resulting relative skill between members. The point was to generate significant diversity and then to examine the skill of the resulting ensemble. Analysis revealed that the relative skill among the UWME+ members was primarily controlled by each member’s IC/LBC and not by the model variations. One could argue that the fixed model variations adversely constrains UWME+ since randomizing the variations from day to day might capture more of the possible model error. However, since forecast bias is dependent upon model options and the initializing analysis, such randomness was purposefully avoided to allow for bias correction based on the previous performance of each member (discussed below).

Figure 4 also shows that two very different components of model variability exist in UWME+. Use of different model options assumes that the differences between model options are reasonable approximations of model uncertainty—a difficult assumption to support since
model error is poorly understood. Accounting for surface parameter uncertainty is accomplished more rigorously in UWME+ through random perturbations that mimic the suspected errors.

In the UW MM5, SST is updated each forecast cycle but is held constant during forecast integration, creating error since over the open ocean the diurnal variation of SST averages 0.3°C and can be much higher near land or in turbulent waters (Clancy and Sadler 1992). Larger possible errors result from the objective analysis of SST in which buoy and satellite observations are combined with a model first guess from the Optimum Thermal Interpolation System. The typical SST rmse is 0.5°C –1.0°C in the Eastern Pacific (Clancy and Sadler 1992) and the errors have a high degree of spatial correlation with a length scale of roughly 150 km (J. A. Cummings 2002, personal communication). Figure 5 shows the random SST perturbation field used for UWME+ member #1 that uses a conservative average perturbation of 0.7°C and a coherent spatial structure that approximates the spatial correlation inherent to the errors.

MM5 represents the spatial variability of moisture availability, albedo, and roughness length by assigning two seasonal values (winter and summer) to twenty four land use categories in a master land use table. Each model grid box is classified with a single land use category and thus fixed (by season) values of the surface parameters, which can result in significant error on any particular forecast cycle. Each member of UWME+ was given a unique land use table constructed using random draws from PDFs designed to represent the possible values of the surface parameters. The size and shape of the PDFs were based on the seasonal values in the standard land use table, empirical data from Pielke (2002), and a subjective evaluation of the typical variability within a grid box. Figure 6 provides a few example surface parameter PDFs.

c. Multimodel Multianalysis Ensemble

For medium-range ensembles, several recent studies have demonstrated the value of the
multimodel approach, which assumes that the differences between the model forecasts are representative of model uncertainty. Evans et al. (2000) concluded that benefits provided by the multimodel technique are likely “due to the sampling of different, skillful populations provided by the individual systems.” Studies by Richardson (2001) and Mylne et al. (2002) showed that a multimodel multianalysis ensemble outperformed the ECMWF Ensemble Prediction System.

In this research, a multimodel multianalysis ensemble, commonly called a Poor Man’s Ensemble (PME), was available as a byproduct of the design of UWME and UWME+. The PME, which consisted of the eight operational forecast models shown in Figure 3, were fit to the 36-km grid using bilinear interpolation as part of the MM5 preprocessing. Ziehmann (2000) and Ebert (2001) showed that a small PME can outperform the 51-member ECMWF Ensemble Prediction System since the PME has larger spread that helps identify genuine possibilities of the true future state of the atmosphere. While the PME in this research has little mesoscale information, its performance on the large-scale was examined to explore the benefits of a multimodel multianalysis ensemble.

d. Verification

The SREF systems were analyzed using a variety of parameters—500-mb geopotential height ($Z_{500}$), mean sea level pressure ($MSLP$), 10-m wind speed ($WS_{10}$), and 2-m temperature ($T_2$). Model-based, gridded analyses were used for verification to provide complete coverage over the model domain for generation of a large sample of forecast/observation data pairs—essential for assessing the quality of $FP$. When using gridded analyses as truth, it is important to consider that: 1) model first guesses are used in making such analyses, which can result in an underestimate of error, 2) verification results are dependent on the quality of the analysis (e.g., biases in the analysis can lead to misleading results), and 3) the scales resolved by the analysis
must be compatible with those of the forecast.

Rather than using one of the operational analyses for verification over the 36-km domain, the centroid analysis (mean of all the analyses) was used since it should possess the least amount of error and bias (Richardson 2001). This conclusion was supported by a separate experiment in which the MM5 forecast initialized by the centroid analysis outperformed all other members of UWME. Because of its low resolution, the centroid analysis is not appropriate for verifying the mesoscale forecast of the inner domain. Therefore the independent mesoscale analysis provided by the Rapid Update Cycle 20-km resolution modeling system (RUC20, Benjamin et al., 2004) was used. The RUC20 produces a new analysis every hour using a 3-Dimensional Variational Data Assimilation (3D-Var) scheme that combines a first guess from its 50-level mesoscale model with a large variety of observational assets. The 12-km MM5 forecasts were fit to the RUC20 grid for verification using bilinear interpolation.

3. Impact of Systematic Model Error

Richardson (2001) showed that over a large number of cases, bias correction improves the skill of the ensemble mean and the quality of $FP$ by shifting the ensemble’s forecast PDF toward the PDF of the verification values. Similarly, Atger (2003) focused on improving reliability by applying a spatially and temporally dependent bias correction to the ensemble mean. Our research further explored the need for bias correction, revealing that: 1) it is critically important for SREF since mesoscale models often exhibit larger biases, 2) it is essential for removing misleading ensemble spread, and 3) it increases $FP$ quality by improving both reliability and resolution.

To eliminate the bulk of the systematic error in real-time, a grid-based two-week running mean bias correction was applied to each ensemble member using as truth the centroid analysis
for the 36-km domain and the RUC20 analysis for the 12-km domain. Each forecast cycle had a unique bias correction, as shown in Figure 2, and was verified with independent data (not used in training). This technique was similar to the approach used in Stensrud and Yussouf (2003) in which a 7-day running mean bias correction was applied to individual surface observation locations, rather than individual grid points for a domain wide correction, and was found to be competitive with MOS. The grid-based bias-correction technique used in this research, considers various sources of systematic error:

1) **The model.** When using model diversity in an ensemble, each model generates different biases from its numerics and physical parameterizations.

2) **ICs and LBCs.** Bias for a limited area model is inherited from the bias of the large-scale model used to force it.

3) **Surface characteristics.** Model bias depends on elevation, land use, proximity to water, etc., thus requiring a separate bias correction at every model grid point.

4) **Temporal.** Bias can shift by season or by synoptic regime. In this research, short-term bias variations were accounted for by using the mean bias from recent forecasts. A long training period (e.g., 60 d) provided a stable statistical sampling but only minor forecast improvement while a very short training period produced highly erratic results. A 14-d training period was found to produce a reliable and significant correction that sufficiently captured temporal variability.

