6.7 SPREAD-SKILL RELATIONSHIP IN THE CANADIAN ENSEMBLE PREDICTION SYSTEM

Franck Pithois
Météo-France

Richard Verret*
Louis Lefaivre
Gérard Pellerin
Marc Klasa

Peter Houtekamer
Lawrence Wilson

Meteorological Research Branch

Canadian Meteorological Centre

1. INTRODUCTION

The usefulness of an Ensemble Prediction System (EPS) resides mostly in the variety of possible solutions that the system can offer to a given meteorological forecast problem. The differences between all weather scenarios presented by each member of the EPS, or the variance within the ensemble forecasts, lead toward the study of the spread-skill relationship. If such a relationship exists, it would then be possible to associate higher skill and better confidence in the forecasts when the ensemble variance is low, and vice versa. The outcomes of such a spread-skill relationship, amongst others, include the possibility of forecasting the forecast skill on one hand and to use the spread of the ensemble as an indication of the confidence on the deterministic forecast on the other hand.

The Perfect Prog (PP) (Klein et al, 1959) statistical adaptation system operational at the Canadian Meteorological Center (CMC) has been run on each of the sixteen members of the CMC EPS. The ensemble variance of the statistical 12-h probability of precipitation (PoP) forecasts is being evaluated as a proxi for a confidence index. The 12-h PoP forecasts generated from each member of the ensemble have been verified at all projection times to ten days, at 264 Canadian stations over the period extending from June 2000 to February 28 2001 inclusively. The skill of the ensemble forecasts average converges toward that of climatology by the 180 hour projection time. This is an indication that skill with respect to climatology can be expected up to seven days. Contingency tables of Brier scores of the 12-h probability of precipitation forecasts of the control model versus the ensemble variance have been constructed to study the spread-skill relationship. The results indicate that it is possible to use the ensemble variance as a proxi for a confidence index. Based on the chi-squared test, the spread-skill relationship is statistically significant to the 240 hour projection time.

A brief description of the CMC EPS will be presented. The spread-skill relationship applied to the PP statistical 12-h PoP forecasts will be demonstrated and results of the cross-validation of the confidence index in forecast mode will be presented.

2. THE CANADIAN ENSEMBLE PREDICTION SYSTEM

The methodology used at CMC to generate the ensemble members of the Canadian EPS is described in Houtekamer et al (1996). The basis of the methodology is to produce perturbed analyses through data assimilation procedures. In order to produce *n* perturbed analyses, *n* parallel analysis cycles are run quasi-independently. The characteristics of these perturbed cycles are as follows: each eigenvector of the covariance matrix for the observational error is multiplied with a random value; the resulting perturbation vector is then added to the observations used in the data assimilation cycles. The random values are different for each piece of information and are different from one perturbed cycle to the other. Currently *n* is set to eight, which means that eight perturbed data assimilation cycles are run producing eight perturbed analyses. The data assimilation model is a global spectral (SEF) model (Ritchie, 1991). The number of perturbed analyses is doubled by adding the opposite

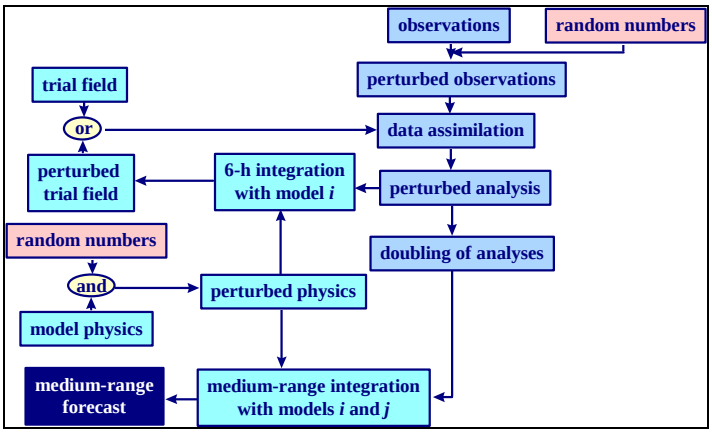

Figure 1: Flow chart of the Canadian Ensemble Prediction System. See text for more details.

