The CMC Ensemble Prediction System

L. Lefaivre[1], P. L. Houtekamer[2], G. Pellerin[1] ,R. Verret[1] F. Pithois[3] and M. Klasa[1]
[1]Canadian Meteorological Centre [2]Meteorological Research Branch [3]Météo-France Dorval, Canada

Summary: Ensemble forecasts are run at the Canadian Meteorological Centre since 1996, operationally since February 1998, with outputs up to ten days. The ensemble size was increased from eight to sixteen members in August 1999 and the resolution was increased in July 2001. The method of producing the perturbed analyses consists of running independent assimilation cycles that use perturbed sets of observations and are driven by eight different models, mainly different in their physical parametrizations. Perturbed analyses are doubled by taking opposite pairs. A multi-model approach is then used to obtain the forecasts. The ensemble outputs have been used to generate several products. The ensemble output is also used to feed a Perfect Prog statistical package to forecast probability of precipitation used to construct a confidence index. Verifications of the ensemble forecasts will be presented with examples of the usefulness of the technique.

1. Introduction

The ensemble prediction system is running operationally at the Canadian Meteorological Centre (CMC) since February 1998. In August 1999, the ensemble size was doubled to 16, and in July 2001, their resolution was increased horizontally to 300 x 150 (equivalent of T149). Two different models are used in the production of the ensemble outputs. Section 2 will describe the method used and explain the multi-model approach. Verifications of selected products, 500 hPa heights, 10-day temperatures and precipitation, will be shown in section 3. Section 4 will sketch a confidence index based on the spread in the ensemble. Section 5 will discuss the future work planned.

2. The CMC Ensemble Method

2.1 Description of the method

The ensemble method used at CMC is described in Houtekamer et al (1996). The basis of the method is to produce perturbed analyses through data assimilation procedures. In order to produce 8 perturbed analyses, 8 parallel analysis cycles are run quasi-independently of the high-resolution operational analysis cycle. To obtain 8 additional analyses to produce medium-range forecasts, pairs of opposing initial conditions are used[1].

The schematic flow chart of the perturbed cycles is shown in Figure 1. First, each observation (i.e. radiosondes, satems, satobs, synops, ships, buoys, aireps, acars / amdars, humsat) used to feed the analysis is perturbed: 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. The random values are different for each piece of information and are different from one perturbed cycle to the other. The analysis scheme used is the Optimal Interpolation (O/I) technique (*Rutherford* 1972), which is efficient computerwise since the weights for the innovations are calculated once for the control, using a cholesly decomposition, and used for the 8 ensemble members as well. Second, each model used in the assimilation cycles, although based on the same SEF model (Ritchie 1991) with a horizontal resolution of T149 and a horizontal diffusion in (4, have different switches activated. The different configurations, used to produce the trial fields of each cycle, are described in the top half of Table 1. In addition, 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 perturbations of the fields for sea surface temperature, albedo and roughness length.

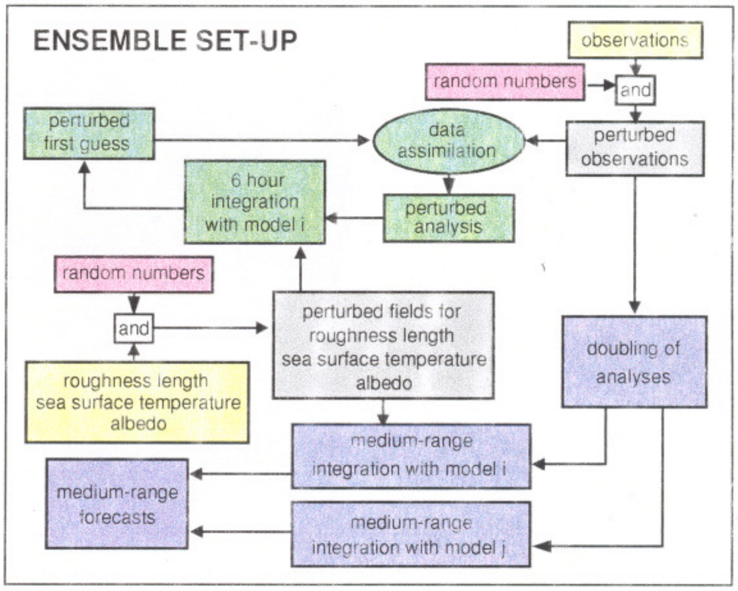

ENSEMBLE SET-UP
observations
random numbers
and
perturbed first guess
data assimilation
perturbed observations
6 hour integration with model i
perturbed analysis
random numbers
and
perturbed fields for roughness length sea surface temperature albedo
roughness length sea surface temperature albedo
doubling of analyses
medium-range integration with model i
medium-range forecasts
medium-range integration with model j