5) **Diurnal.** Diurnal variability in bias can be captured by a separate bias correction for each hour of the day, or as in this research, at every forecast lead time since all forecasts were initialized at 00Z. Bias does evolve over the forecast period, so for example, the 12-h (12Z) bias correction is different than the 36-h (also 12Z) bias.
Figure 7 through Figure 9 present bias-correction results of the individual members for 
MSLP and $T_2$. The more pronounced (both consistent and large) the bias in a parameter, the 
more improvement was realized from the bias correction. Compared to the magnitude of the 
bias, bias consistency over the training (not directly displayed in the figures) period is likely the 
more influential aspect on bias correction effectiveness. The PME members generally exhibited 
less improvement than did the UWME members (c.f., Figure 7 and Figure 8) since the large-
scale models are normally better tuned than mesoscale models. On the other hand, the similarity 
in bias and rmse of the PME members and the corresponding UWME members (e.g., PME-cmcg 
and UWME-cmcg) suggests that an important source of the mesoscale model bias is derived 
from the IC and LBCs.

Inaccuracies in the verification values undermines bias correction. A poor forecast 
adjustment can actually verify as an improvement when the same truth data type is used for 
training and verification. To confirm that the grid-based bias correction made a real 
improvement, raw and bias-corrected (by the grid-based method) forecasts of MSLP and $T_2$ were 
verified using ~600 station observations over the Northwest for a one-week period in February 
2003. The percent improvement (not shown) was roughly two-thirds of that shown in Figure 7 
through Figure 9, confirming the validity of the grid-based bias correction.

Several interesting conclusions can be drawn from examining the $T_2$ results (Figure 9). First, 
in contrast to Figure 8, the rmse and bias differences among the UWME members for $T_2$ are 
negligible. For a parameter strongly influenced by the surface such as $T_2$, the forecast error and 
bias results primarily from model deficiencies and not from IC errors (i.e., $T_2$ error is the same no 
matter what IC is applied), a conclusion supported by the weak error growth (discussed further 
below). The disparity between the 12/36-h bias and the 24/48-h bias show the impact of diurnal
effects and reveals basic flaws in the radiation and PBL schemes. Lastly, reduction of average bias is not a complete measure of the value of bias removal. For example, for the ngps member of UWME, 36-h forecast $T_2$ shows negligible average bias before and after correction, but a 16% rmse improvement. This likely indicates that opposing biases, which averaged out over space and time, were successfully captured and corrected.

To examine the effects of model bias, only uncalibrated $FP$ was applied in this research. A post processing calibration technique, which can partially correct for both model bias and poor ensemble dispersion, would have blurred the differences in the raw and bias-corrected results. In this study, $FP$ was calculated using a simplified form of the method in Hamill and Colucci (1997) in which the rank-ordered position of a parameter’s threshold (e.g., $T_2 < 0^\circ C$) is found among the ordered forecasts from the ensemble. Then, $FP$ is calculated by summing up the probability of occurrence in each rank that exceeds the threshold, plus an appropriate fraction of probability from the rank in which the threshold occurred (see Hamill and Colucci 1997, for thorough explanation). Instead of using rank probability from a verification rank histogram, which creates a calibration, this research used uniform probability among the ranks (i.e., probability of occurrence in each rank for an $n$-member ensemble is $1 / n+1$).

The skill of $FP$ from a SREF system is significantly increased by bias correction. Figure 10 and Figure 11 show that bias correction not only improved reliability (presumably by shifting the ensemble’s PDF towards the verification values) but also improved resolution by narrowing the ensemble’s PDF, especially when members had opposing biases. A narrower PDF is more likely to produce $FP$ toward the extreme values (i.e., 0% and 100%) for any given event threshold, which increases resolution. The data in Figure 10 shows that the bulk of resolution improvement of *UWME+ (* indicates bias-correction) over UWME+ came from a 17% increase in the
number of forecasts in the highest $FP$ bin. The roughly 1% decreased weighting in the lower extreme $FP$ by *UWME+ resulted from the shift of the PDF to the right.

Bias correction normally reduces ensemble spread since members often have different biases, both in magnitude and direction w.r.t. the verification value, but are all corrected toward the same location (i.e., the verification value). This decrease in ensemble spread, evinced in Figure 12d, should be expected if spread is a reflection of uncertainty. In effect, removing model bias from an ensemble eliminates bogus uncertainty since systematic errors are not uncertain. Notice in Figure 12d that bias correction notably decreased the UWME+ spread but only slightly decreased the UWME spread. The highly varied biases among UWME+ members created much excess spread (not representative of forecast uncertainty) that was appropriately removed by bias correction.

4. Impact of Model Error

This section discusses how model error and related issues impact SREF dispersion and the skill of ensemble-derived $FP$. To diagnose ensemble underdispersion, two new metrics were introduced: the standardized verification ($V_Z$) and the verification outlier percentage ($VOP$) (see Appendix I). Basically, the $V_Z$ normalizes model error by the ensemble standard deviation and the $VOP$ is the percent occurrence of $V_Z > 3.0$.

a. Impact on Mesoscale

The results shown in Figure 12b–d confirm that failure to account for model error in mesoscale SREF contributes to underdispersion and poor statistical consistency. The concept of statistical consistency, formalized by Talagrand et al. (1999), is that for the verification value to be considered a random sample from the ensemble’s forecast PDF, the mse of the ensemble mean must match the ensemble variance, averaged over many cases. The results in Figure 12
were adjusted for small ensemble size \( n \) by using the unbiased variance (dividing by \( n - 1 \) instead of \( n \)) and by correcting the mse by a factor of \( n / (n + 1) \) (see Appendix II).

Figure 12d shows that not only did *UWME+ provide a needed increase in spread over *UWME, but also improved statistical consistency by simultaneously reducing the mse of the ensemble mean. Figure 12d also shows a pronounced diurnal signal in the \( T_2 \) mse but very little error growth. Stensrud et al. (2000) showed that error growth for surface variables can be very fast initially but Wandishin et al. (2001) found that growth to be relatively small compared to the growth for synoptically driven variables (e.g., 500-mb geopotential height). The small error growth in \( T_2 \) after the first few hours is not due to error saturation since the mse is only about 1/3 of the climatic variance \( (\sigma_c^2) \), but rather due to the model’s inability to represent the variability of \( T_2 \). Even in the cool season, \( T_2 \) variability comes mostly from the diurnal cycle and only partly by the synoptic-scale flow that generates the large growing errors. Notice in Figure 12d that the diurnal signal of the mse correlates strongly with the *UWME+ spread but only weakly with the *UWME spread—another indication that *UWME+ is better at representing forecast uncertainty.

