3.3 PROGRESS IN THE DEVELOPMENT OF THE CANADIAN UPDATEABLE MODEL OUTPUT STATISTICS (UMOS) SYSTEM

Laurence J. Wilson* and M. Vallée
Meteorological Service of Canada, Dorval, Québec

1. INTRODUCTION

Recently, the operational regional model GEM was changed at the Canadian Meteorological Center. This time, for the first time ever as far as we know, the new model was implemented along with MOS equations that are partially based on output from the new version of the model. This was made possible by the Canadian UMOS system, which is designed specifically to allow the incorporation of new model output quickly into MOS equations, and to ensure the rapid and continuous adaptation of the weather element forecasts to changes in the statistical characteristics of the model output.

The model change took place on September 11, 2001 (an inauspicious day), and we were able to make use of data from the parallel run period of about six weeks prior to the change for the MOS development. When the model became operational, forecasts for spot temperature, POP and wind direction and speed were issued, using equations based on about 35 cases from the new model and several hundred cases from the previous version of the model, which had been in operation for nearly three years. The application of a weighting scheme described below ensured an early response to the new model data by assigning relatively greater weight to new model cases.

This paper summarizes the design of the Canadian UMOS system, and presents the most recent verification results as obtained since the implementation of the new model.

2. UMOS SYSTEM DESIGN

The basic idea of updateable MOS was proposed by Ross (1987, 1989). Although our system differs in significant ways from the system developed by Ross, the fundamental principles are the same. There are two main characteristics that distinguish UMOS from a more standard MOS system, as for example Glahn and Lowry (1972). First, an updateable system includes a capability to prepare the data for input to statistical algorithms such as linear regression in near real time. In our system, we store the observation and forecast data in the form of sums of squares and cross-products matrices (SSCP) instead of raw predictor and predictand values. The matrices are updated daily with the latest data. This means that new equations can be run at any time, and will be guaranteed to take account of the latest model output predictor values.

The second distinguishing feature of a UMOS system is that it is able to treat the cases that make up the dependent sample with different and controllable weights. Thus, data from a new model can be assigned relatively higher weight to ensure a rapid response to the new model. Data from the old model is retained at gradually decreasing weights to maintain statistical stability of the equations and to ensure a smooth transition from dependence on the old model to dependence on the new model.

We store data in the form of SSCP matrices for over 700 Canadian locations, one matrix for each projection time, for each of two seasons, for each run time (00, 12 UTC), for each model, and for each of 5 predictands: 3 h spot temperature, 6 h POP, 3 h wind speed, and U and V components of the wind. All predictors are from the regional GEM model, covering projections out to 48 h. Whenever the model changes, we start accumulating a new set of SSCP matrices.

Equations are updated approximately weekly, by rerunning the multiple linear screening regression on the latest SSCP matrices. The advantage of a weekly update cycle is that it allows the equations to change in small increments, so that forecasts adjust smoothly to the bias characteristics of the new model. In preparation for equation redevelopment, the SSCP matrices from new and old model are combined using weights that vary according to the number of cases of new and old model that are available.

The parameters of the weighting scheme were chosen according to the following principles: 1. We need at least 30 cases from the new model before beginning to use the new model data; 2. A sample size of 300 or more is needed to develop statistically stable equations (Carter, 1986; Wilson, 1985); 3. The effect of the old model data should be phased out as the sample size from the new model approaches 300; and 4. Data from the new model should be emphasized so that the equations respond to the new model characteristics as quickly as possible. With these principles in mind, we chose minimum new model sample sizes of 30, 35 and 30 respectively for temperature, POP and wind. The sample size needed for complete dependence on the new

*Corresponding author address: Laurence J. Wilson, RPN, 2121 N. Service Road, Transcanada Hwy, Suite 500, Dorval, Québec, H3Y 2N2, e-mail: lawrence.wilson@ec.gc.ca

model is set at 300, 350 and 325 for temperature, POP and wind respectively. For POP, we found by experiment that it took longer for the equations to stabilize, so set the thresholds higher.