Fig 1: Flow chart of the Canadian Ensemble Prediction system.

[1] The number of perturbed analyses is doubled by the following artificial perturbation: the mean of the analyses is subtracted to the operational analysis and a fraction of this difference is added to the original perturbed analyses (see second column of Table 1).

Table 1: Combination of modules for different model versions.

SEF (T149)	Add ops analysis	Convection/Radiation	GWD[1]	GWD[1] version	Orography	No. of levels	Time level
1	yes	Kuo/ Garand	Strong	High	0.3 σ	23	3
2	no	Manabe/ Sasamori	Strong	Low	0.3 σ	41	3
3	no	Kuo/ Garand	Weak	Low	Mean	23	3
4	yes	Manabe/ Sasamori	Weak	High	Mean	41	3
5	yes	Manabe/ Sasamori	Strong	Low	Mean	23	2
6	no	Kuo/ Garand	Strong	High	Mean	41	2
7	no	Manabe/ Sasamori	Weak	High	0.3 σ	23	2
8	yes	Kuo/ Garand	Weak	Low	0.3 σ	41	2
control	mean	Kuo/ Garand	Mean	Low	0.15 σ	41	3
GEM (1.2°)	**Add ops analysis**	**Deep convection**	**Shallow convection**	**Soil moisture**	**Sponge**	**No. of levels**	**Coriolis**
9	no	Kuosym[2]	new	Less 20%	global	28	Implicit
10	yes	RAS[3]	old	Less 20%	equatorial	28	Implicit
11	yes	RAS[3]	old	Less 20%	global	28	Implicit
12	no	Kuosym[2]	old	More 20%	global	28	Implicit
13	no	Kuosym[2]	new	More 20%	global	28	Explicit
14	yes	Kuosym[2]	new	Less 20%	global	28	Implicit
15	yes	Kuosym[2]	old	Less 20%	global	28	Implicit
16	no	Kuo	new	More 20%	global	28	Implicit

[1] Gravity Wave drag (see *McFarlane* 1987 and *McLandress and McFarlane*, 1993)

[2] See *Wagneur* (1991)

[3] Relaxed Arakawa-Schubert (see *Moorthi and Suraez*, 1992)

Once a day, at 00 UTC, 10-day forecasts are produced using:

- 8 perturbed analyses, half of them obtained by taking the opposite addition to the operational analysis (see 2nd column of Table 1), using the same model options as the ones used to produce the trial fields and detailed in top half of Table 1;
- the control run, obtained from an analysis with unperturbed observations and with intermediate model options;
- 8 perturbed analyses, half of them obtained by taking the opposite addition to the operational analysis (see 2nd column of Table 1), using this time the GEM model (*Côté et al* 1998) with its own set of perturbations, detailed in bottom half of Table 1.

2.2 *Operational set-up*

The latest verification of the ensemble is seen in Figure 2, where several Root Mean Square (RMS) scores of 500 hPa for Northern Hemisphere are shown for September 2001. The solid line indicates the RMS score for the deterministic model (GEM 0.9° L28), while the small dashed line represents the ensemble mean RMS score; the dotted line represents the spread of the ensemble. One would like to see the RMS error for the ensemble mean smaller than for the high resolution model and the spread of the ensemble close to the ensemble mean error. Both the increase of the number of members in the ensemble and the increase in resolution have reduced the mean RMS error and increased the spread in the ensemble. However, the deterministic system has benefited lately from improvements in the data assimilation treatment of observations (e.g. TOVS instead of SATEMS, analysis on model's levels, etc.) that the ensemble system did not get. This issue will be addressed in Section 4.

