11A.4

EVALUATION OF A MESOSCALE SHORT-RANGE ENSEMBLE FORECAST SYSTEM OVER THE NORTHEAST UNITED STATES

Brian A. Colle*[1], Matthew S. Jones[1], and Jeffrey S. Tongue[2]

[1] Institute for Terrestrial and Planetary Atmospheres
Stony Brook University, Stony Brook, New York
[2] NOAA/National Weather Service, Upton, NY

1. INTRODUCTION

Significant model errors can develop for relatively short-range predictions (0-48h forecasts), such as January 2000 "surprise" East Coast snowstorm (Zhang et al. 2002) and the major numerical forecast errors over the Northeast Pacific (McMurdie and Mass 2004). These errors in numerical weather prediction (NWP) result from uncertainty in initial conditions (ICs) and imperfect physical (PHYS) parameterizations. As a result, several recent studies have explored the benefits and shortcomings of short-range ensemble forecast (SREF) modeling systems. Developers of these SREF systems have quantified the impact of initial condition uncertainty, model dynamics diversity, and model physics variability on short-term forecasts.

Most SREF studies have focused over the Pacific Northwest or the central U.S., while there have been few long-term SREF verification studies over the Northeast U.S. Stensrud and Yussouf (2003) and Yussouf et al. (2004) focused on summer temperature prediction over the Northeast, but other low-level parameters also need to be evaluated in the Northeast, such as 10-meter wind and precipitation. The Northeast U.S. weather also poses different challenges than other regions where SREF systems have been verified. The Great Lakes, Appalachian Mountains, urban centers, irregular coastline, Gulf Stream and Labrador currents all add mesoscale complexity and result in model errors that vary significantly from season to season (Colle et al. 2003a, 2003b). Thus, a SREF system over this region requires evaluation for both the warm and cool seasons in order to qualify the relative importance of IC and PHYS uncertainty.

This paper summarizes the verification of a SREF ensemble forecast system that was developed at Stony Brook University (SBU) over the Northeast U.S. in collaboration with several of NOAA's National Weather Service (NWS) forecast offices as part of a COMET (Collaborative Program for Operational Meteorology, Education, and Training) collaborative project. The 18-member SREF system utilizes both IC and physics (PHYS) uncertainty in the MM5 at 12-km grid spacing. At the time of this research, this was the highest resolution SREF ensemble over the Northeast.

2. ENSEMBLE AND VERIFICATION SETUP

A mesoscale SREF system was constructed using 18 members of the MM5 (version 3.6). The MM5 was integrated over an outer 36-km domain that extended from the Rocky Mountains to the western Atlantic Ocean and a 12-km (one-way) nested grid that covered much of Northeast U.S. (Fig. 1). Thirty-three sigma levels were used in the vertical, with a maximum resolution in the boundary layer. The terrain for the 36-/12-km grids was analyzed using a 5' and 30" terrain dataset, while a 30" land use dataset from NCAR was used to initialize 25 land surface categories.

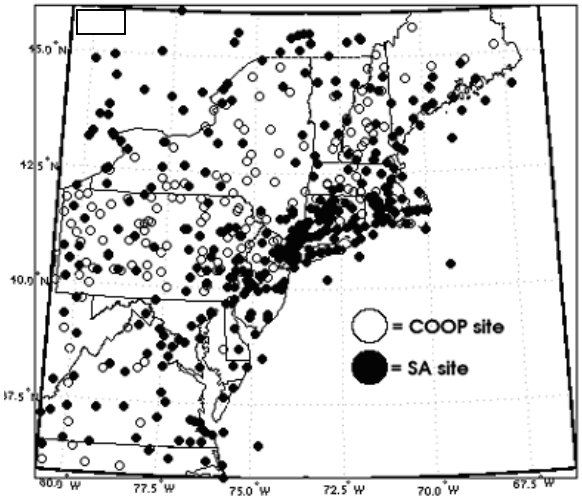

Figure 1. Location of the 12-km MM5 domain, which is nested within a larger 36-km domain.. The surface and C-MAN (SA) and cooperative observation (COOP) sites used for the verification are plotted in using black and white circles, respectively. The 2-meter temperature, 10-meter wind speed and direction, and sea-level pressure verification statistics use the SA sites, whereas the 24-hour precipitation verification includes both SA and COOP sites.