For \( W_{S10} \) (Figure 12c), there is notable error growth because while \( W_{S10} \) is strongly influenced by surface properties and physics, it also depends greatly upon surface pressure, which in turn depends upon the deep atmosphere. The \( W_{S10} \) error growth of *UWME (from only IC diversity) results in superior statistical consistency compared to the \( T_2 \) results (Figure 12d), but the large increase in \( W_{S10} \) spread of *UWME+ over *UWME shows that model uncertainty is also significant.

The importance of including model diversity varies by parameter, as demonstrated in Figure 13. *UWME+ provided only minor improvement for the synoptic-scale parameters \( Z_{500} \) and
MSLP) because forecast error is dominated by error growth from the ICs. Underdispersion is not a serious problem since even without model diversity, ensemble members are able to generate a great deal of spread from the IC differences alone. For surface parameters like $WS_{10}$ and $T_2$ where forecast error depends less on IC uncertainty, underdispersion is a large problem and model diversity makes a significant positive impact. Compared to *UWME, the $WS_{10}$ and $T_2$ verification rank histograms of *UWME+ are more uniform and truth escaped only about half as often, as evidenced by the lower $VOP$.

Figure 11 shows that the additional dispersion provided by the varied-model technique is indeed good spread since both reliability and resolution were improved, indicating that adding model diversity identified valid possible solutions not previously represented. The improvement by model diversity is greatest during the late night to early morning hours (0600-1500 UTC) when the UWME results are limited by the use of a single, deficient representation of the PBL. It is only after bias correction that *UWME+ stands out as superior to *UWME at all lead times, further emphasizing the need for bias correction.

Figure 14 shows relative operating characteristic skill score ($ROCSS$) results for two noteworthy events to demonstrate the potential value of $FP$ derived from mesoscale SREF. The $ROCSS$ is considered to be an upper bound of overall forecast value whereas the $BSS$ is the lower bound (Jolliffe and Stephenson 2003), which is consistent with the higher $ROCSS$ compared to the $BSS$ for $T_2 < 0^\circ$C in Figure 11. The notably higher $ROCSS$ of *UWME+ in Figure 14 confirms the importance of including model diversity for $WS_{10} > 18$ kt. Bias correction did not work as well for $WS_{10}$ as for $T_2$ because the higher error variance associated with $WS_{10}$ forecasts made bias difficult to identify.

Ocean-masked data (i.e., excluding ocean regions) were used in Figure 11 and Figure 14 to
emphasize the greater model uncertainty over land versus water. Over land, the model must represent more variable and complex interactions in the PBL, resulting in larger errors and biases. Figure 15 demonstrates that the model diversity of *UWME+ made a bigger positive impact over land.

It is clear from the $W_{S10}$ and $T_2$ results in Figure 12 and Figure 13 that even after including model diversity in the UW SREF, the problem of underdispersion remained, consistent with the results of Stensrud et al. (2000). The failure of *UWME+ to reach statistical consistency and fully represent the forecast uncertainty indicates the need for additional good spread, which should also increase $FP$ skill. It may be that significant aspects of model error are neglected in the varied-model method—a topic of the next section. There is also a question of how model resolution affects ensemble spread.

Smagorinsky (1969) demonstrated that using higher model resolution increases dispersion as smaller scales of motion are represented. For an ensemble, differences among the members can only exist at the scales represented by the model. Part of the inadequate dispersion of UWME+ may be due to the limited capability of the 12-km members to simulate and reveal different possibilities for smaller scale features; thus, increasing model resolution should generate additional good spread. To test this hypothesis, the ensemble spreads for the common grid points between the 12-km and the 36-km domain solutions were compared. Figure 16 shows that the ensemble spread on the 12-km domain is on average 27% higher than on the 36-km domain. There may be an asymptotic limit to how much more dispersion can be produced by finer model resolution, but significantly higher dispersion could likely be realized by increasing model resolution to a few km. The importance to the overall error growth from the additional small-scale spread could not be explored in this research since 1-way nesting was used. Future
ensemble research comparing the error growth between a 1- and 2-way nested system could address the question of up-scale error growth from the mesoscale.

b. Multimodel vs. Varied-model

The superior performance of the PME displayed in Figure 10 brings into question the merit of running UWME+. However, the relatively low resolution (50-100 km grid spacing) of PME provides limited mesoscale information. Most of the PME members do not even contain output of surface variables, which is why PME was not included in the \( W_{S10} \) and \( T_2 \) results. It may well be that a PME of distinct mesoscale models would also out perform UWME+, but that remains a question for future research.

In this section, PME (a multimodel ensemble) and UWME+ (a varied-model ensemble) are compared using MSLP on the 36-km domain to examine 1) the greater dispersion of the multimodel approach over the varied-model approach, and 2) whether the increased dispersion improves FP quality (i.e., does it add good spread?). In the multimodel approach, each member has a considerably different model attractor, resulting in solutions that may provide unique skillful information to the ensemble (Evans et al. 2000). In the varied-model approach, the members share many model characteristics so the set of model attractors may be too constrained.

Figure 12a and Figure 13a and b show that not only is PME more dispersive than UWME+, but actually slightly overdispersive, indicating that PME members may over-represent model uncertainty. However, the severity of PME’s overdispersion is far less than the large underdispersion of UWME+. It appears that even with the extensive model variations of UWME+, the varied-model approach cannot compete with the multimodel approach for representing model uncertainty. However, there are other disparities between PME and UWME+ besides their approach to model diversity that contribute to the different dispersive
characteristics.

A contributor to the lower spread of UWME+ is the spin up period early in the forecast cycle. As information from the various ICs is likewise adjusted toward the MM5 model attractor, increased similarity among UWME+ members can restrict error growth. The effect can be seen in Figure 12a and b in which the PME members increase in spread from 0 to 12 h but the UWME+ members decrease in spread as the effects of topographic and surface contrast project themselves on all of the UWME+ members. A “hot start” of UWME+ in which each member is allowed to spin up in a data assimilation cycle prior to the forecast period might eliminate the initial drop in spread.

After the spin up period, the error growth by UWME+ is still slower than PME, which may be due to the insufficient model diversity of UWME+ but also may be due to the application of LBCs. Nutter (2003) showed that an ensemble using a limited area model has lower dispersion compared to an ensemble that uses a large model domain because the use of periodically updated, coarse LBCs can filter out short waves as well as reduce the amplitude of nonstationary waves entering from the large-domain model.