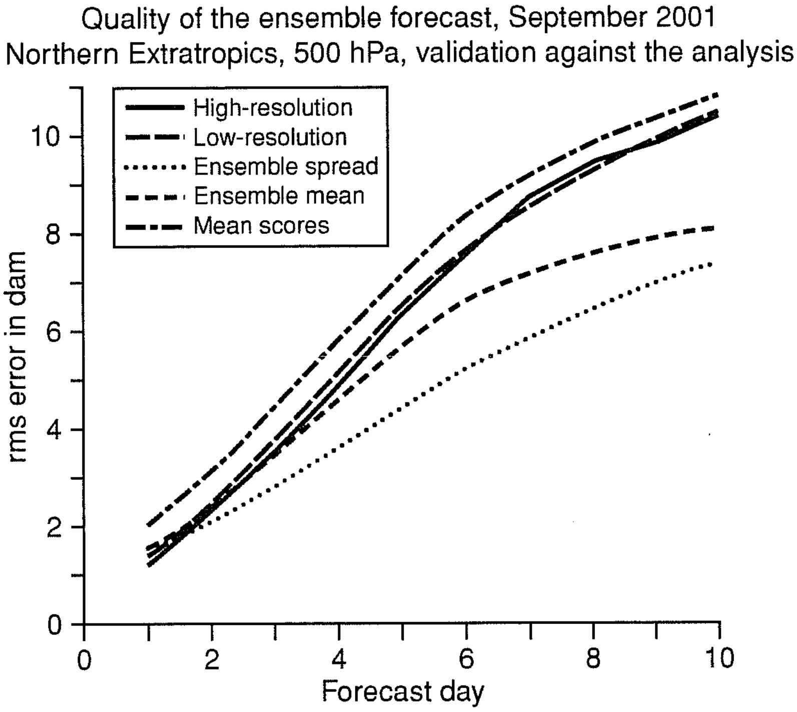


Fig 2: Verification (RMS score) of 500 hPa over northern extra–tropics (N of 20(N) for the ensemble (dashed line) as compared to the deterministic model (solid line); spread in the ensemble is dotted line (also shown: control and mean of the ensemble).

3. Products and verification results

All the products from the CMC ensemble are available via the web at the following address: http://weatheroffice.ec.gc.ca/ensemble/index_e.html. Every day the following products are made:

- Spaghetti plots;
- 10 day mean temperature anomaly;
- Calibrated probability of equivalent precipitation;
- Accumulated quantity of precipitation;
- Sea level pressure centres;
- Geopotential heights 500 hPa maps;
- Ensemble spread of trial fields.

3.1 500 hPa spaghetti plots

Spaghetti plots show several control lines for the different members, along with the one from the operational model. Also shown are the ensemble mean in thick red line and the ensemble spread in coloured background. After the fact, the verifying analysis is also displayed.

To verify if the spread of the ensemble encompasses the verifying analysis, Olivier Talagrand proposed a statistical method of displaying bias and dispersion within the ensemble (*Talagrand et al* 1997). Talagrand diagrams (or rank probability diagrams) are obtained by checking where the verifying analysis usually falls with respect to the ensemble forecast data (arranged in increasing order at each grid point). Note that the first (last) bin is selected if the analysed value is lower (higher) than any of the values in the ensemble.

Since all perturbations are intended to represent equally likely scenarios, this distribution should be flat. Several common problems can be diagnosed from the diagrams:

- a U-shaped diagram is obtained if the spread in the ensemble is typically too small;
- if the spread in the ensemble is too big one obtains an n-shape (highest in the middle);
- if the diagram is asymmetric the model has a bias to one side. An L-shape would correspond to a warm bias for the model.