Figure 1 shows the weighting scheme for precipitation, for the old model data. The maximum weight is 1.6 for new model data, and the old model weights are set so that the effective sample size is 300. That is, the higher the sample size from the old model, the lower the weight for each case. As long as the new and old model

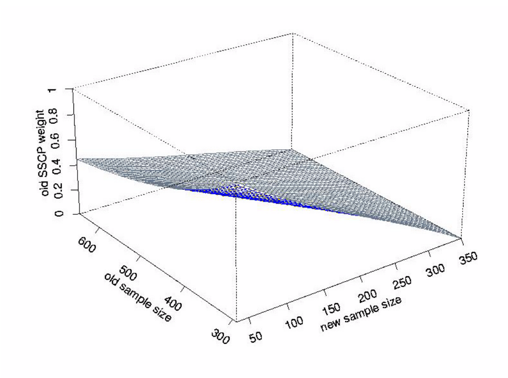

Figure 1. Plot of the weighting scheme for old model data for precipitation equations as a function of old and new model sample size.

samples total more than 300, we can drop the data from the oldest model version at the next model change. Otherwise, we keep model data from more than one version of the old model, weighting it equally.

The impact of the weighting scheme was tested following a major model change in 1998 (Wilson and Vallée, 2002). The tests on all three predictands showed that the equations responded rapidly to the model change, often showing a complete change in the selected predictors with the use of relatively little data from the new model. Once the initial changes occurred, the coefficients would change each week with the addition of more new model data, and the order of the second and subsequent predictors might be permuted, but major changes in the equation structure were rare. We also found that the total reduction of variance changed little during the transition period.

We deal with the seasonal variations in the statistical properties of the development samples in three ways. First, we develop separate equations for warm (April 23 to November 6) and cold (November 7 to April 22) seasons. Second, we offer the sun angle as a variable in the equation development; and third, we apply a weighting scheme to 6-week spring and fall transition periods, blending the winter and summer SSCP matrices prior to equation development. This is intended to ensure a smooth transition between winter and summer regimes, and seems to work well for this purpose (Wilson and Vallée, 2002).

One might expect equations developed from samples representing two or more versions of the model to contain unexplainable variance due to systematic differences in the statistical properties of the output from the different model versions. This might lead to lower quality forecasts during a transition period. Independent sample comparisions of the UMOS forecasts have been conducted following the 1998 model change, which showed that the forecasts remained superior to the older operational perfect prog (PPM) forecasts for all three elements. Wind forecasts proved most sensitive to blended samples; that may have been caused by changes in the model resolution, which would strongly affect predictors involving pressure gradients or wind shear.

3. EARLY RESULTS FROM TESTS FOLLOWING THE FALL 2001 MODEL CHANGE.

Although the September, 2001 change involved only one model component, it was a major change from the point of view of surface weather element forecasts: The “force-restore” land surface module was replaced with the so-called ISBA (interactions, surface, biosphere, atmosphere) surface modelling scheme, along with changes to the assimilation of soil variables in the model. The new scheme involves six new prognostic variables that might be of use to future versions of UMOS.

Based on subjective evaluations during the six week parallel run period just prior to the implementation, the new surface scheme appeared to handle the diurnal surface temperature variation better than the old model (stronger cycle). The new scheme appeared to be biased slightly dry, reversing the bias of the old model. Stability indices from the new model also were significantly different than those from the old model. Finally, differences were noted in the precipitation forecasts between the new and old models, but precipitation is too variable to draw any firm conclusions about the relative accuracy of these forecasts. Differences in the msl pressure and 500 mb height fields were small.

In terms of the predictors used by UMOS, we would expect to see significant changes in the temperature and precipitation forecasts and perhaps less so in the wind forecasts. For our operations, it is most important to ensure that the forecasts based on blended samples retain enough quality to be superior to the PPM forecasts, which were operational throughout the 1990s, and were recently replaced by UMOS. The results shown below are compared to results for the same period from the PPM system. The PPM temperature equations are described in Brunet (1987), the PPM POP equations are described in Verret (1987), and the PPM winds are described in Sarrazin and Wilson (1989).

As mentioned above, we were able to use data from the parallel run, so that equations developed at the time of the implementation contained data from the new model. The results shown below are based on about three weeks of independent data for 222 Canadian stations. The independent sample was accumulated by running the latest UMOS equations operationally for a week after their development, and saving the forecasts and corresponding observations for verification purposes. As a result, we are able to summarize not only the performance of the UMOS equations, but also the performance of the predictors from the model rather quickly after implementation of the model, which provides early feedback to the model developers. These results, obtained only three weeks after the implementation, constitute the first summary verification of the new version of the model.