Analyzing plots of $V_Z$ revealed that LBC filtering may only be a small contributing factor to the lower dispersion of UWME+ compared to PME, and that lesser model diversity in UWME+ is likely a bigger factor. Figure 17 shows an example case in which a short wave, initialized (not shown) near the west boundary, is near 150°W at the 12 h lead time and subsequently part of a long wave trough along the West Coast at 36 h. The high $V_Z$ values at 36 h along the West Coast in both ensembles likely indicates a deficiency in the ICs. The region in which truth escaped from the *UWME+ members is roughly double is size compared to *PME, perhaps from a loss of information in the *UWME+ members as the short wave entered the outer domain. Now
consider the regions at 12 h on the Oregon coast and British Columbia where truth escaped
greater for *UWME+ compared to *PME. At only 12 h into the forecast cycle and thousands of
kilometers from the lateral boundaries, this large impact is more likely due to the lesser model
diversity of UWME+ than LBC filtering, which suggests that model diversity is the larger of the
two effects.

While LBC filtering may only be a minor effect, it is still a contributor to the insufficient
dispersion of UWME+. A possible technique to boost the synoptic-scale dispersion of UWME+
is to periodically nudge (i.e., four-dimensional data assimilation, FDDA) each member toward
the large-scale model from which it was forced, thus imposing the beneficial large-scale
dispersion of the PME onto the mesoscale SREF. The PME produces a wider range of likely
possibilities (i.e., good spread) on the synoptic scale by modeling large waves over global
domains. Using FDDA in a mesoscale ensemble system would project the additional good
synoptic spread down into the mesoscales where it is needed.

To determine the efficacy of the higher PME spread, the quality of $FP$ for an arbitrary
threshold of $MSLP$ was examined. Because there is significant error growth in $MSLP$, a steady
drop in $BSS$ with increased lead time is seen in Figure 18. The skill of *UWME+ is the same as
that of *UWME two hours before, which may be interpreted as a 2-h improvement in skill by the
varied-model approach. In contrast, the greater model diversity and synoptic error growth of the
*PME resulted in ~18 h improvement over *UWME+.

An interesting question is how can *PME display better resolution than *UWME+ when
*PME also has larger spread? It was previously noted that resolution was improved by bias
correction as reduction in spread narrowed the forecast PDF, thus increasing the weight (i.e.,
number of forecasts) of $FP$ toward the extremes. The resolution term of the $BSS$ is (Wilks 1995):
\[
\frac{1}{M} \sum_{i=1}^{L} N_i (ORF_i - SC)^2
\]  \hspace{1cm} (1)

where \( M \) is the total number of data pairs, \( i \) is the index for the \( I \) bins of \( FP \), \( N_i \) is the number of forecasts within the \( i^{th} \) bin, \( ORF_i \) is the observed relative frequency, and \( SC \) is the sample climatology. Eq. (1) shows that besides increasing the weights toward extreme probabilities, resolution may also be improved by shifting of the observed relative frequency away from the \( SC \) (i.e., zero resolution) and toward the extremes. By either means, an ensemble is then better at discriminating between whether an event will occur or not. Figure 10 shows that the higher resolution of *PME over *UWME+ resulted from the effect of shifting observed relative frequency, which was slightly offset by reduced weighting of the extreme probabilities.

Another interesting question is how can the PME be so skillful when, as shown by Figure 7, it violates the fundamental principle that ensemble members be equally likely? On the subject of using unequally skilled members, Mylne et al. (2002) reasoned that “…the benefits of whichever is the better system at a particular time and place may be obtained all the time through better probabilities.” Perhaps if the relative skill among members varies over time and space, only a member that is consistently inferior adds no value. To test this hypothesis, the member with the highest rmse, tcwb, was removed from *PME, and then the \( BSS \) was recomputed. While a slight reduction in \( BSS \) is expected due to the decrease from 8 to 7 members, Table 1 shows that the \( BSS \) actually increased, indicating that tcwb was harmful to *PME. Removing a member with average skill (ngps) did reduce the ensemble’s skill and removing the best member (ukmo) reduced skill the most. The conclusion is that the higher the deterministic skill of a member, the more value it adds to an ensemble. Lower skilled members can add value to an ensemble but must perform well a significant portion of the time to be useful.
5. Summary and Recommendations

A short-range ensemble forecast (SREF) system was implemented using eight analyses from different operational forecast centers as initial conditions (ICs) for running the MM5 over the Pacific Northwest. Analysis of a large collection of forecast periods indicated the utility of forecast probability ($FP$) from mesoscale SREF, even without a calibration. Comparison of the poor man’s ensemble (PME), the University of Washington Mesoscale Ensemble (UWME), and UWME+ systems yielded insight into ways to further improve SREF.

The relative importance of IC and model uncertainty depends upon the parameter and scale of interest. Accounting for model uncertainty is most important for variables strongly influenced by the land surface, which includes sensible weather parameters (such as surface temperature, relative humidity, and winds) that greatly affect human activity. Thus inclusion of model diversity in SREF is critically important. Although the varied-model technique provides valuable model diversity, the multimodel technique produces greater dispersion that results in more thorough, realistic model uncertainty and higher quality $FP$.

UWME+, a mesoscale SREF system with both IC and model diversity, suffered from low dispersion with clear potential for further improvement. Besides the use of alternative and expanded IC diversity and/or more complete representation of model uncertainty, SREF dispersion and skill may also be increased by:

1) Using Four Dimensional Data Assimilation or statistically consistent lateral boundary conditions (Nutter 2003) to correct for reduced dispersion from use of lateral boundary conditions.

2) Avoiding the spin-up problem of a mesoscale model by adding a data assimilation/pre-forecast period.
3) Using finer model resolution to generate dispersion at the smaller scales. It is also clear that mesoscale SREF would greatly benefit from improvements in mesoscale modeling, thus reducing the need for and challenges of representing model uncertainty. Plotting standardized verification, a new tool for diagnosing an underdispersive SREF system, reveals structures not represented by the ensemble that may be traced to model or analysis deficiencies, or to specific inadequacies of their perturbations. The verification outlier percentage (VOP) should be used in place of the missing rate error (MRE) as a measure of the degree to which truth escapes from an ensemble.

Systematic model errors can severely inhibit the quality of SREF since they contribute to forecast error and appear to be part of the uncertainty. A bias correction applied to the SREF data benefited $FP$ by: 1) improving reliability by adjusting the mean of the ensemble’s PDF to better approximate the mean of the verification’s PDF, and 2) improving resolution by narrowing the ensemble’s PDF where members had opposing biases. Bias correction is imperative for SREF since mesoscale models often exhibit large biases and model diversity (needed for useful SREF) can create varying biases among ensemble members. Bias correction improves skill by narrowing the forecast PDF away from values where the verification is unlikely to occur. Analogously, including realistic model diversity improves skill by widening the forecast PDF toward values where the verification may occur. A question for future research concerns how best to post processing ensemble data to optimize $FP$ skill. This research suggests that perhaps a separate bias correction should be performed on individual members followed by a calibration technique (as in Hamill and Colucci 1997, or Eckel and Walters 1998) to adjust for lack of statistical consistency.