* *Corresponding author address*: Richard Verret, Development Branch, Canadian Meteorological Centre, 2121 Trans-Canada Highway, Dorval (Quebec), Canada, H9P 1J3 (E-Mail: Richard.Verret@ec.gc.ca)

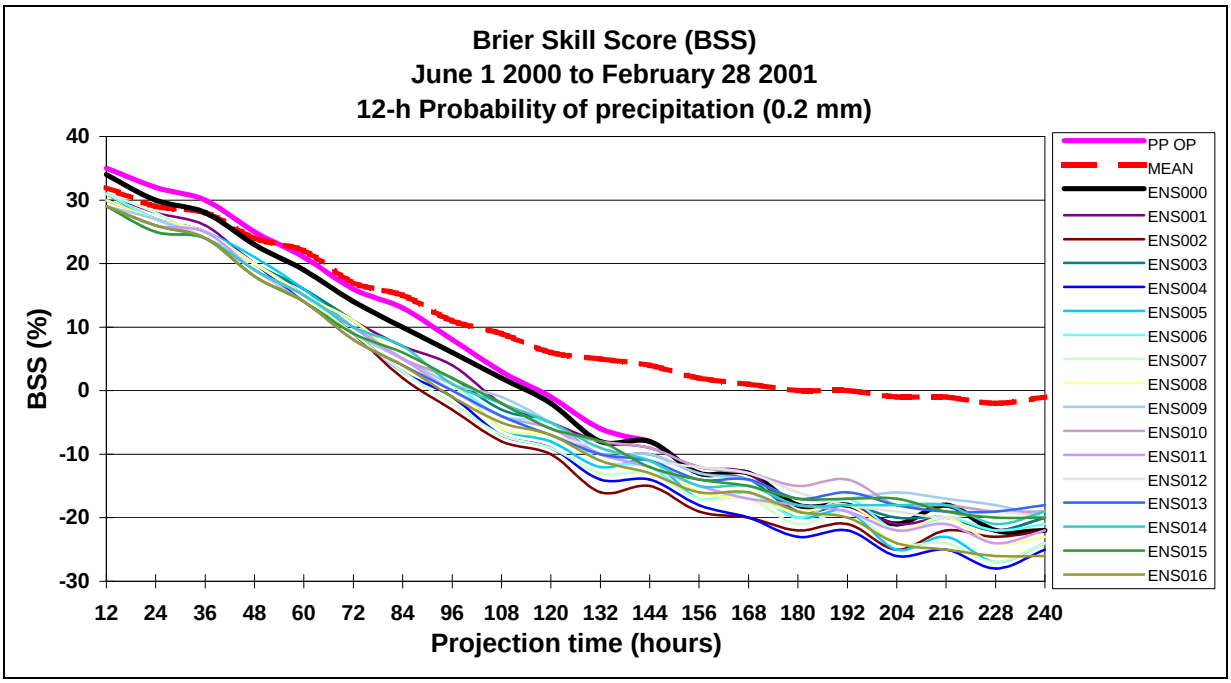

Figure 2: Brier Skill Score (BSS) for the 12-h Probability of precipitation from each of the ensemble members, the control model (ENS000, thick dark line), the operational deterministic model (out to 144-h projection time only) and for the mean of the ensemble forecasts (thick dashed line).

of the residuals between the control analysis and the average of the eight original perturbed analyses. At the end of the process, sixteen analyses are available for model initialization.

Each of the sixteen perturbed analyses are used to run eight versions of the SEF model and eight versions the GEM model (Côté et al, 1998), one perturbed analysis being coupled with one specific model version. Each of these models have different switches activated in their physics parameterization, and some physical parameters are set with random values (horizontal diffusion, minimal roughness length over sea and time filter). Perturbations are also introduced in the surface forcing through perturbation of the fields for sea surface temperature, albedo and roughness length. Details on the different model configurations can be found in Lefaivre et al (1997).

Figure 1 shows the current set-up of the Canadian EPS. The main advantage of this set-up is that it truly models the observational errors and the numerical weather prediction errors. The resolution of the SEF model is T95 at the time of this study (currently T150 since June 2001) and that of the GEM model is 1.875° or approximately 200 km at the time of this study (currently 1.2° or 150 km since June 2001). The EPS is integrated out to ten days (240 hours) in the 00 UTC production cycle at the CMC.

3. PROBABILITY OF PRECIPITATION ON THE EPS

The CMC has a long history in statistical post-processing of numerical weather prediction model outputs. Two systems are currently operational, one based on the PP approach (Verret, 1992) and a more recent one based on the MOS approach (Ghlan et al, 1972), but in an updateable (UMOS) framework (Vallée et al, 1998). The statistical guidance includes spot temperatures at every three hour intervals, total cloud opacity at three hour intervals and 6- and 12-h PoP forecasts at different precipitation amount thresholds. Since the PP approach is independent of the driving numerical model, any PP system can be run off any driving model. This makes it the perfect candidate to run statistical post-processing as part of the EPS. In that context, all PP statistical weather element forecasts are being generated using the outputs of each of the sixteen members of the EPS. This study looks only at the 12-h PoP forecasts.