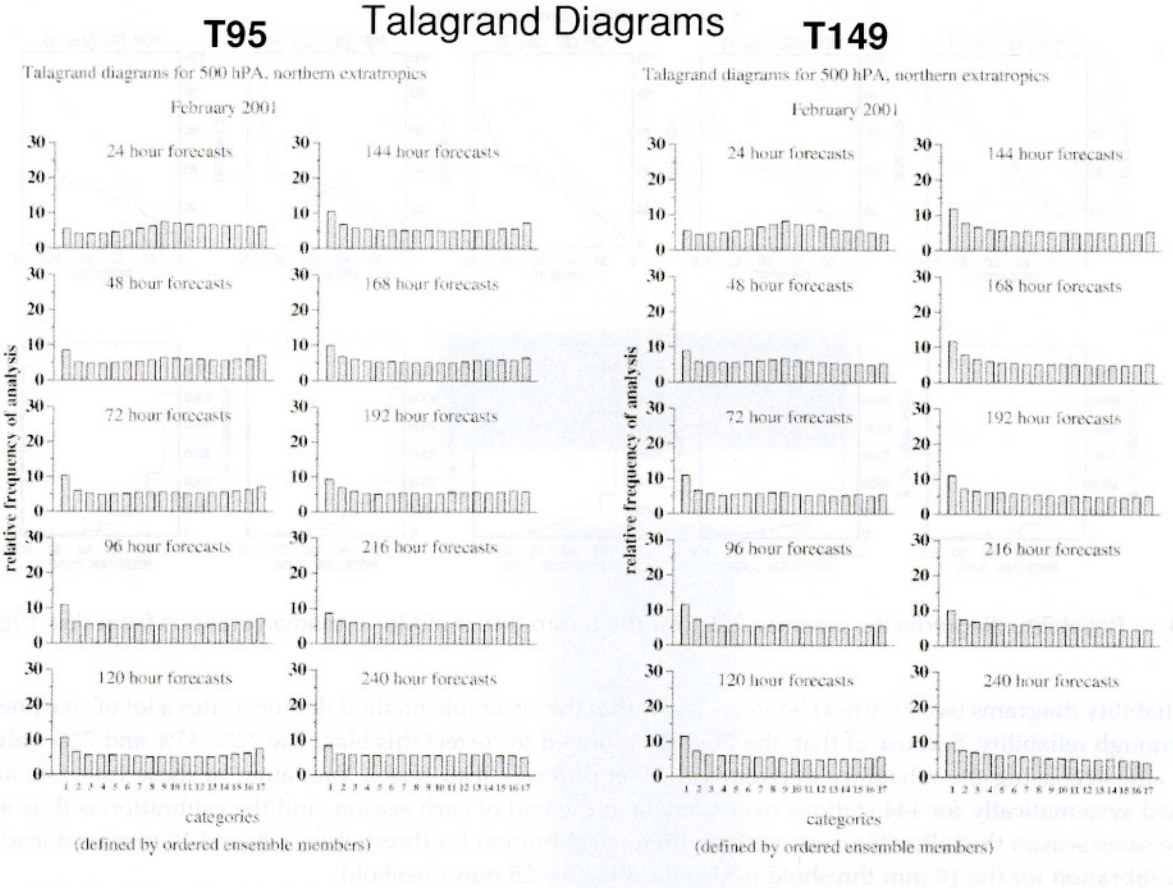


Fig 3: Comparison of ensemble resolutions at T95 ensemble and T149 for Talagrand diagram distributions at 500 hPa over Northern Hemisphere for February 2001 for lead times from 24 to 240 hours.

These diagrams are produced every month. For example, Figure 3 shows diagrams for February 2001 for all the lead times, comparing the T95 (old) to the new T149 (new) resolutions. It is encouraging to see that the diagrams encompass nearly 80% of all the analyses and do not seem to demonstrate any systematic biases.

3.2 *Ten-day mean temperatures*

The ten-day mean temperature forecast is a mean temperature product deduced from the linear regression between 1000-500 hPa thicknesses and surface temperatures. Three class anomaly maps are produced. The improvement of the mean ensemble forecast of the 500 hPa height over the operational model can give hope for a gain over the equivalent product obtained from the deterministic one. This was done for the year of 1998 by *Palany et al* (1999), where forecasts were verified through a 3 x 3 contingency table, from which the Percent Correct (PC) score was calculated. The majority of the months showed an improvement in skill from the ensemble system. Work was pursued to further improve this score, by correcting for the bias present in the forecast and by using the distribution from the ensemble rather than just the mean to produce a probabilistic forecast.

