9.10 FORECAST EVALUATION OF A MIXED-PHYSICS ENSEMBLE

Matthew S. Wandishin*, Steven L. Mullen
University of Arizona, Tucson, Arizona

David J. Stensrud, Harold E. Brooks
NOAA/National Severe Storms Laboratory, Norman, Oklahoma

1. INTRODUCTION

The importance of accounting for model error in ensemble forecasting has been demonstrated by recent case studies for both medium- (Harrison et al. 1999) and short-range (Stensrud et al. 2000) forecasts and by the implementation of a stochastic representation of precipitation processes in an operational setting (Buizza et al. 1999).

The present study is a first step toward extending the work of Stensrud et al. (2000) and beginning a systematic examination of the mixed-physics ensemble. A mixed-physics ensemble is one in which the members contain different formulations of physical parameterizations. As such, the ensemble can be used to examine the relative importance of different physical processes and relative effectiveness of individual parameterizations. Herein, the emphasis will be on the probabilistic forecasts of precipitation.

As explained by Murphy (1993), the "goodness" of a forecast involves not only quality but value, as well. Furthermore, there exist several different aspects of forecast quality, including the familiar measures of accuracy and skill along with aspects such as reliability, sharpness, and discrimination, which are based on the joint distribution of forecasts and observations (Murphy 1993). An exhaustive examination of forecast goodness is beyond the scope of this paper but an attempt will be made to give a broad evaluation of the performance of the mixed-physics ensemble.

2. ENSEMBLE DESCRIPTION

The ensemble system used in this study consists of nine members using identical initial conditions and all the different possible combinations of three convection parameterizations (CP) and three planetary boundary layer (PBL) schemes. The three CPs are the Kain-Fristch (KF; Kain and Fristch 1993), Betts-Miller (BM; Betts and Miller 1986), and Grell (GR; 1993) schemes. The PBLs employed are those developed by Blackadar (BL; 1979) and Burk and Thompson (BT; 1989), and the scheme used in the National Centers for Environmental Prediction's (NCEP) Medium Range Forecast model (MF; Hong and Pan 1996). These parameterizations are plugged into the Pennsylvania State University-National Center for Atmospheric Research Mesoscale Model version 5 (MM5; Grell et al. 1994) with 30 km horizontal grid spacing and 23 vertical sigma levels, covering most of the contiguous United States. The ensemble members are initialized from the 00Z NCEP Eta model analyses. The data sets consist of 43 cases between April and June 1999 for which 36 h forecasts were made. Twenty-four hour rainfall accumulations from the 4-km gridded analyses of nearly 3000 hourly raingage observations, available as part the NCEP's Stage IV rainfall data, are used for verification. The verification data are valid at 12Z, corresponding to the 12-36h model forecast period.

3. RESULTS

3.1 Attributes diagram

Several aspects of forecast quality can be examined through the use of an attributes diagram. The reader is referred to Wilks (1995) for a complete discussion of the features of attributes diagrams. Figure 1 shows an attributes diagram for forecasts of at least 0.01" of rain from the nine possible 8-member ensembles achieved by excluding, in turn, each member of the full ensemble.

A primary feature of these curves is the lack of significant variation between the curves; the notable exception being the bifurcation at forecasts of 37.5%. Insight into the cause of this bifurcation is provided by the distribution of forecast probabilities for each ensemble (Fig. 1). Again a split is seen at the forecast probabilities of 25 and 37.5%. The split distinguishes the six ensembles containing all three members using the BT PBL scheme and the three ensembles containing only two of the BT members. The BT members possess a distinct wet bias for low precipitation thresholds (not shown). Thus, the ensembles containing all three BT members are more likely to produce a forecast of 37.5% (3/8), whereas those containing only two of the BT members will favor forecasts of 25% (2/8).

Another striking feature of Fig.1 is the high degree of sharpness to the forecasts. Nearly 70% of all the forecasts come from the forecast probabilities zero and one. In other words, the ensemble frequently provides very confident predictions of rain or no rain. However, the large number of forecasts of probability zero or one could also indicate a lack of ensemble spread. For a majority of the forecasts the ensemble members are in complete agreement.

Curves along the 45° line indicate perfect reliability (forecast probability equals the observed relative frequency), whereas curves falling along the

*Corresponding author address: Matthew Wandishin, NSSL, 1313 Halley Circle, Norman, OK, 73069. Email: Matt.Wandishin@nssl.noaa.gov

diagonal lying between the perfect reliability line and the climatological frequency (the horizontal line) indicate zero skill relative to a forecast based solely on climatology. Thus the ensembles exhibit marginal skill for the forecasts probabilities 0, 12.5, 25, and 100% and no skill for all other probabilities.