Twelve PHYS members were initialized using the 12-km Eta interpolated to the NCEP-221 grid (32-km grid spacing, 25-mb interval vertical levels). These analyses were interpolated bilinearly to the MM5, while boundary conditions were obtained by linearly interpolating the 3-h Eta-104 model forecasts (90-km grid spacing, 25-mb vertical levels). The 12 PHYS members were chosen using a combination of three MM5 PBL schemes: Blackadar, Miller-Yamada-Janjic (Eta-PBL), and MRF, as well as four MM5 CPs: Betts-Miller (BM), Grell (GR), Kain-Fritsch (KF), and Kain-Fritsch-2 (KF2).

Five IC members were initialized at 0000 UTC using the 3-h forecast from the NCEP SREF Eta bred members at 2100 UTC. Boundary conditions for these members were obtained by linearly interpolating the bred Eta-104 model forecast grids

*Corresponding author address: Dr. Brian A. Colle, MSRC, Stony Brook University / SUNY, Stony Brook, NY 11794-5000.
email: brian.colle@stonybrook.edu

to the MM5 grid at 3-h intervals. A sixth IC member was initialized using the NCEP Global Forecast System (GFS) model initialized at 0000 UTC at one-degree resolution, with boundary conditions from the GFS at 6-h intervals. The physics and IC ensembles were not mixed (different IC use different physics) in order to better understand the different physics impact on the simulations.

The ensemble system was run once daily at 0000 UTC, and the ensemble mean and spread data was transferred in real-time to the Advanced Weather Interactive Processing System (AWIPS) at several regional NWS offices and the Northeast River Forecast Center. Real-time data archived during the warm season of 01 May - 31 September 2003 (warm season) and during the cool season of 01 October 2003 – 31 March 2004 (cool season) was examined for this study.

Verification results were compiled over the warm and cool seasons using the MM5 and Eta verification system that has been operational at SBU since 1999 (Colle et al. 2003a, 2003b). This paper primarly focuses on the surface verification of temperature, wind, and precipitation for each member in order to illustrate the challenges one faces in constructing an ensemble over this region. For each surface parameter standard measures of forecast skill were calculated for each ensemble member and ensemble-mean forecast, such as mean error (ME) and mean absolute error (MAE). These errors were averaged over the 12-km model domain. Verification rank histograms were constructed for lower-level metrics to examine the dispersion qualities of the ensemble.

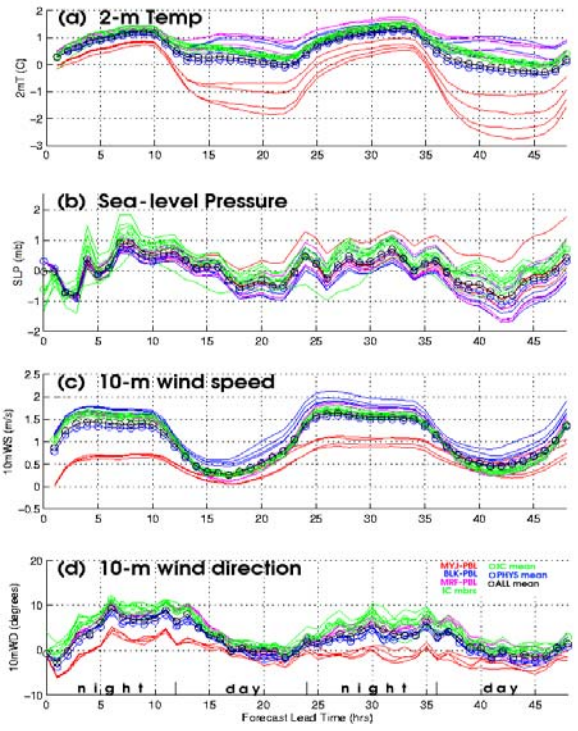

Figure 2. Diurnal mean errors (MEs) every 1-h for (a) 2-meter temperature, (b) sea-level pressure, (c) 10-meter wind speed, and (d) 10-meter wind direction for the warm season 0000 UTC forecasts for all members and 3 ensemble means (PHYS, ICs, ALL) averaged for the 12-km domain.