Twelve hour PoP's are forecast for three thresholds, 0.2, 2 and 10 millimeters. The predictand is the total observed precipitation accumulations over two consecutive 6 hour periods converted to binary form for each of the three categories. PP linear regression is used which has been developed from twenty-two years of historical data (1963-1984) stratified into three month seasons with the exception of the 10 mm threshold where there are two 6 month seasons due to the fact that this is a rarer

occurrence. PP forecasts are prepared for 12 hour intervals out to 240 hours based on each member of the EPS at an ensemble of 264 Canadian stations.

The statistical PP 12-h PoP forecasts from each member of the EPS and the average of the 12-h PoP forecasts have been verified over the period from June 1 2000 to February 28 2001, using the Bier Skill Score (BSS) as a measure of skill. Figure 2 shows the evolution of the BSS as a function of projection time out to 240 hours, for each member of the EPS (thin lines in different shades of gray), for the average of the forecast (thick dashed line) and for those generated from the operational high resolution deterministic model (thick line). It can be seen that the behavior of PP PoP forecast is very consistent between all sixteen members of the EPS. In general, the skill of the forecasts crosses the zero line at approximately 96-h projection time, but the skill of the forecasts keeps falling beyond and appears to level off around the 200-h projection time. This implies that, although on average, climatology seems better than the actual forecast after 96-h, there is still information embedded in the forecasts because there is still a dependency between the skill and projection times. The forecasts generated from the high resolution deterministic model show a BSS that crosses the zero line at approximately 120 hours showing an overall skill better than any of the EPS members. This has to be expected since the deterministic model has a better resolution and also because it runs off an unperturbed high resolution analysis. But what is striking in Figure 2 is that the skill of the ensemble averaged 12-h PoP forecasts over all sixteen members converges asymptotically toward zero, thus toward that of climatology, after 156-h projection time. This implies that the EPS weather element forecasts converge toward climatology as the usable information falls with projection time. On average there is little information left in the 12-h PoP forecasts beyond 156-h projection time and climatology should be used as a better forecast, but that does not preclude the fact that there are a few cases where there is still valuable information in the forecasts that can be used, if there is a relationship between the variance of the forecasts and their skill. The question is how to pinpoint these cases.

4. SPREAD-SKILL RELATIONSHIP

Following Lefaivre et al. (1997), the spread-skill relationship amongst the PP 12-h PoP forecasts based on the EPS has been studied using two by two contingency tables of the variance of the forecasts against a measure of accuracy. In this case, the Brier score (BS) has been used as an estimate of the accuracy of the forecasts. These contingency tables are constructed by classifying the variance of the forecasts above or below the median value and by categorizing the accuracy of the forecasts into two classes, above or below the median value of the BS. Such two by two contingency tables were constructed for each station independently and then summed up together for each projection time. The hypothesis is that higher skill will be associated with lower variance amongst the forecasts and vice versa. Figure 3 shows the fraction of occurrence of good spread-skill association (thick curve marked N2+N3) as a function of projection time out to 240 hours and the fraction of the forecasts which show a better skill associated with a larger variance and a lower skill associated with a lower variance (dashed line marked N1+N4). A clear separation can been seen between the two curves, with a larger proportion of the forecasts being correctly associated with the variance of the forecasts at all projection times. This is an indication that a spread-skill relationship exists amongst the forecasts. The chi-squared test shows that the spread-skill relationship is statistically significant at the 95% confidence level at all projection times out to 240 hours.

These results serve as a basis for the development of a confidence index. It was assumed that when the variance in the forecasts is low or in the lower tercile, the likelihood of having skillful forecasts will be high. On the opposite, when the variance of the forecasts is large or in the third (higher) tercile, the probability of having good forecasts is low. All cases in the second (middle) tercile are neutral, or in other words, the probability of having a good forecast is just as high as that of having a poor forecast. Figure 4 shows the expected BS for each of the variance terciles over the period from June 1 2000 to December 31 2000. There is one curve for each 12-h period out to 240 hours, the uppermost curve being for the 12-h forecasts, while the bottom one is for the 228-240-h forecasts. Each curve has a definite and