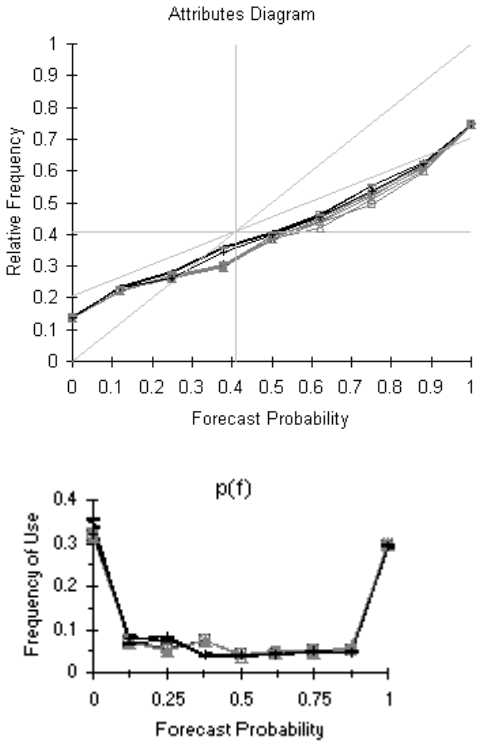

Fig. 1. Attributes diagram for 0.01" precipitation threshold and 8-member ensembles (top) and distribution of the relaitve use of forecast probabilities (bottom).

Finally, a reliability curve located beneath the 45° line indicates an overforecast (wet bias) while a curve above the 45° line indicates an underforecast (dry bias). The ensembles overforecasts for the lowest forecast probabilities and underforecast the moderate and high probabilities. The wet bias of the BT members is evidenced once again as the six ensembles with all three BT members overforecast at the 37.5% probability while the other ensembles are nearly perfectly reliable at that probability.

Similar behavior is seen for each of these forecast quality aspects for ensembles comprised of fewer ensemble members and, to a lesser degree, at higher precipitation thresholds. An illustration of this similarity is provided the behavior of the Brier score (BS) and Brier Skill score (BSS) (Fig. 2). The BS and BSS are based entirely on information derived from the attributes diagram, namely the reliability, resolution, and uncertainty (Wilks 1995). The top panel shows the BS and BSS for 0.01" precipitation forecast for different ensembles comprised of eight down to four members. The BS is nearly completely insensitive to the ensemble size or the different combinations of members for each ensemble size. Somewhat greater sensitivity to both the inter- and intragroup differences is seen for the BSS, with skill decreasing with decreased ensemble size but the sensitivity is still relatively slight. The BS and BSS both drop more significantly as the precipitation threshold is increased to 1.00" (Fig. 2, bottom), but the description given above for the behavior of the different ensemble groups at the lower threshold applies equally for forecasts of higher-end events. Note that the ensembles exhibit negative skill for these forecasts of more significant rainfall amounts, meaning that a forecast based on climatology is more skillful than the ensemble.

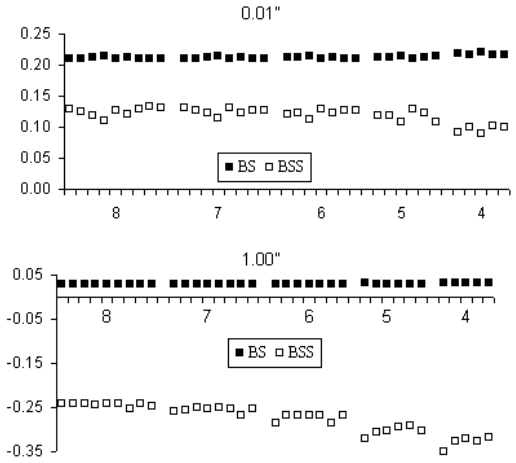

Fig. 2. Brier Score (BS, closed box) and Brier Skill Score (BSS, open box) for 0.01" (top) and 1.00" (bottom) thresholds. The number along the abscissa indicates the number of ensemble members in that group.