3.3 *Experimental precipitation forecasts*

The ensemble approach is a natural tool for forecasts of the probability of precipitation (POP), where for a sixteen member ensemble, the classes could be defined as follows, for thresholds of 2, 5 10 and 25 mm of precipitation in a 24 hour period:

- below 22%, if 3 members or less forecasts a precipitation amount above the threshold;
- between 22% and 47%, if 4 to 7 members forecast a precipitation amount above the threshold;
- between 47% and 72%, if 8 to 11 members forecast a precipitation amount above the threshold;
- above 72%, if more than 12 members forecast a precipitation amount above the threshold..

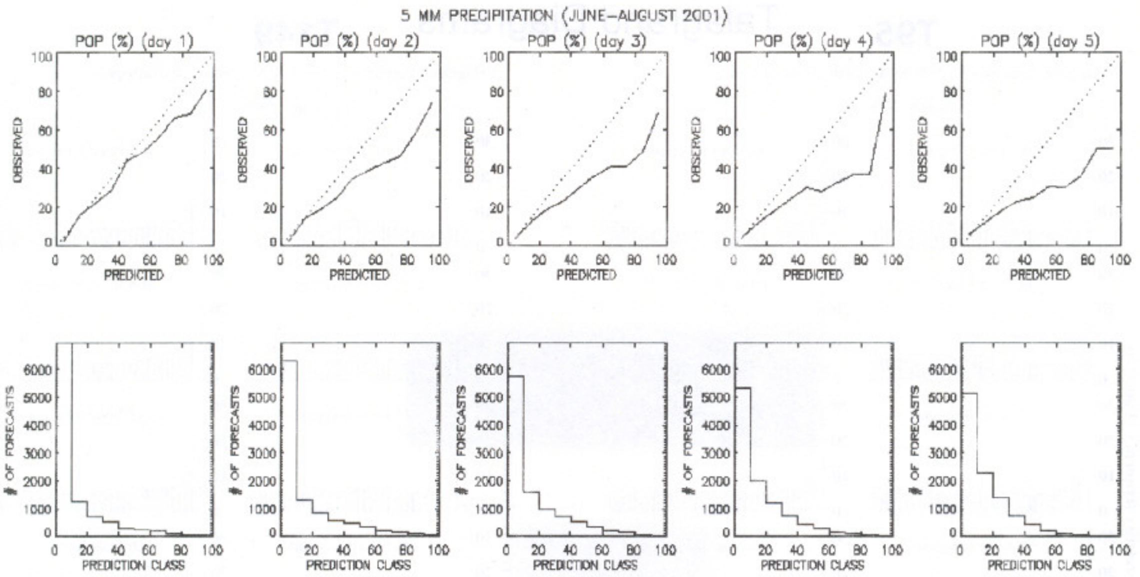
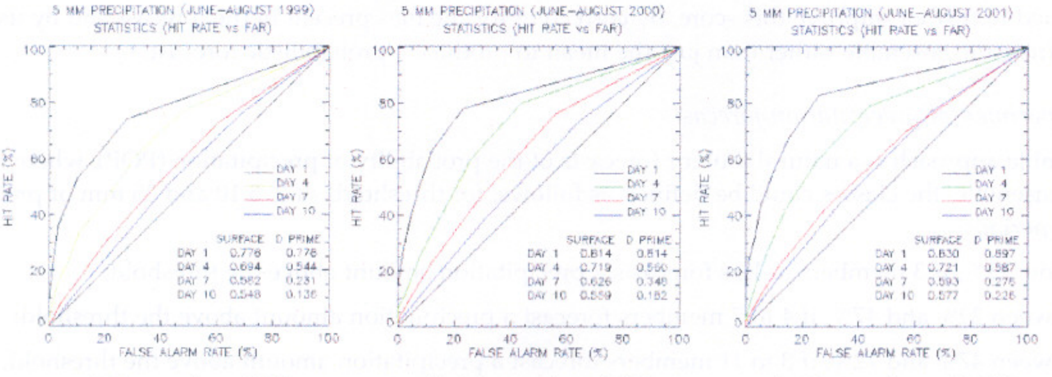


Fig 4: Reliability diagrams for summer 2001 for the 5 mm threshold for Canadian stations from day 1 to day 5

Reliability diagrams (see Figure 4) however show that the ensemble method demonstrates a lot of sharpness, but not enough reliability. Because of that, the POP is calibrated to correct this bias. The 22%, 47% and 72% values are thus adjusted using the reliability diagram values at different lead times. Validation of these forecasts are performed systematically for 144 stations over Canada at the end of each season, and the calibration is thus applied to the same season the following year, with a different calibration for thresholds (2, 5, and 10 mm) and lead times. The calibration for the 10 mm threshold is also used for the 25 mm threshold.