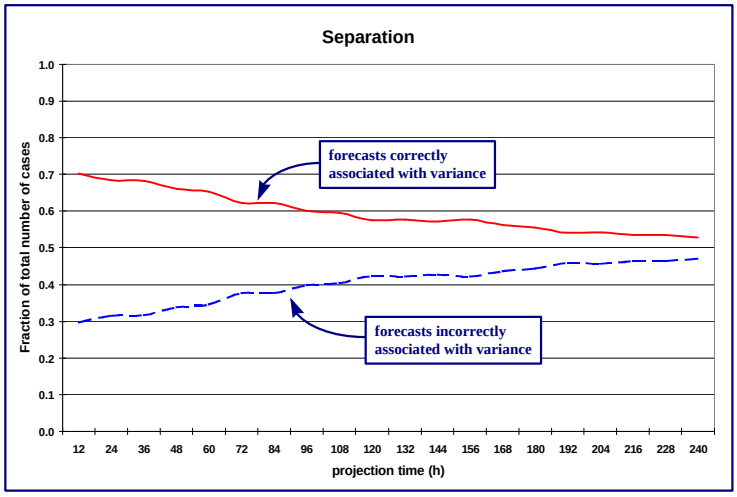

Figure 3: Fraction of the forecasts with a correct association between variance and skill (upper curve). The lower curve shows the percentage of the forecasts with an incorrect association between skill and variance.

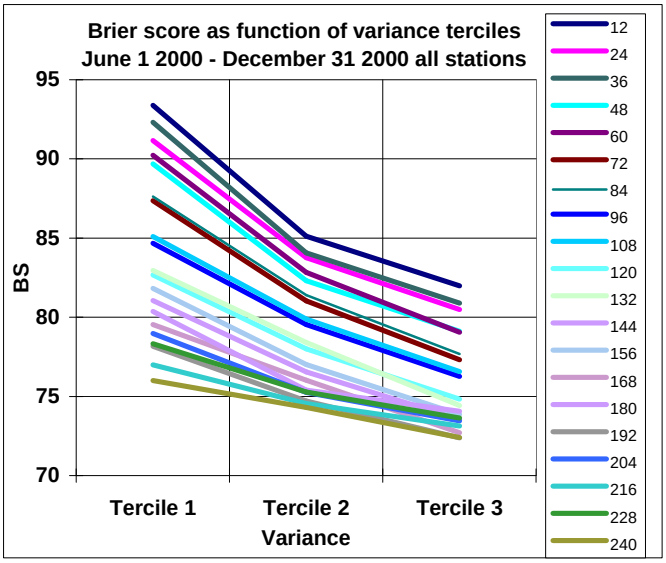

Figure 4: Expected Brier score for each tercile of variance amongst the 12-h PoP forecasts. There is one curve for each 12-h period between the 0- and 240-h projection time. The uppermost curve corresponds to the 0- to 12-h forecasts while the bottom one is for the 228-to 240-h forecasts.

statistically significant negative slope, although the slope of the curves decreases with projection time. The confidence index is naturally defined with the terciles.

The generation of the confidence index is done with respect to reference tercile thresholds. These threshold values are calculated over the previous season/year and used to determine in which tercile the current variance of the 12-h PoP forecasts falls. Tercile one will correspond to an index of one (high confidence) and so on. The confidence index has been verified in cross-validation mode, over the period from June 2000 to February 2001, on a seasonal and monthly stratification basis. The cross-validation shows a decreasing performance of the 12-h PoP forecasts with increasing tercile of variance. There is also indication that the strength of the spread-skill relationship is independent of the season, particularly at the shorter projection times out to 144-h. However, the nature of precipitation varies from Winter to Summer, being more of the synoptic scale nature in Winter as opposed to convective precipitation during Summer. The variance of the 12-PoP forecasts is generally low in Summer, and particularly in July. This means that the confidence index generation in operational mode must take into account this seasonal variation in the variance of the forecasts.