Figure 2 shows the average error (bias) for temperature forecasts as a function of projection time, for the three weeks of independent data. These are compared to the direct model forecasts (new model) and PPM forecasts, which also used data from the new model. This figure indicates the both the model and PPM forecasts are negatively biased (too cold), and that the bias is

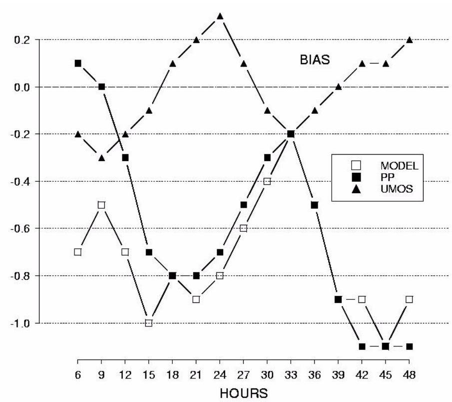

Figure 2. Surface temperature bias as a function of projection time for the last three weeks of September, 2001, for 220 Canadian stations. Sample size is about 3600 cases. Curves are for UMOS (triangles), perfect prog (black squares), and direct model output (open squares).

greatest near maximum temperature time, 21 and 45 hours after the 00 UTC model initialization time. The PPM forecasts are biased similarly to the model forecasts, except for the shortest ranges. This is expected and indicates the upper air predictors used in the PPM equations are biased consistently with the surface temperature forcasts from the model. The MOS forecasts show smaller bias, in the opposite direction. Thus it would appear that the UMOS has overcorrected the bias slightly. Since the UMOS equations are still largely based on old model data, where the diurnal biases were larger, this tendency would be expected during the transition period before the equations have had a chance to adjust to the smaller diurnal bias cycle in the new model. The direct model output (DMO) verification, however, suggests that the new model scheme has not really corrected the temperature bias as well as might have been expected from the parallel run results.

Figure 3 shows the reduction of variance (RV) with respect to the sample climatology of the three week test period. In the calculation of the RV, the variances have been computed with respect to the independent sample mean for each station, thus the bias has been removed. Both the PPM and the UMOS forecasts improve on the DMO, and the UMOS forecasts are superior to the PPM forecasts by about 3% at the early projections, narrowing to 2% by the 48 h projection. The RV is highest in the

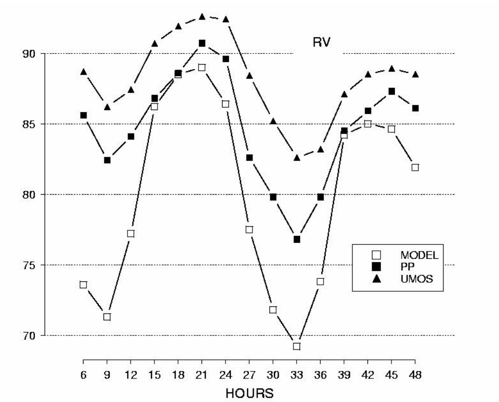

Figure 3. Same as Figure 2, but reduction of variance with respect to sample climatology.

daytime hours, which confirms the parallel run result that the new model does a better job of predicting maximum daytime temperatures. Both the PPM and UMOS forecasts show greatest improvements over the model forecasts where the latter are weakest, near minimum temperature time. For MOS, the RV improvement is as high as 13%. In summary, for temperature, UMOS has slightly overcorrected the model temperature bias and has explained two to three percent more variance than the PPM forecasts.

The performance of the POP forecasts during the three week test period is demonstrated by figures 4 to 7. Figure 4 shows the Brier score as a function of projection time for all 220 stations. In interpreting these results, it should be remembered that the model forecasts are deterministic (categorical) while the PPM and UMOS forecasts are probabilistic. The PPM forecasts have also been post-processed to optimize their characteristics.