3. VERIFICATION RESULTS

Figure 2 shows the 12-km domain-averaged MEs and MAEs for the warm season for all 18 members and the three ensemble-means (PHYS, IC, and ALL). All members have a warm (0.5-1.0 ^{o}C) 2-m temperature bias at night (Fig. 2a), while the MYJ-PBL (Eta) members develop a 1-2 ^{o}C cool bias during the day. The MYJ-PBL cool bias partially offsets the warm biases prevalent in the BLK and MRF PBL members, producing an overall lower bias during the day for both the PHYS and ALL means as compared to the IC mean. As a result, the PHYS mean has a greater 2-m temperature skill than the IC mean across all forecast lead times (not shown), while the full (ALL) ensemble is more skillful than the PHYS during the day.

The strong diurnal biases are likely due to imperfect land-surface and boundary layer physics. For example, the MYJ-PBL cool bias during the day is also associated with a large (20-30%) moist bias at the surface (Jones 2004). Also, too much mixing at night favors a near surface warm bias as well as the nocturnal high wind speed bias (0.5–1.75 m s^{-1}) and positive (clockwise or too geostrophic) wind direction bias (5-10^{o}) in all members (Figs. 2c,d), with the MYJ PBL having a smaller bias than the other PBL members. For 10-m wind speed (not shown), the ensemble means are among the best performing members only during the day, since some of the large errors in the MRF and BLK PBLs at night result in the MYJ-PBL outperforming the ensemble means on average. The sea-level pressures tend to have a weak (~0.5 mb) negative bias during the day for most members (Fig. 2b). For sea-level pressure and wind direction (Figs. 2b,d), the errors gradually increase through the 0-48 h period, with the Eta-bred members having the largest error for most time periods. In fact, some of the Eta-bred sea-level pressure errors during the first 12 h are as large as the ensemble ALL mean at hour 48.

Using the MYJ-PBL with different CPs results in relatively large spread among members during the day. This variation in low-level temperature among the MYJ-PBL members is related to those days that have large amounts of precipitation over the Northeast U.S (not shown). It was found that the amount of low-level clouds produced by a given explicit precipitation and convective scheme combination reduces the incoming solar radiation and results in a particular surface cool bias.

During the cool season (not shown), the MEs of 2-m temperature for the PHYS members are clustered together more than the warm season, with members grouping according to PBL scheme. Cool biases are more prevalent during the day than in the warm season (not shown), and a moist bias exists for all members during the day.

The diurnal pattern of temperature MAEs for the cool season (Fig. 3a) is similar to the warm season (not shown); however, as compared to the warm season, the cool-season PHYS and ALL means have a smaller skill advantage over the individual members. The sea-level pressure MAEs have the largest spread of all the parameters, with the Eta-bred members having 20-50% larger errors than the other members (Fig. 3b), while the GFS member is the most skillful member. Clearly, all members are not equally skillful in sea-level pressure, and because of the large Eta-bred errors, the ensemble means do not improve upon the GFS. The 10-m

wind speed MAEs are clustered according to the above noted bias errors (Fig. 3c), with the MYJ-PBL being the more skillful than even the ensemble means. The Eta-bred members are also the worst set of members for 10-m wind direction on average (Fig. 3d), with the PHYS and ALL mean being the best member on average for this parameter.