The relative operating characteristic (ROC) curves have been proposed by *Mason* (1982) as a verification measure for ensemble forecasts. In a ROC curve, the hit rate is shown as a function of the false alarm rate. The surface under the curve gives a good indication of the performance of the system, with the value of 0.7 being accepted as the limit below which forecasts are useless.

Figure 5 shows the results for the 5 mm threshold for 4 lead times (day 1, day 4, day 7 and day 10) for 3 summer seasons (June - July -August). Apart from the interannual variations, the improvements in the ROC scores seen in the three figures, correspond to the improvement in the ensemble system over the years: from 8 to 16 members and then increased resolution. The summer verification is especially interesting, with ROC score of .694 at day 4 (a) in 1999, while it is of .719 (b) in 2000 to 0.721 in 2001 (c). The area under the curve of the ROC can also be used to validate the improvement in the ensemble system. This was done during the winter season 2000/2001 (December - January - February) period during which the new ensemble system was tested before implementation to show a better skill at the higher resolution.



a) 8 members T95 b) 16 members T95 c) 16 members T150

Fig 5: Example of a 120 hour calibrated probabilistic QPF using the 16-member ensemble set-up. The maximum QPF for this lead-time and this season (winter) is 42%

4. Confidence index

The Perfect Prog (PP) statistical adaptation system operational at the 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 proxy for a confidence index.

4.1 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 Model Output Statistics (MOS) approach, but in an updateable (UMOS) framework (*Vallée et al*, 1998). The statistical guidance includes spot temperatures at every three-hour interval, 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.

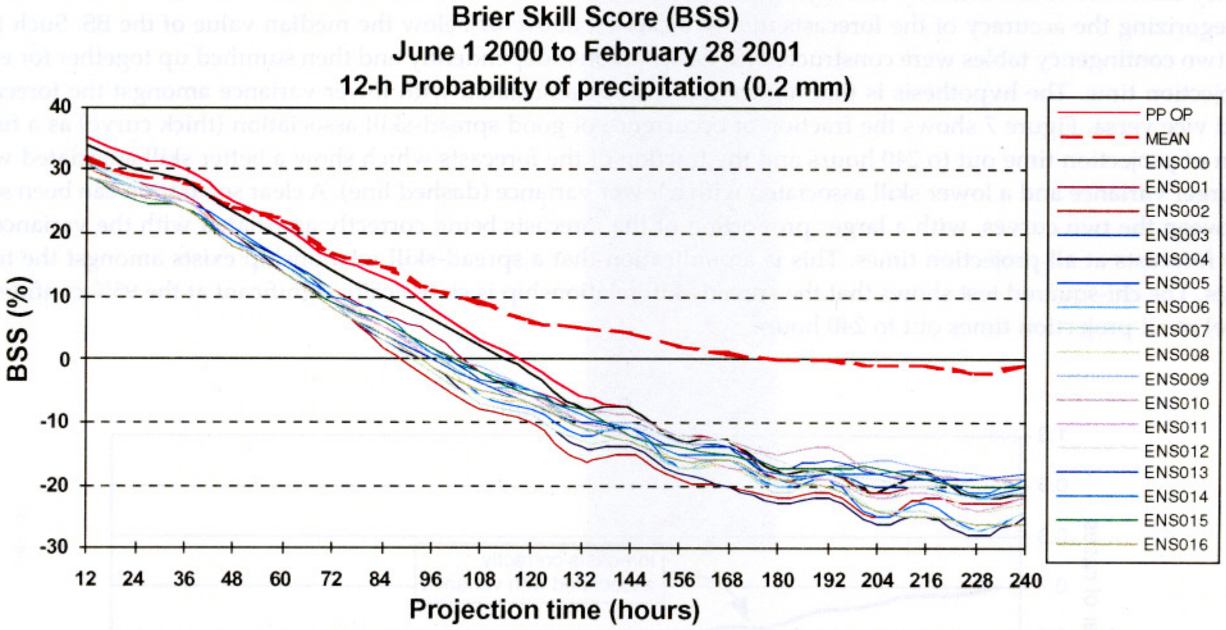


Fig 6: Brier Skill Score (BSS) for the 12-h Probability of precipitation from each of the ensemble members, the control model (ENS000, thick dark line), the operational deterministic model (out to 144-h projection time only) and for the mean of the ensemble forecasts (thick dashed line).