Both PPM and UMOS improve on the model forecasts,

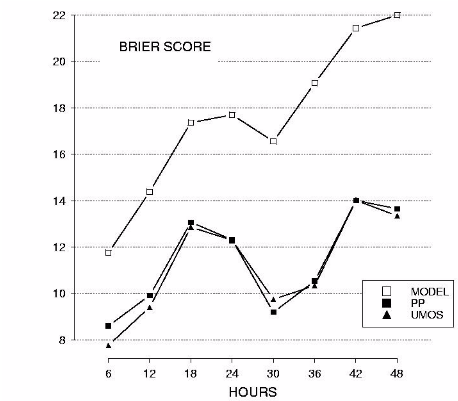

Figure 4. Brier scores (multiplied by 100) for the new model (open squares), PPM (black squares) and UMOS forecasts (triangles) for 220 Canadian stations for three weeks in September, 2001, as a function of projection time. Sample size is about 3600 cases.

though much of the advantage may be due to the greater flexibility available to probability forecasts compared to categorical forecasts. The UMOS forecasts are only very slightly superior to the PPM forecasts, except at 30 h. The improvement of UMOS over PPM is more noticeable at the shortest ranges. Given that the model precipitation forecasts and predictors are sharper (more spatial detail) than the predictors used in the PPM equations,

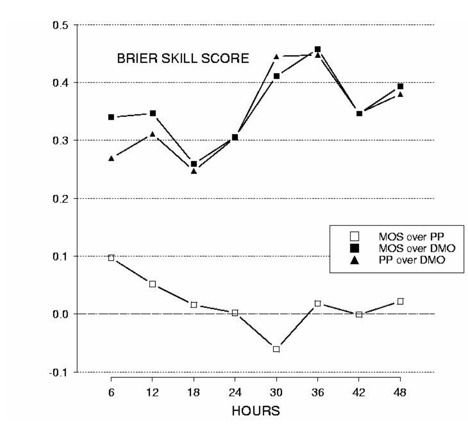

Figure 5. Brier skill scores for MOS vs. PPM (open squares), MOS vs. DMO (black squares) and PPM vs. DMO (triangles) for POP forecasts. Same sample as figure 4.

this result would suggest that the additional detail is useful in a statistical sense only for day 1 forecasts.

Figure 5 shows Brier Skill scores for the three sets of POP forecasts, calculated using one of the forecasts as a reference standard. Both MOS and PPM show significant skill with respect to the model forecasts, but the most interesting curve is for the UMOS forecasts against the PPM forecasts as a standard. UMOS skill is positive with respect to PPM out to 18h, then there is little difference in skill overall. It should be noted that the UMOS POP forecasts do not use the model's precipitation as predictors. As revealed by earlier tests, these fields proved to be too statistically unstable (not stationary) to provide good forecasts, even though they were sometimes chosen as predictors. These predictors are under investigation. The predictor sets used by UMOS for precipitation are therefore closer to the predictors used by the PPM equations.

Figure 6 shows reliability tables for the POP forecasts for 12 to 18 h and 42 to 48 h POP. Both UMOS

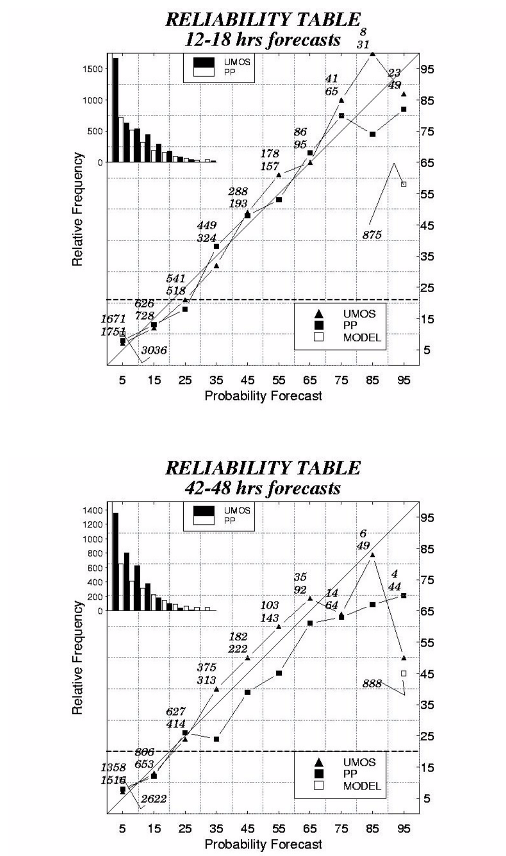

Figure 6. Reliability tables for POP forecasts for 12 to 18 h (above) and 42-48 h (below). Sample sizes in each bin are plotted next to the points, with the UMOS sample size on top and the PPM below. Categorical model forecasts are represented by points in the lowest and highest bins (open squares).

and PPM forecasts are generally quite reliable, though the MOS forecasts are more reliable at the longest forecast range. For this short verification period, there were relatively few occasions where either method attempted to forecast over 80% POP; thus the reliability table values are noisy for the highest probability bins. Judging from the distribution of forecast probabilities, the PPM forecasts are sharper at the longest forecast range, while the UMOS forecasts are both more reliable and sharper at the shortest range. Differences are not large, however. The model's categorical forecasts of precipitation occurrence are about 57% correct at 12-18 h and 45% correct at 42-48 h. Case-by-case examination of the data from the parallel run suggested that the new model might produce better convective precipitation forecasts than the old model. There does not seem to be clear evidence to support this as yet from these early verification results.