Use of ensemble members that are not equally likely is not problematic but should be done
with care. This is especially true when applying model diversity in which varying skill among members can be more extreme. Poorly skilled members add value to an ensemble only when they perform well a significant portion of the time, so when designing an ensemble, such members should be tested to see if they improve $FP$ when added to the ensemble.

**Acknowledgments.** This work represents a portion of the first author’s Ph.D. research under the supervision of the second author at the University of Washington. This research was supported by United States Air Force as well as the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745. The authors extend special thanks to Eric Grimit for his collaboration in designing the SREF systems, and to Dave Ovens for his computer system help.
Appendix I

Standardized Verification and Verification Outlier Percentage

Since ensemble forecasts frequently display low dispersion, it is desirable to have a metric that conveys the degree to which the true state of the atmosphere “escapes” from the ensemble members. A commonly applied tool, called the missing rate (MR), is derived from a verification rank histogram by totaling the percentage of verification values that occurred in the outer ranks (i.e., rank 1 and rank n+1):

\[ MR = 100 \left( \frac{N_1 + N_{n+1}}{M} \right) \]  

where \( N_x \) is the number of verification values that occurred in a rank \( x \), and \( M \) is the total number of verification values. It is better to consider the missing rate error (MRE) since the statistically consistent value of the MR depends on \( n \):

\[ MRE = 100 \left( \frac{N_1 + N_{n+1}}{M} - \frac{2}{n+1} \right) \]

A larger positive (negative) MRE reveals a more underdispersion (overdispersion) ensemble.

The limitation of MRE is that it only conveys an ensemble’s ability to completely encompass truth and gives no information on when the truth really escapes from the ensemble. With only a finite number of ensemble members, the truth can occur only slightly beyond the smallest or largest member and still be considered to be drawn from the same PDF. To measure whether the verification value is an outlier with respect to the ensemble, the value can be transformed into units of ensemble standard deviation following the statistical calculation of the standard normal random variable (Devore 1995):
\[ V_Z = \frac{V - \bar{e}}{s} \]

where \( V_Z \) is standardized verification, \( \bar{e} \) is the ensemble mean, \( V \) is the verification value, and \( s \) is the ensemble standard deviation, all at a single grid point.

Generally, \(|V_Z| > 3\) is considered an outlier and an indication that the truth escaped from the ensemble. By plotting \( V_Z \) for a single forecast case, it is possible to analyze patterns of how and when truth escapes from an ensemble. Note however that \( V_Z \) is not truly standardized since its expected value is not 1.0 but depends on ensemble size and shape of the PDF. Such factors would have to be considered in using the average \( V_Z \) to check for statistical consistency

An overall measure that can be used to compare ensemble systems is termed the verification outlier percentage (\( VOP \)):

\[
VOP = \frac{100}{M} \sum_{m=1}^{M} \left\{ 0: 3s_m \geq |V_m - \bar{e}_m| \right\} + \left\{ 1: 3s_m < |V_m - \bar{e}_m| \right\}
\]  (5)

or simply the average percentage of the data pairs in which the verification value is considered an outlier with respect to the ensemble. The cutoff of 3.0 for designating an outlier is somewhat arbitrary and there is a small but nonnegligible percentage of expected occurrences beyond three standard deviations. For example, for a normal PDF outliers are expected \(~0.3\%) of the time so the amount that \( VOP \) exceeds 0.3\% is a measure of how often truth escapes from the ensemble. However, for the same reason that \( V_Z \) is not actually a standardized measure, that rule is only a rough guide. The superiority of \( VOP \), compared to the \( MRE \), is that \( VOP \) is a more effective measure of the underdispersiveness of an ensemble. Note that Equations (4) and (5) apply only to normally distributed parameters so for a parameter with a non-normal PDF (e.g., precipitation), a transformation to a normal PDF would be required prior to calculation of \( V_Z \) or \( VOP \).
Appendix II

Statistical Consistency for a Small Ensemble

Talagrand et al. (1999), explained that for an ensemble forecast to exhibit statistical consistency,

\[ \frac{\text{MSE}_e}{\langle \text{Var}_e \rangle} = 1 \quad (6) \]

where \( \langle \text{MSE}_e \rangle \) is the mean square error of the ensemble mean and \( \langle \text{Var}_e \rangle \) is the average ensemble variance, given by:

\[ \text{MSE}_e = \frac{1}{M} \sum_{m=1}^{M} (\bar{e}_m - o_m)^2 \quad (7) \]

\[ \langle \text{Var}_e \rangle = \frac{1}{M} \sum_{m=1}^{M} \left( \frac{1}{n} \sum_{i=1}^{n} (e_{m,i} - \bar{e}_m)^2 \right) \quad (8) \]

where \( M \) is the number of verification values (i.e., forecast/observation data pairs), \( \bar{e}_m \) is the ensemble mean and \( o_m \) is the verifying observation for verification \( m \), \( n \) is the number of ensemble members, and \( e_{m,i} \) is the \( i \)th ensemble member’s value at verification \( m \). Ziehman (2000) explained that Eq. (6) only applies to an ensemble of infinite size but can be altered to account for a finite ensemble by adding an adjustment term:

\[ \frac{\text{MSE}_e}{\langle \text{Var}_e \rangle} = 1 + \frac{2}{n-1} = \frac{n+1}{n-1} \quad (9) \]

To assist in making coherent plots of \( \langle \text{MSE}_e \rangle \) and \( \langle \text{Var}_e \rangle \), the adjustment factor can be broken up as follows:

\[ \frac{1}{n+1} \text{MSE}_e = \frac{1}{n-1} \langle \text{Var}_e \rangle \]
The quantity on the right is known as the sample variance, designed to approximate the population variance by accounting for sample size (Devore 1995). The quantity on the left is the mse of the ensemble mean (adjusted for ensemble size), also known as the theoretical skill of a finite ensemble as described by Leith (1974). For analyzing statistical consistency of an ensemble, Eq. (7) and (8) should be used as:

\begin{equation}
\frac{1}{n+1}MSE = \frac{1}{n-1} \left\{ \frac{1}{n} \sum_{m=1}^{M} \sum_{i=1}^{n} (e_{m,i} - \bar{e}_m)^2 \right\}
\end{equation}

\begin{equation}
\frac{1}{n+1}MSE = \frac{1}{n} \left\{ \frac{1}{M} \sum_{m=1}^{M} \left( \frac{1}{n-1} \sum_{i=1}^{n} (e_{m,i} - \bar{e}_m)^2 \right) \right\}
\end{equation}

\begin{equation}
\frac{n}{n+1}MSE = \frac{1}{M} \sum_{m=1}^{M} \left( \frac{1}{n-1} \sum_{i=1}^{n} (e_{m,i} - \bar{e}_m)^2 \right)
\end{equation}

\begin{equation}
\langle \text{Var}_e \rangle = \frac{1}{M} \sum_{m=1}^{M} \left( \frac{1}{n-1} \sum_{i=1}^{n} (e_{m,i} - \bar{e}_m)^2 \right)
\end{equation}

\begin{equation}
\langle \text{MSE}_e \rangle = \left( \frac{n}{n+1} \right) \frac{1}{M} \sum_{m=1}^{M} (\bar{e}_m - o_m)^2
\end{equation}
REFERENCES