Twelve hour PoPs are forecast for three thresholds, 0.2, 2 and 10 millimetres. 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 6 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 grey), 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 behaviour 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 6 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.

4.2 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 7 shows the fraction of occurrence of good spread-skill association (thick curve) 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). 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.

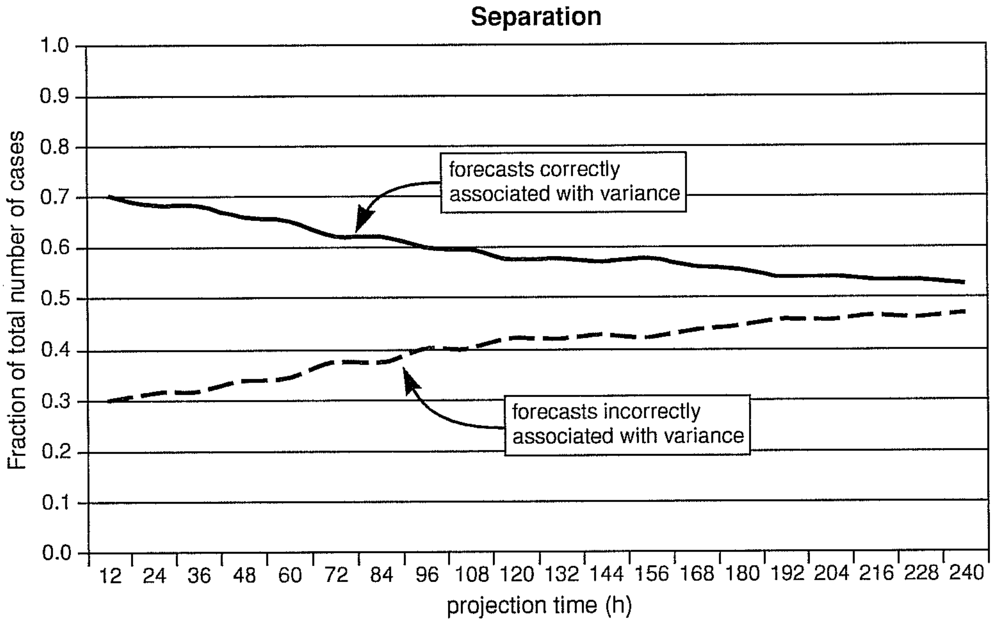


Fig 7: Fraction of the forecasts with a correct association between variance and skill (upper curve). The lower curve shows the percentage of the forecasts with an incorrect association between skill and variance.