Figure 7 shows bias results for wind speed. These results are similar to the temperature results: Both the model and PPM tend to underforecast wind speed overall, and UMOS seems to have corrected the bias. Underforecasting is greatest during the day when wind speeds are typically the highest. Once again, the fact that the PPM bias is similar to the DMO bias shows that

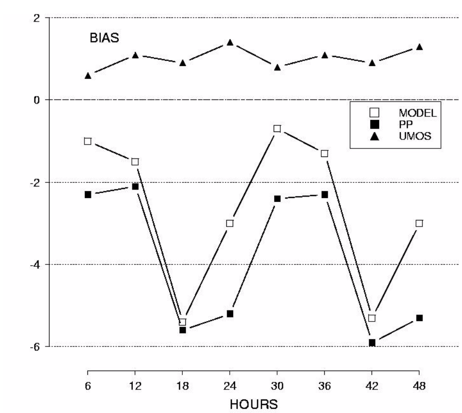

Figure 7. Bias of wind speed forecasts in kmh^{-1} as a function of projection time for DMO (open squares), PPM (black squares) and UMOS forecasts (triangles). Independent sample of about 3000 cases, 222 Canadian stations, Sept, 2001.

the model's surface wind forecasts are consistent with the upper air predictors used by the PPM equations. Also similarly to temperature, the UMOS development on a blended sample of new and old model data, along with an independent sample of only new model data might be a reason for the slight overcorrection of the bias. The bias in the UMOS wind forecasts is quite small, about 1 kmh^{-1}, which is consistent with the fact that the pressure and upper air height fields were not expected to change significantly following the model change.

Figure 8 shows the root mean square error (RMSE) for wind speed. UMOS errors are the lowest, ranging from 6 kmh^{-1} to a little over 7 kmh^{-1} at 48 h. The diurnal cycle in the RMSE of the PPM and DMO forecasts is most likely due to the bias component.

The characteristics of the wind forecasts can be fur-

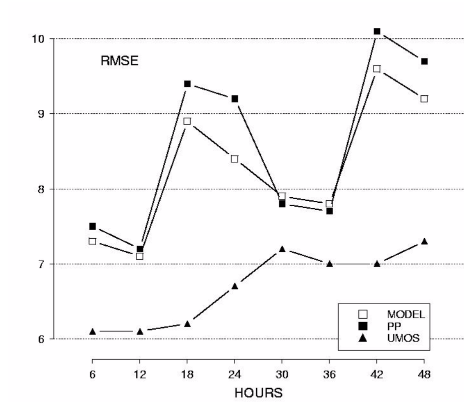

Figure 8. Same as figure 7, but RMSE.

ther evaluated by means of contingency tables, Tables 1,2 and 3 are for 18 h forecasts of wind speed. Categories were chosen to be those significant for operational forecasting; the thresholds shown in the tables are the upper bounds of each category. The DMO and PPM forecasts are of spot wind speed values, while the UMOS forecasts are for the maximum speed over three consecutive hours centered on the valid time. Each was evaluated consistently with its predictand definition. A scan of the figures in the tables suggests that the DMO and PPM tend to underforecast winds, consistent with figure 7, especially the higher categories, by about one category. UMOS, on the other hand, is more balanced overall, but tends to overestimate the lowest two categories and mildly underestimate the higher categories.

The percent correct is 39, 36 and 53 for DMO, PPM and UMOS respectively. Heidke skill scores with respect to chance are 19, 15 and 34 respectively. Thus, UMOS scores considerably higher than the other two methods for these categorized forecasts. We have to admit, though, that the overforecasting of the light wind (first) category by UMOS does have the effect of causing our automated forecast generator to mention winds too often in text forecasts. Even with a more accurate forecast, some adjustments have to be made to optimize the interpretation of the forecasts.