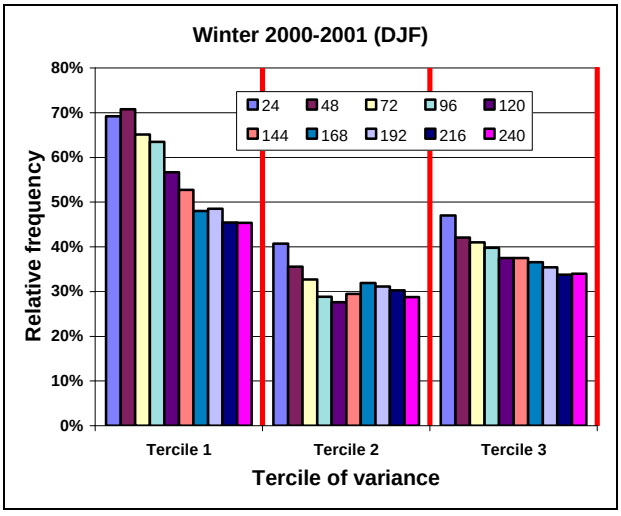

Figure 5: Relative frequency of the forecasts in each tercile of variance with a Brier score in the corresponding tercile. There is one histogram for each 24-h projection time out to 240-h. The first bar in each tercile of variance is for the 12- to 24-h forecasts, while the last one is for the 228 to 240-h forecasts.

It should be expected that the forecasts in tercile one (or with a confidence index of one) should have a BS that also falls in the lower tercile of the BS scores. Figure 5 shows the relative frequency of the 12-h PoP forecasts per tercile of variance of the forecasts, that did show a BS in the corresponding correct tercile. Figure 5 has been generated in cross-validation mode. It is encouraging to see that forecasts in tercile 1, do show a better skill in general than the forecasts in the other two terciles. It is also encouraging to see that the expected relationship between the confidence index and the BS tercile is very good, particularly at the shorter projection times. More that 70% of the forecasts in tercile one have a BS in the corresponding tercile of the verification scores at the 48-h projection time, and more that 50% of these forecasts show a BS in the lower tercile out to the 144-h projection time. This implies that the confidence index can be used at least out to day 6 and perhaps even beyond, since the percentage of the forecasts with a BS in the correct lower tercile is above 33% out to 240-h.

5. CONCLUSIONS

Statistical PP 12-h PoP forecasts on each of the sixteen members of the Canadian EPS have been used to study the spead-skill relationship of the EPS. The choice of the statistical 12-h PoP forecasts as a proxi for confidence index may seem arbitrary but is justified by the fact that most of the predictors (if not all of them) for the 12-h PoP's come from the model mass fields. Consequently, the variance of these statistical forecasts is likely to provide a realistic estimate of the overall variance amongst the ensemble members. It has also to be realized, that the ultimate goal of a confidence index is to provide

users the level of confidence on public forecasts, and probability of precipitation is this context appears as the most important and appropriate parameter.

It has been shown that a spread-skill relationship does exist amongst the members of the EPS, and that this relationship is statistically significant at all projection times out to 240-h. From these results, a confidence index has been developed and validated in cross-validation mode and the results appear promising. Further work is needed before such a confidence index is implemented.

The current set-up uses as the forecast, the one generated off the operational high resolution deterministic model, while the confidence index is calculated from the EPS. There may be discrepancy between what the deterministic model says, and what the EPS proposes as forecasts. The generation of the tercile thresholds in operational mode also has to be studied. Finally, this study needs to be put in perspective with other techniques, such as tubing (Atger, 1999) for example.

6. REFERENCES

Atger, F., 1999: Tubing - an alternative to clustering for the classification of ensemble forecasts. *Wea. Forecasting*, 14, 5, 741-757.

Côté J., S. Gravel, A. Méthot, A. Patoine, M. Roch and A. Staniforth, 1998: The Operational CMC/MRB Global Environmental Multiscale (GEM) Model: Part I - Design Considerations and Formulation. *Mon. Wea. Rev*, 126, 1373-1395.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. We. Rev.*, 124, 1225-1242.

Ghlan, H. R. and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, 11, 1203-1211.

Klein, W. H., B. M. Lewis and I. Enger, 1959: Objective prediction of five-day mean temperature during winter. *J. Meteor.* 16, 672-682.

Lefaivre, L., P. L. Houtekamer, A. Bergeron and R. Verret, 1997: The CMC Ensemble Prediction System. Proc. *ECMWF 6th Workshop on Meteorological Operational Systems*, Reading, U. K., ECMWF, 31-44.

Ritchie, H., 1991: Application in a semi-Lagrangian method to a multi-level spectral primitive-equations model. *Quart. J. Roy. Meteor. Soc.*, 117, 91-106.

Vallée, M. and L. J. Wilson, 1998: The new Canadian updateable MOS forecast system. *Preprints 14th AMS Conference on Probability and Statistics in the Astmospheric Sciences Statistics in the Atmospheric Sciences*. Phoenix Arizona, 183-189.

Verret R., 1992: CMC operational statistical products. *Preprints 4th AES/CMOS workshop on operational meteorology*. Whistler B.C., 119-127.