Stensrud, D. J., and N., Yussouf, 2003: Short-range predictions of 2-m temperature and


-----, J. Bao, and T. T. Warner, 2000: Using initial conditions and model physics perturbations in
short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, 128,
2077–2107.

*Proc. from the Workshop on Predictability*, Reading, UK, European Center for Medium-


Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National

Tribbia, J. J. and D. P. Baumhefner, 2004: Scale interactions and atmospheric predictability: An


**FIGURE CAPTIONS**

Figure 1. Grid domains (Lambert conformal projections) of the UW SREF systems.  (a) 151×127 domain with 36-km grid spacing.  (b) 103×100 nested domain with 12-km grid spacing.

Figure 2. The 129 forecast case days of the research dataset over the 2002-2003 cool season.  A 48-h forecast cycle was initialized at 0000 UTC on each date.  Darkly shaded dates contain incomplete data, so were dropped from the dataset.  The lightly shaded 2-week period illustrates the training period used to compute a bias correction for the indicated forecast period.  The bias-corrected dataset consists of 112 cases, which are the complete cases beginning 25 Nov 2002.

Figure 3. The eight analysis/forecast modeling systems of the PME, which also provide ICs for the UWME and UWME+.  (T, triangular spectral truncation wave number;  L, vertical levels; 3D-Var, 3-Dimensional Variational Data Assimilation;  SSI, Spectral Statistical Interpolation;  OI, Optimal Interpolation)

Figure 4. MM5 model versions of UWME+ with the control MM5 version used in UWME shown in the first row.  Descriptions of the various MM5 schemes (e.g., Eta PBL, Goddard microphysics, and cloud radiation) can be found in Grell et al. (1994).  The unique land use tables for the UWME+ members include random perturbations to moisture availability, albedo, and roughness length.

Figure 5. Plot of the randomly generated SST perturbations over the outer domain for member plus01 of UWME+ (i.e., pert.01).  Isopleths are positive (solid) and negative (dashed)
perturbation values in °C. Over each forecast cycle the SST analysis field was different (as provided by the Optimum Thermal Interpolation System) and the eight perturbation fields for the UWME+ members remained fixed. The perturbation fields for the inner domain matched the outer domain perturbations.

Figure 6. Sample surface parameter PDFs. The solid curve is for summer and the dashed curve is for winter. The peaks of the curves correspond to the standard MM5 land use table values. The width of a PDF is proportional to the suspected uncertainty in the parameter.

Figure 7. Results of MSLP bias correction for each PME member and the ensemble mean averaged over the bias-corrected dataset, using the outer, 36-km domain. (a) and (b) show rmse (in non-filled bars) and bias (in shaded bars by forecast lead time) before and after bias correction, and the percent improvement (also shaded) in rmse is given in (c).

Figure 8. As in Figure 7 but for UWME. Note that UWME ensemble member names refer to the IC and LBC source of each MM5 model run (e.g., jma was run using the jma IC and LBCs).

Figure 9. As in Figure 8 but for UWME $T_2$ data from the inner, 12-km domain, fit to the RUC20 20-km analysis grid.

Figure 10. Reliability diagram and components of BSS calculation for 36-h FP of the event MSLP < 1001.0 mb. All three SREF systems have the same number of forecasts (# of Fcsts), number occurrences (# of Occ.), and sample climatology of 0.2273 since they covered the same
set of forecast cases. Shaded values of \textit{res}, \textit{rel}, and \textit{BSS} indicate an improvement over the results of the SREF system to the left in the table. The improvement by bias correction is seen by comparing UWME and *UWME+ (* indicates bias-correction). The advantage of a multimodel system over a varied-model system is seen by comparing *PME and *UWME+.

Figure 11. \textit{BSS} and its components for \textit{FP} of $T_2 < 0^\circ\text{C}$ using ocean-masked data. Note that the ordinate axis of the reliability plot is reversed so that upward is better on all plots. Error bars are omitted since they are so small (roughly $\pm 0.01$ for \textit{BSS}) due to the large dataset. For completeness, the uncertainty term of the \textit{BSS} is included in the \textit{BSS} plot.

Figure 12. Ensemble variance (open points) for raw and bias-corrected (*) forecasts, and mse of the ensemble mean (filled points) for (a) \textit{MSLP} by *PME, (b) \textit{MSLP} by *UWME and *UWME+, (c) \textit{WS}_{10}, and (d) $T_2$. In panel (b) the mse of the *UWME+ mean is nearly identical to that of the *UWME mean so only one is plotted. The ? marks in (a) and (b) denote a likely underestimate of mse since the verification (centroid analysis) contains much forecast information at the 12-h lead time. Panel (d) includes ensemble variance results before and after bias correction.

Figure 13. Verification rank histograms that show the impact of including model diversity for bias-corrected, 36-h forecasts of (a) $Z_{500}$ and (b) \textit{MSLP} on the 36-km domain, and (c) \textit{WS}_{10} and (d) $T_2$ on the 12-km domain. Dashed-outlined histograms are for the *PME (only available in (a) and (b)), thick-outlined histograms are for *UWME, and shaded histograms are for the *UWME+. The variables are ordered from most synoptic ($Z_{500}$) to most surface-influenced ($T_2$) to show that model diversity plays a larger role for surface variables.
Figure 14. ROCSS for FP for the event (a) $WS_{10} > 18$ kt, and (b) $T_2 < 0^\circ$C, using ocean-masked data.

Figure 15. Raw (solid) and percent (dashed) BSS improvement by *UWME+ over *UWME for FP of the event $T_2 < 0^\circ$C using the full 12-km domain (curves with no marks) and the 12-km domain with ocean grid points omitted (curves with o’s).

Figure 16. Comparison of UWME+ spread for $WS_{10}$ from the 12-km and 36-km domains using only shared grid points.