4. SUMMARY

The Canadian UMOS system is currently going through its first model transition period since becoming operational about a year ago. For the first time ever, we were able to make use of the data from a six week parallel run to enable us to implement MOS equations which use data from the new model along with the model implementation. This paper describes the first verification results from the blended MOS equations, based on three weeks of independent data for 222 Canadian stations.

Table 1: Contingency table for wind speed forecasts - DMO, September, 2001.

OBS \ FCST (kmh^{-1})	8.96	15.0	25.0	40.0	60.0	>60
8.96	384	60	12	0	0	0
15.0	488	298	96	6	0	0
25.0	255	452	395	73	1	0
40.0	37	114	314	164	9	0
60.0	6	3	15	30	12	2
>60	0	0	0	0	5	0

Table 2: Contingency table for wind speed forecasts - PPM, September, 2001

OBS \ FCST	8.96	15.0	25.0	40.0	60.0	>60
8.96	320	109	24	3	0	0
15.0	426	359	98	4	1	0
25.0	283	505	356	31	1	0
40.0	32	184	299	112	11	0
60.0	0	4	19	33	12	0
>60	0	0	0	1	2	2

Table 3: Contingency table for wind speed forecasts - UMOS, September, 2001

OBS \ FCST	8.96	15.0	25.0	40.0	60.0	>60
8.96	96	279	77	4	0	0
15.0	50	408	414	16	0	0
25.0	12	164	832	166	2	0
40.0	3	22	235	364	14	0
60.0	0	0	10	34	24	0
>60	0	0	0	0	2	3

The performance of the UMOS equations is clearly superior to DMO for all three predictands, temperature, wind and POP, and is superior to the older PPM forecasts for temperature and wind at all projection times. For POP, the accuracy of the UMOS forecasts is about the same as the PPM forecasts (based on the Brier score) after day 1; UMOS shows slight superiority over PPM for the first 18 h of the forecast.

These results seem to support the case study results from the parallel run period with respect to the diurnal cycle in temperature bias, but the overall cold bias of the old model has apparently not yet been eliminated. The MOS equations have been able to correct for the bias in both temperature and wind speed. For POP, more data will be needed before any conclusions can be drawn about changes in the quality of precipitation forecasts with the new model.

All of the equations described here use multiple linear regression. A multiple discriminant analysis module is under development for use with multiple-category predictands such as cloud amount. Once this is ready, we will extend UMOS to the full set of surface weather element predictands.

5. REFERENCES.

Brunet, N., 1987: Development of a perfect prog system for spot time temperature forecasts. CMC Technical Document No. 30, 55 pp. [Available from Environment Canada, Dorval, Quebec]

Carter, G., 1986: Towards a more responsive statistical guidance system. Preprints, *11th Conference on Weather Forecasting and Analysis*, Amer. Meteor. Soc., Boston, Mass., 39-45.

Glahn, H.R. and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.* **11**, 1203-1211.

Ross, G.H., 1987: An updateable model output statistics scheme. World Meteorological Organization, Programme on short- and medium range weather prediction, PSMP Report Series No. 25, 25-28.

Ross, G.H., 1989: Model output statistics - an updateable scheme. Preprints, *11th Conference on Probability and Statistics in Atmospheric Sciences*, Amer. Meteor. Soc., 93-97.

Sarrazin, R., and L.J. Wilson, 1989: A tuned perfect prog wind forecast system for Canada. Preprints, *11th Conference on Probability and Statistics in Atmospheric Sciences*, Amer. Meteor. Soc., Boston, Mass., 50-55.

Verret, R., 1987: Development of a perfect prog system to forecast probability of precipitation and sky cover, CMC Technical Document No. 29. [Available from Environment Canada, Dorval, Quebec]

Verret, R., 1988: Post-processing of statistical weather element forecasts. CMC Monthly Review, 7, 5, 2-16. [Available from Environment Canada, Dorval, Quebec]

Wilson, L.J., 1985: Application of statistical methods to short-range operational weather forecasting. Preprints, *Ninth Conference on Probability and Statistics in Atmospheric Sciences*, Amer. Meteor. Soc., 1-10.

Wilson, L.J. and M. Vallée, 2002: The Canadian updateable model output statistics (UMOS) system: Design and development tests. Accepted for publication in Weather and Forecasting.