3.2 Relative Operating Characteristics

Another aspect of forecast quality is discrimination, which examines the ability of forecasts to distinguish between events and nonevents. For example, strong discrimination would result if events typically are preceded by high forecast probabilities and nonevents are typically preceded by low forecast probabilities (Murphy 1993). A measure of forecast discrimination that has become increasingly popular in recent years is the relative operating characteristic (ROC; Mason 1982). For each possible forecast probability threshold, the probabilistic forecast can be turned into a deterministic forecast by treating each probability greater than the threshold as a 'yes' forecasts and each probability below the threshold as a 'no' forecast. In this way, a hit rate (HR) and false alarm rate (FAR) can be calculated for each probability threshold. Plotting the array of HRs versus the array of FARs gives a ROC curve. The area beneath this curve is a measure of the forecasts discrimination. A ROC area equal to unity denotes perfect discriminatory ability while an area of 0.5

denotes no discriminatory ability. There are two methods for calculating the ROC area. The most commonly used method is the trapezoidal rule, which is sensitive to the number of points along the curve (i.e., the number of precipitation thresholds, or in the present case the number of ensemble members). The second method involves fitting a line to the ROC curve transformed into normal deviate space (Mason 1982), resulting in a smoothed curve that is not sensitive to the number of ensemble members. The former method can be thought of as a measure of the actual performance of the forecast system while the latter method can be thought of as a measure of the potential performance of the forecast system (e.g., of an infinite-member ensemble where each member possess similar ability to the present members) (Bamber 1975).

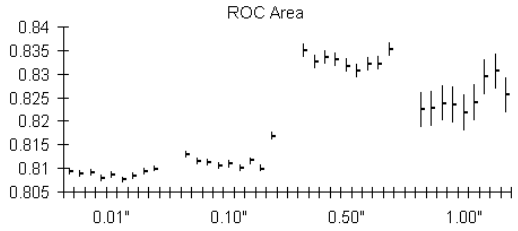

Fig. 3. Fitted ROC area for the 8-member ensembles for the four precipitation thresholds.

Figure 3 gives the smoothed ROC areas for the 8-member ensembles for four precipitation thresholds (0.01", 0.10", 0.50", and 1.00"). Slightly more variability is seen in the performance of the different ensemble configurations, but most of that variability is not statistically significant to the 95% level, as shown by the error bars. (The longer error bars for the higher thresholds reflect the substantially smaller sample size for events of that magnitude.) Similar to the behavior of the BSS, the ROC areas do exhibit moderately greater variability between the different precipitation thresholds than for the different configurations within each threshold. The discrete ROC areas as determined by the trapezoidal rule (not shown) are somewhat lower in magnitude overall and display a decrease as the ensemble size decreases, as expected. However, the different ensemble configurations within each size group behave in the same manner as for the smoothed curves.

The peak in areas at 0.5" may at first appear counter-intuitive, as it is generally expected that rainfall forecasts become more difficult as the accumulation increases. Namely, heavy rain events are very difficult to forecast. However, 0.50" reflects more of a moderate rain event such that a strong signal may be present (e.g., a strong forcing mechanism) but the event is not so rare as to exceed the ability of the model to capture it.

3.3 Value

As stated in the introduction, forecast goodness involves value as well as quality. It could be argued that value is a more important indicator of goodness than quality as it takes into account the users of the forecasts, as well. Forecast value can be calculated directly from HRs and FARs used in the ROC curves. See Richardson (2000) for a comprehensive discussion on calculating the value of ensemble forecast systems. As with the BSS, forecast value is a relative measure such that a value of zero means that the forecasts provide no improvement over climatology and a value of unity denotes that the forecasts supply maximum benefit to the user.

Figure 4 presents value curves for the different ensemble sizes as function of the cost-loss ratio (C/L). Every preventative action taken against an event incurs a cost. The cost-loss ratio expresses this cost as