Figure 17. $Z_{500}$ ensemble mean and $V_Z$ for *UWME+ and *PME forecasts initialized 20 Dec 2002 at 00Z, along with the corresponding centroid analyses. The inset VOP values are specific to this forecast case rather than averaged over all cases. $V_Z < -3$ is lightly shaded and $V_Z > 3$ is darkly shaded to show regions where the verification value is an outlier w.r.t. the ensemble.

Figure 18. BSS and its components for FP of $MSLP < 1011$ mb using ocean-masked data

**TABLE CAPTIONS**

Table 1. Comparison of BSS results for the event $MSLP < 1001$ mb for *PME when different members are omitted (i.e. x-tcwb is the 7-member version of *PME with the tcwb member omitted). While the results do not test as statistically significant, they do suggest that *PME without tcwb performs better and that without ngps or ukmo, *PME performs worse.
Figure 1. Grid domains (Lambert conformal projections) of the UW SREF systems. (a) 151×127 domain with 36-km grid spacing. (b) 103×100 nested domain with 12-km grid spacing.

Figure 2. The 129 forecast case days of the research dataset over the 2002-2003 cool season. A 48-h forecast cycle was initialized at 0000 UTC on each date. Darkly shaded dates contain incomplete data, so were dropped from the dataset. The lightly shaded 2-week period illustrates the training period used to compute a bias correction for the indicated forecast period. The bias-corrected dataset consists of 112 cases, which are the complete cases beginning 25 Nov 2002.
<table>
<thead>
<tr>
<th>Abbreviation, Model, Source</th>
<th>Type</th>
<th>Resolution (~ @ 45° N)</th>
<th>Objective Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>avn.</strong> Global Forecast System (GFS), National Centers for Environmental Prediction</td>
<td>Spectral</td>
<td>T254 / L64 ~55 km / 1.0° / L14 ~80 km</td>
<td>3D-Var (SSI)</td>
</tr>
<tr>
<td><strong>cmcg.</strong> Global Environmental Multi-scale (GEM), Canadian Meteorological Centre</td>
<td>Finite Diff.</td>
<td>0.9°×0.9° / L28 ~70 km / 1.25° / L11 ~100 km</td>
<td>3D-Var</td>
</tr>
<tr>
<td><strong>eta.</strong> limited-area mesoscale model, National Centers for Environmental Prediction</td>
<td>Finite Diff.</td>
<td>32 km / L45 / 90 km / L37</td>
<td>3D-Var (SSI)</td>
</tr>
<tr>
<td><strong>gasp.</strong> Global Analysis and Prediction model, Australian Bureau of Meteorology</td>
<td>Spectral</td>
<td>T239 / L29 ~60 km / 1.0° / L11 ~80 km</td>
<td>3D-Var</td>
</tr>
<tr>
<td><strong>jma.</strong> Global Spectral Model (GSM), Japan Meteorological Agency</td>
<td>Spectral</td>
<td>T106 / L21 ~135 km / 1.25° / L13 ~100 km</td>
<td>OI</td>
</tr>
<tr>
<td><strong>ngps.</strong> Navy Operational Global Atmos. Pred. System, Fleet Numerical Meteorological &amp; Oceanographic Cntr.</td>
<td>Spectral</td>
<td>T239 / L30 ~60 km / 1.0° / L14 ~80 km</td>
<td>OI</td>
</tr>
<tr>
<td><strong>tcwb.</strong> Global Forecast System, Taiwan Central Weather Bureau</td>
<td>Spectral</td>
<td>T79 / L18 ~180 km / 1.0° / L11 ~80 km</td>
<td>OI</td>
</tr>
<tr>
<td><strong>ukmo.</strong> Unified Model, United Kingdom Meteorological Office</td>
<td>Finite Diff.</td>
<td>5/6°×5/9° / L30 ~60 km / same / L12</td>
<td>3D-Var</td>
</tr>
</tbody>
</table>

Figure 3. The eight analysis/forecast modeling systems of the PME, which also provide ICs for the UWME and UWME+. (T, triangular spectral truncation wave number; L, vertical levels; 3D-Var, 3-Dimensional Variational Data Assimilation; SSI, Spectral Statistical Interpolation; OI, Optimal Interpolation)
Figure 4. MM5 model versions of UWME+ with the control MM5 version used in UWME shown in the first row. Descriptions of the various MM5 schemes (e.g., Eta PBL, Goddard microphysics, and cloud radiation) can be found in Grell et al. (1994). The unique land use tables for the UWME+ members include random perturbations to moisture availability, albedo, and roughness length.

<table>
<thead>
<tr>
<th>IC</th>
<th>ID#</th>
<th>Soil</th>
<th>vertical diffusion</th>
<th>Cloud Microphysics</th>
<th>Cumulus 36-km Domain</th>
<th>12-km Domain</th>
<th>shallow cumulus</th>
<th>Radiation</th>
<th>SST Perturbation</th>
<th>Land Use Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>UWME</td>
<td>MRF</td>
<td>5-Layer</td>
<td>Y</td>
<td>Simple Ice</td>
<td>Kain-Fritsch</td>
<td>Kain-Fritsch</td>
<td>N</td>
<td>cloud</td>
<td>none</td>
<td>standard</td>
</tr>
<tr>
<td>UWME+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avn</td>
<td>plus01</td>
<td>MRF</td>
<td>LSM</td>
<td>Y</td>
<td>Simple Ice</td>
<td>Kain-Fritsch</td>
<td>Kain-Fritsch</td>
<td>Y</td>
<td>RRTM pert.01</td>
<td>tbl.01</td>
</tr>
<tr>
<td>cmcg</td>
<td>plus02</td>
<td>MRF</td>
<td>5-Layer</td>
<td>Y</td>
<td>Reisner II</td>
<td>Grell</td>
<td>Grell</td>
<td>N</td>
<td>cloud pert.02</td>
<td>tbl.02</td>
</tr>
<tr>
<td>eta</td>
<td>plus03</td>
<td>Eta</td>
<td>5-Layer</td>
<td>N</td>
<td>Goddard</td>
<td>Betts-Miller</td>
<td>Grell</td>
<td>Y</td>
<td>RRTM pert.03</td>
<td>tbl.03</td>
</tr>
<tr>
<td>gasp</td>
<td>plus04</td>
<td>MRF</td>
<td>LSM</td>
<td>Y</td>
<td>Shultz</td>
<td>Betts-Miller</td>
<td>Kain-Fritsch</td>
<td>N</td>
<td>RRTM pert.04</td>
<td>tbl.04</td>
</tr>
<tr>
<td>jma</td>
<td>plus05</td>
<td>Eta</td>
<td>LSM</td>
<td>N</td>
<td>Reisner II</td>
<td>Kain-Fritsch</td>
<td>Kain-Fritsch</td>
<td>Y</td>
<td>cloud pert.05</td>
<td>tbl.05</td>
</tr>
<tr>
<td>ngps</td>
<td>plus06</td>
<td>Blackadar</td>
<td>5-Layer</td>
<td>Y</td>
<td>Shultz</td>
<td>Grell</td>
<td>Grell</td>
<td>N</td>
<td>RRTM pert.06</td>
<td>tbl.06</td>
</tr>
<tr>
<td>tcwb</td>
<td>plus07</td>
<td>Blackadar</td>
<td>5-Layer</td>
<td>Y</td>
<td>Goddard</td>
<td>Betts-Miller</td>
<td>Grell</td>
<td>Y</td>
<td>cloud pert.07</td>
<td>tbl.07</td>
</tr>
<tr>
<td>ukmo</td>
<td>plus08</td>
<td>Eta</td>
<td>LSM</td>
<td>N</td>
<td>Reisner I</td>
<td>Kain-Fritsch</td>
<td>Kain-Fritsch</td>
<td>N</td>
<td>cloud pert.08</td>
<td>tbl.08</td>
</tr>
</tbody>
</table>