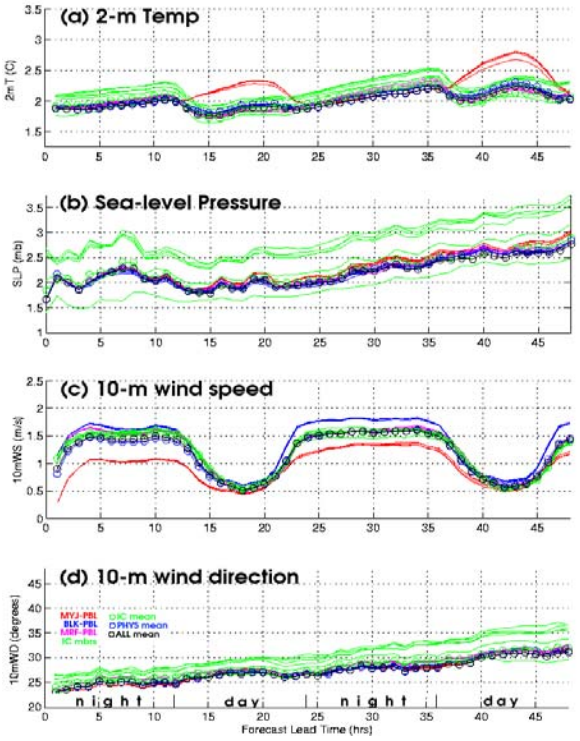

Figure 3. Diurnal mean absolute errors (MAEs) every 1-h for (a) 2-meter temperature, (b) sea-level pressure, (c) 10-meter wind speed, and (d) 10-meter wind direction for the warm season 0000 UTC forecasts for all members and 3 ensemble means averaged for the 12-km domain.

Because of the model wind speed and temperature biases, all parameters show an overpopulation of the extreme ranks of the histograms (i.e., the histograms are U-shaped or L-shaped) during the night (0-12 h and 25-36 h) and day (13-24 h and 37-48 h) periods (not shown). The pronounced L-shape histogram during the night indicates a positive bias, and this feature is most prevalent in the 2-m temperature and 10-m wind speed distributions. A 14-day bias calibration improves the dispersion of the ensemble forecast by reducing the frequency of misses due to an ensemble-wide bias (not shown). However, even after calibration, the ensemble remains under-dispersed. This highlights the need to improve the overall dispersion of the raw ensemble as well as to develop better bias calibration techniques.

The ability of the 18-member ensemble to predict the skill of the mean of all 18-member (ALL) ensemble forecasts for the warm and cool seasons was evaluated. Each point in the scatterplot on Fig. 4 represents the 12-km domain-averaged ALL ensemble variance versus ALL mean MAE averaged for a given diurnal period for both the raw ensemble and using the 14-day bias calibration. For the warm season 2-m temperature, sea-level pressure, and surface wind speed (Figs. 4a-f), the MAEs differ greatly compared to its variance, producing a "column" pattern in the scatterplot. In other words, as a result of the model biases, a wide range of errors is associated with little variance between members. This results in spread-error correlations that are relatively poor, with correlation coefficients between 0.20-0.40 (Figs. 4i,j). The 2-m temperature error-variance patterns vary less for the cool season forecast periods than the warm season (not shown), resulting in a more pronounced column pattern than the cool season and correlation coefficients only ranging from 0.07 to 0.09. Unfortunately, the correlation results do not change substantially after the calibration is applied.