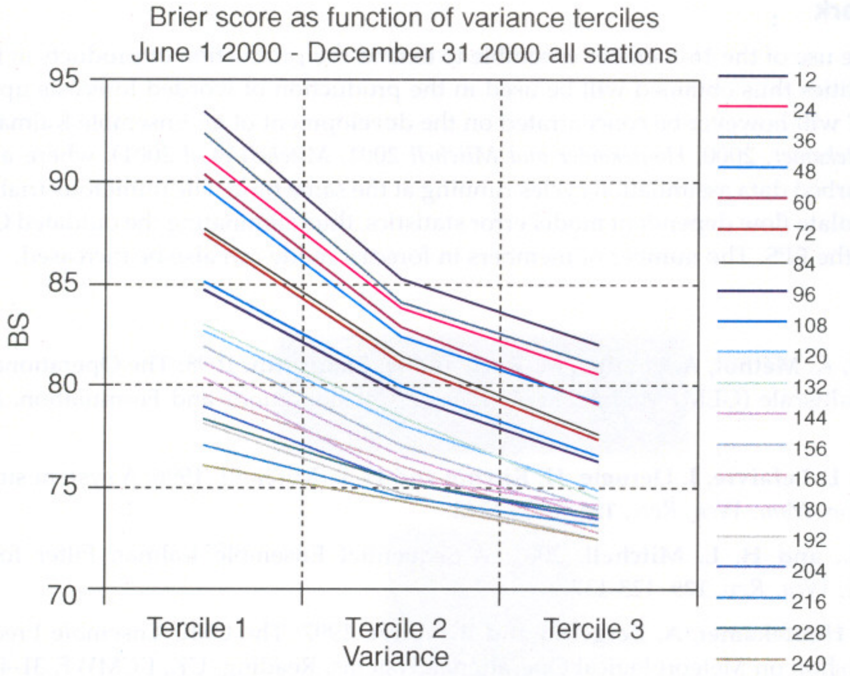


Fig 8: Expected Brier score for each tercile of variance amongst the 12-h PoP forecasts. There is one curve for each 12-h period between the 0- and 240-h projection times. The uppermost curve corresponds to the 0- to 12-h forecasts while the bottom one is for the 228-to 240-h forecasts.

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 skilful 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 8 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 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.

The choice of the statistical 12-h PoP forecasts as a proxy 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 PoPs 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.

5. Future work

There will be more use of the 16-member ensemble to extend the production of products in terms of probabilistic forecasts. Probabilities thus obtained will be used in the production of worded forecasts up to day 10. The main EPS effort at CMC will however be concentrated on the development of an Ensemble Kalman filter (EKF) system (*Mitchell and Houtekamer*, 2000, *Houtekamer and Mitchell* 2001, *Mitchell et al* 2001), where a large number (~100 members) of perturbed data assimilation cycles running at the same time. The numerous trial fields thus produced will permit to calculate flow dependent model error statistics, thus eliminating the outdated Optimal Interpolation presently used in the EPS. The number of members in forecast mode can also be increased.

References

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. Wea. Rev.*, **124**, 1225-1242.

Houtekamer, P. L. and H. L. Mitchell, 2001: A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation, *Mon. Wea. Rev*, **129**, 123-137.

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, UK, ECMWF, 31-44.

Mason, I., 1982: A model for the assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291-303.

McFarlane, N.A., 1987: The effect of orographically excited gravity wave drag on the general circulation of the lower stratosphere and troposhere. *J. Atmos. Sci.*, **44**, 1775-1800.

McLandress, C. and N. A. McFarlane, 1993: Interactions between orographic gravity wave drag and forced stationary planetary waves in the winter Northern Hemisphere middle atmosphere. *J. Atmos. Sci.*, **50**, 1966-1990.

Mitchell, H. L. and P. L. Houtekamer, 2000. An Adaptive Ensemble Kalman Filter, *Mon. Wea. Rev.*, **128**, 416-433.

Mitchell, H. L., P. L. Houtekamer and G. Pellerin, 2001: Ensemble Size, Balance, and Model-Error Representation in an Ensemble Kalman Filter, Submitted to *Mon. Wea. Rev.*

Moorthi, S. and M. J. Suarez, 1992: Relaxed Arakawa-Schubert: A parametrization of moist convection for general circulation models. *Mon. Wea. Rev.*, **120**, 978-1002.

Palany, P., G. Richard, R. Verret, L. Lefaivre, G. Pellerin and P. Houtekamer, 1999: Ten day Temperature anomaly forecast based on ensemble prediction system, Proceedings of the 17th Conference on Weather and Forecasting, Denver, Co, 13-17 Sept. 1999.

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

Rutherford, I. D., 1972: Data assimilation by statistical interpolation of forecast error fields. *J. Atmos. Sci.*, **29**, 809-815.

Talagrand, O., R. Vautard and B. Strauss, 1997: Evaluation of probabilistic prediction systems. Proceedings, ECMWF Workshop on Predictability, 1997.

Vallée, M. and L. J. Wilson, 1998: The new Canadian updateable MOS forecast system. Preprints 14th AMS Conference on Probability and Statistics in the Atmospheric 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.

Wagneur, N., 1991: Une évaluation des schémas de type Kuo pour le paramétrage de la convection, *Msc Thesis*, **UQAM**, 76 pp.