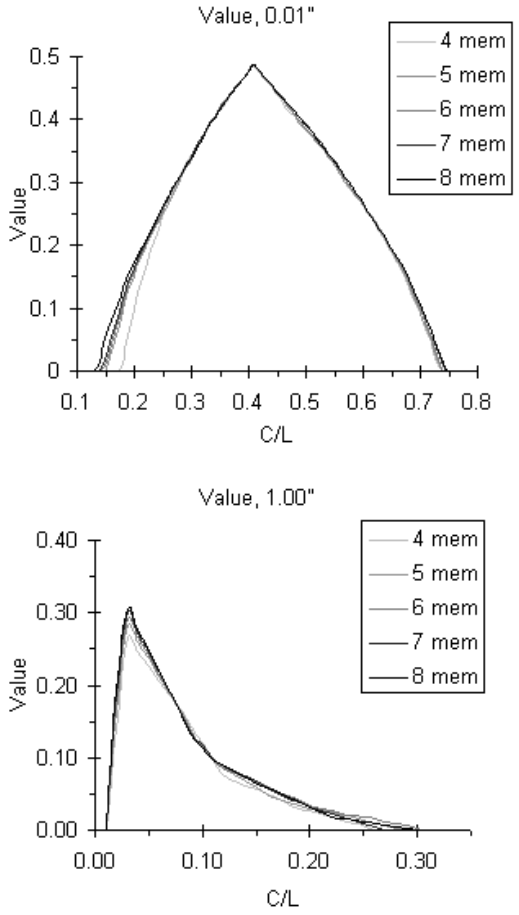

Fig. 4. Value curves for the 0.01" (top) and 1.00" (bottom) thresholds as a function of cost-loss ratio for the different ensemble sizes.

a fraction of the potential loss protected by that action (Thompson 1950). Once again the variation between different ensemble sizes is quite small, with the only discernable difference in the value curves occurring in the left wing of the 0.01" curves and in the maximum value of the 1.00" curves. However, most real world users have low C/L (H. Brooks, personal communication) and so these differences can have

significant impact. For example, a user with a C/L=0.17 would get no value out of the 4-member ensemble, but would realize a 15% improvement with the 8-member ensemble. Also, note that despite the negative skill possessed by the ensembles at the higher precipitation threshold (Fig. 2), the ensembles do offer positive value to a substantial segment of users.

4. SUMMARY

A preliminary evaluation of a mixed-physics ensemble has been presented following the framework of Murphy (1993). The ensemble performs rather poorly for several aspects of forecast quality such as reliability, accuracy and skill. However, the ensemble is strong in the areas of sharpness and discrimination. Apparently these strengths balance the previously mentioned deficiencies such that the ensemble does provide substantial value to a wide range of potential users of the forecasts. This points out the inadequecy of using only a few select measures of forecast goodness.

Further research into the relationship between value and the various aspects of forecast quality is warranted. Examination of the quality and value of the individual forecast members could aid in this endeavor. Such analysis also could help in explaining the surprisingly small variability in the quality and value measures as a function of ensemble configuration or as a function of ensemble size. Selective groupings of ensemble members would allow for comparisons between the different representations of the PBL and convective processes and help determine the relative importance of the PBLs versus the CPs.

Acknowledgments. M. Wandishin and S. Mullen were supported by NSF Grant ATM-9612487.

REFERENCES

Bamber, D., 1975: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.,* **12,** 387-415.

Betts, A.K., and M. Miller, 1986: A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, ATEX and arctic air-mass data sets, *Quart. J. Roy. Meteor. Soc.*, **112,** 693-709.

Blackadar, A.K., 1979: High resolution models of the planetary boundary layer. Vol. 1, No. 1, J. Pfafflin and E. Ziegler, Eds., Gordon and Breach, 50-85.

Buizza, R., M. Miller, and T.N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **125,** 2887-2908.

Burk, S.D., and W.T. Thompson, 1989:, A vertically nested regional numerical weather prediction model with second-order closure physics, *Mon. Wea. Rev.*, **117,** 2305-2324.

Grell, G. , 1993, Prognostic evaluation of assumptions used by cumulus parameterizations, *Mon. Wea. Rev.*, **121,** 764-787.

____, J. Dudhia, D.R. Stauffer , 1994: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR/TN-398+STR, 121 pp. [Available from MMM Division, NCAR, P.O. Box 3000, Boulder, CO 80307].

Harrison, M.S.J., T.N. Palmer, D.S. Richardson, and R. Buizza, 1999: Analysis and model dependencies in medium-range ensembles: two transplant case-studies. *Quart. J. Roy. Meteor. Soc.*, **125,** 2487-2515.

Hong, S.-Y., and Pan, H.-L. , 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model, *Mon. Wea. Rev.*, **124,** 2322-2339.

Kain, J.S, and J.M. Fritsch, 1993: Convective parameterization for mesoscale convective systems: The Kain-Fritsch scheme. *The representation of cumulus convection in numerical models. Meteor. Monogr.*, No. 24, Amer. Meteor. Soc., 165-170.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **37,** 291-303.

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8,** 281-293.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126,** 649-667.

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

Thompson, J.C., 1950: A numerical method for forecasting rainfall in the4 Los Angeles area. *Mon. Wea. Rev.*, **78,** 113-124.

Wilks, D.S., 1995: Statistical Methods in the Atmospheric Sciences. Academic Press, 467 pp.