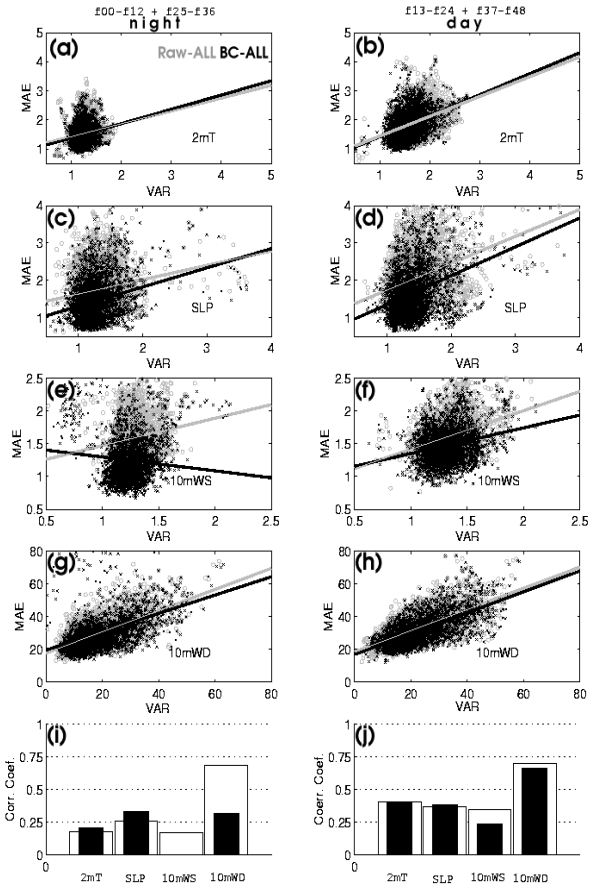

Figure 4. Scatterplots of domain-average variance (abscissa) versus domain-average MAE (ordinate) for before (gray circles) and after (black x's) a 14-day bias calibration is applied. Scatterplots for night and day averaged during the warm season are (a,b) 2-meter temperature, (c,d) sea-level pressure, (e,f) 10-meter wind speed, and (g,h) 10-meter wind direction forecasts for the 18-member ensemble mean. The MAE-variance correlation coefficients (i,j) are shown for (white) before and after (black) the bias calibration is applied. Plotted lines in panels (a) to (h) represent the least-squares linear fit of (gray) raw and (black) bias-calibrated scatterplots.

Figures 4g,h show the 10-m wind direction error-variance patterns for the warm season during the night and day, respectively. In general, the errors tend to increase with increasing variance, producing a "fan" pattern, with correlation coefficients much higher (0.68 to 0.70) than other variables. The cool season 10-m wind direction forecasts tend to have less variance than the warm season (not shown), with

correlation coefficients ranging from 0.60 to 0.64. The 10-m wind direction correlations are changed only slightly during the day by applying bias calibration, while there is a large (50%) reduction at night. Overall, surface wind direction forecast skill is more predictable in the warm season than in the cool season over the Northeast U.S.

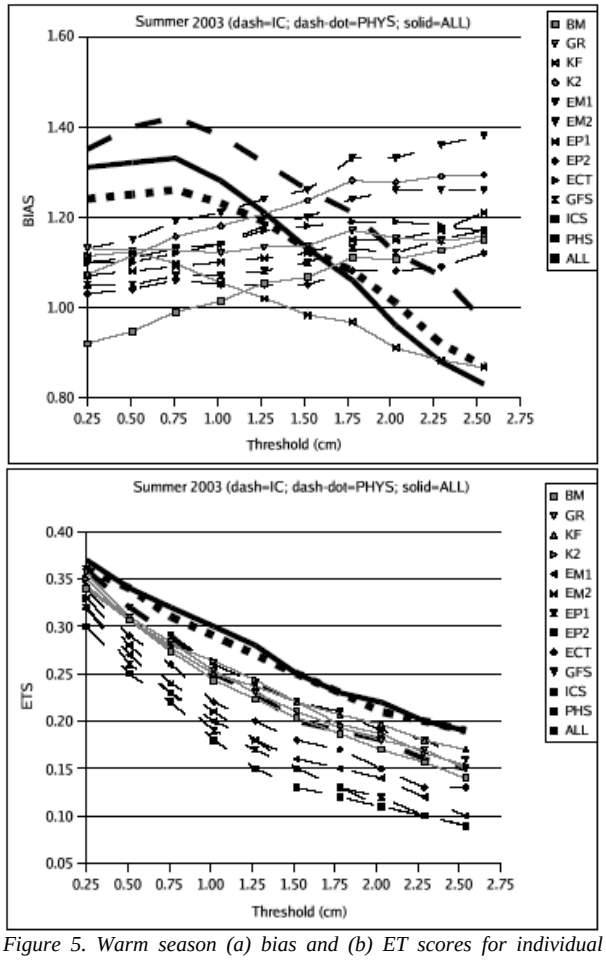

Figure 5. Warm season (a) bias and (b) ET scores for individual ensemble members and ensemble means versus 24-hour QPF event threshold.