Figure 5. Plot of the randomly generated SST perturbations over the outer domain for member plus01 of UWME+ (i.e., pert.01). Isopleths are positive (solid) and negative (dashed) perturbation values in °C. Over each forecast cycle the SST analysis field was different (as provided by the Optimum Thermal Interpolation System) and the eight perturbation fields for the UWME+ members remained fixed. The perturbation fields for the inner domain matched the outer domain perturbations.
Figure 6. Sample surface parameter PDFs. The solid curve is for summer and the dashed curve is for winter. The peaks of the curves correspond to the standard MM5 land use table values. The width of a PDF is proportional to the suspected uncertainty in the parameter.
Figure 7. Results of MSLP bias correction for each PME member and the ensemble mean averaged over the bias-corrected dataset, using the outer, 36-km domain. (a) and (b) show rmse (in non-filled bars) and bias (in shaded bars by forecast lead time) before and after bias correction, and the percent improvement (also shaded) in rmse is given in (c).
Figure 8. As in Figure 7 but for UWME. Note that UWME ensemble member names refer to the IC and LBC source of each MM5 model run (e.g., jma was run using the jma IC and LBCs).
Figure 9. As in Figure 8 but for UWME $T_2$ data from the inner, 12-km domain, fit to the RUC20 20-km analysis grid.
Figure 10. Reliability diagram and components of BSS calculation for 36-h FP of the event MSLP < 1001.0 mb. All three SREF systems have the same number of forecasts (# of Fcsts), number occurrences (# of Occ.), and sample climatology of 0.2273 since they covered the same set of forecast cases. Shaded values of res, rel, and BSS indicate an improvement over the results of the SREF system to the left in the table. The improvement by bias correction is seen by comparing UWME+ and *UWME+ (* indicates bias-correction). The advantage of a multimodel system over a varied-model system is seen by comparing *PME and *UWME+.
Figure 11. \( BSS \) and its components for \( FP \) of \( T_2 < 0^\circ C \) using ocean-masked data. Note that the ordinate axis of the reliability plot is reversed so that upward is better on all plots. Error bars are omitted since they are so small (roughly \( \pm 0.01 \) for \( BSS \)) due to the large dataset. For completeness, the uncertainty term of the \( BSS \) is included in the \( BSS \) plot.
Figure 12. Ensemble variance (open points) for raw and bias-corrected (*) forecasts, and mse of the ensemble mean (filled points) for (a) MSLP by *PME, (b) MSLP by *UWME and *UWME+, (c) WS$_{10}$, and (d) $T_2$. In panel (b) the mse of the *UWME+ mean is nearly identical to that of the *UWME mean so only one is plotted. The ? marks in (a) and (b) denote a likely underestimate of mse since the verification (centroid analysis) contains much forecast information at the 12-h lead time. Panel (d) includes ensemble variance results before and after bias correction.
Figure 13. Verification rank histograms that show the impact of including model diversity for bias-corrected, 36-h forecasts of (a) $Z_{500}$ and (b) MSLP on the 36-km domain, and (c) WS$_{10}$ and (d) $T_2$ on the 12-km domain. Dashed-outlined histograms are for the *PME (only available in (a) and (b)), thick-outlined histograms are for *UWME, and shaded histograms are for the *UWME+. The variables are ordered from most synoptic ($Z_{500}$) to most surface-influenced ($T_2$) to show that model diversity plays a larger role for surface variables.
Figure 14. ROCSS for FP for the event (a) $WS_{10} > 18$ kt, and (b) $T_2 < 0^\circ C$, using ocean-masked data.
Figure 15. Raw (solid) and percent (dashed) BSS improvement by *UWME+ over *UWME for FP of the event $T_2 < 0$°C using the full 12-km domain (curves with no marks) and the 12-km domain with ocean grid points omitted (curves with o’s).

Figure 16. Comparison of UWME+ spread for $WS_{10}$ from the 12-km and 36-km domains using only shared grid points.
Figure 17. $Z_{500}$ ensemble mean and $V_Z$ for *UWME+ and *PME forecasts initialized 20 Dec 2002 at 00Z, along with the corresponding centroid analyses. The inset VOP values are specific to this forecast case rather than averaged over all cases. $V_Z < -3$ is lightly shaded and $V_Z > 3$ is darkly shaded to show regions where the verification value is an outlier w.r.t. the ensemble.
Figure 18. BSS and its components for FP of MSLP < 1011 mb using ocean-masked data.
Table 1. Comparison of BSS results for the event $\text{MSLP} < 1001 \text{ mb}$ for *PME when different members are omitted (i.e. x-tcwb is the 7-member version of *PME with the tcwb member omitted). While the results do not test as statistically significant, they do suggest that *PME without tcwb performs better and that without ngps or ukmo, *PME performs worse.

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>12 h</th>
<th>24 h</th>
<th>36 h</th>
<th>48 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>*PME (x-tcwb)</td>
<td>0.942</td>
<td>0.905</td>
<td>0.868</td>
<td>0.836</td>
</tr>
<tr>
<td>*PME (full)</td>
<td>0.940</td>
<td>0.902</td>
<td>0.866</td>
<td>0.834</td>
</tr>
<tr>
<td>*PME (x-ngps)</td>
<td>0.938</td>
<td>0.899</td>
<td>0.861</td>
<td>0.830</td>
</tr>
<tr>
<td>*PME (x-ukmo)</td>
<td>0.935</td>
<td>0.896</td>
<td>0.858</td>
<td>0.825</td>
</tr>
</tbody>
</table>