The 24-h (12-36h) quantitative precipitation forecasts (QPF) were verified during the warm and cool seasons for the 12-km domain. Conventional bias and equitable threat scores were calculated for each member and ensemble means. Figure 18a shows the precipitation bias (mean error) based on the contingency table for the IC, PHYS, and ALL ensemble means as well as select groupings based on convective parameterization. Although there was some variation of precipitation based on the PBL used (not shown), it was found that the CP scheme had the larger impact. Those 24-h precipitation thresholds greater than 2.54 cm are not shown since the number of events were relatively small. All individual members except the Kain-Fritsch CP have increasing bias with increasing threshold amount. Interestingly, the Kain-Fritsch CP member has a bias of 0.87 for the 2.54 cm threshold, while the Kain-Fritsch2 is near 1.30. This change in Kain-Fritsch performance from under-prediction to over-prediction at higher event thresholds may be the result of implementing a minimum moisture entrainment rate and other modifications into the updated scheme. The warm season bias scores for the ensemble means illustrate one of the disadvantages of using an averaged value for warm-season precipitation. Namely, the intrinsic smoothing created by averaging individual members forecasts leads to over-prediction of low- and mid-thresholds, and under-prediction of high-thresholds.

The PHYS CP members have better (higher) equitable threat scores (ETSs) than the IC members for all thresholds (Fig. 18b). The GFS performs slightly better than the IC mean for all thresholds over 0.762 cm (0.3 inches). Due to the relatively low ETSs of the IC members, the PHYS mean outperforms the IC mean, and the ALL mean is generally comparable to the PHYS mean at all thresholds.

During the cool season (not shown) all members and means have cool season ETSs that are 0.15 to 0.2 larger than the warm season. As in the warm season, the cool season PHYS members have greater skill than the IC members on average, but the percentage benefit for the PHYS is smaller than the warm season.

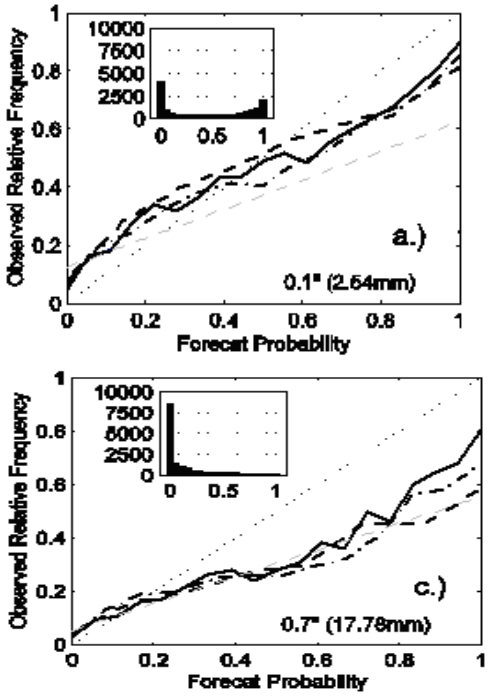

Figure 6. Reliability diagrams for three 24-hour precipitation thresholds: (a) 0.1" (0.254mm) and (b) 0.7" (17.78mm) during the warm season. The dotted straight line represents a perfect forecast, while the straight dashed line denotes no skill. The ALL, IC and PHYS ensembles are represented by the solid, dashed, and dash-dotted black lines, respectively. Inset histograms show the number of occurrences (ordinate) of each forecast probability (abcissa).

In order to examine the accuracy and reliability of the ensemble's QPF probability distribution during the warm season, the Brier score (BS) was broken down into the three components outlined in Appendix. These components are presented in reliability diagrams of forecast probabilities versus observed relative frequencies (Fig. 6). In these diagrams, perfect reliability (REL) is represented by the 1:1 dotted line. The ability of the ensemble to discriminate between events, RES, is measured by the slope of the plotted solid line, with less RES producing a more horizontal line.

The dashed line represents the line of "no skill," where REL=RES, and BSS=0.

For the 0.25 cm (0.1 inch) threshold during the warm season (Fig. 6a), lower probabilities tend to be under-forecast (solid lines lie above dotted line) and higher probabilities tend to be over-forecast (solid lines below dotted line) for PHYS, IC, and ALL. Thus, the ensemble tends to over-predict the 24-hour precipitation probabilities. For larger thresholds, such as 1.78 cm (0.7 inches) in Fig. 6b, the IC ensemble shows no skill at all probabilities, whereas the PHYS and ALL ensembles retain skill and reliability at high probabilities (> 75%) at all thresholds.

For the BS components during the cool season (not shown), all ensemble suites tend to be over-confident in 24-hour precipitation probabilities, except at very low (< 25%) probabilities. The IC ensemble forecast probabilities have greater reliability at all thresholds than the PHYS or ALL ensembles for probabilities over 50%. Each of the ensemble grouping's reliability approaches the "no-skill" line at even the 0.254 cm (0.1 inch) event threshold for low-to-mid forecast probabilities (20-50%), and remain unskillful for all event thresholds. Higher probabilities (75-100%) retain skill and reliability for all thresholds, with the IC ensemble probabilities showing greatest reliability for the higher probabilities. These results illustrate that an ensemble weighted more with IC members can prove beneficial for cool season precipitation forecasting over the Northeast U.S.

4. SUMMARY

A short-range ensemble forecast (SREF) system was constructed over the Northeast United States down to 12-km grid spacing using 18 members from the Penn State University – National Center for Atmospheric Research (PSU-NCAR) Mesoscale Model Version 5 (MM5). The ensemble consisted of 12 members with varying planetary boundary layer (PBL) schemes and convective parameterizations (CPs) as well as seven different initial conditions [five NOAA's National Center for Environmental Prediction (NCEP) Eta-bred members at 2100 UTC and the 0000 UTC NCEP Global Forecast System and Eta runs]. The MM5 SREF system was verified over the warm (May-September 2003) and cool (October 2003-March 2004) seasons for several surface parameters.

All ensemble members have an appreciable diurnal temperature and wind speed bias that varies by PBL type. During the warm season the magnitude of the cool surface bias for the Mellor-Yamada-Janjic (MYJ) PBL during the day depends on the CP scheme utilized. The MYJ-PBL cool biases during the day partially cancel the warm biases from other PBL members, resulting in ensemble mean having the most skill on average. Because of clustering of PBL parameterizations during the cool season, the IC members are more useful than the physics members, but none of the members outperform the GFS-MM5 for sea-level pressure. The ensemble outperforms the NCEP Eta model on average and it has similar skill as the deterministic MM5 initialized 12-hours later.

Spread-error correlations are lowest (0.09 to 0.4) for temperature and wind speed given the large biases and clustering prevalent among ensemble members. A 14-day bias calibration improves the ensemble under-dispersion of temperature and winds, but an appreciable bias still exists. Correlations are largest for 10-meter wind direction (0.6 to 0.7), indicating that ensemble variance can be used as an approximation for ensemble uncertainty for this parameter over the Northeast U.S. Although the reliability of ensemble probability of precipitation is only moderate for most accumulation thresholds, probabilistic precipitation is more skillful than sample climatology. For the 24-hour precipitation forecasts, the physics ensemble has the greatest skill and reliability during the warm season, while the initial condition ensemble provides the largest benefit during the cool season.

4. ACKNOWLEDGEMENTS

The research was supported by ONR (Grant N000014-00-1-0407) and UCAR-COMET (Grant S0238662). Collective insights from the COMET collaborative NWS partners improved the direction of this research.

5. REFERENCES

Colle, B.A., J. B. Olson, and J. S. Tongue, 2003: Multi-season verification of the MM5: Part I, Comparison with the Eta over the Central and Eastern U.S. and impact of MM5 resolution. *Wea. Forecasting*, **18**, 431-457.

_____, J.B. Olson, and J.S. Tongue, 2003b: Multiseason verification of the MM5. Part II: Evaluation of high-resolution precipitation forecasts over the Northeastern United States. *Wea. Forecasting*, **18**, 458-479.

Jones, M. S., 2004: Evaluation of a mesoscale short-range ensemble forecasting system over the Northeast U.S., M.S. thesis, Marine Sciences Research Center, Stony Brook University, 135 pp.

McMurdie, L., and C. F. Mass, 2004: Major numerical forecast failures over the Northeast Pacific. *Wea. Forecasting*., **19**, 338-356.

Stensrud, D. J., and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.*, **131**, 2510-2524.

Yussouf, N., D. Stensrud, and S. Lakshmivarahan, 2004: Cluster Analysis of Multimodel Ensemble Data over New England. *Mon. Wea. Rev.*, **132**, 2452-2462.

Zhang, F., C. Snyder, and R. Rotunno, 2002: Mesoscale predictability of the "surprise snowstorm of 24-25 January 2000. *Mon. Wea. Rev.*, **130**, 1617-1632.