6.3 STATISTICAL POSTPROCESSING OF OPERATIONAL NWP OUTPUT – A CANADIAN RETROSPECTIVE AND PERSPECTIVE

Laurence Wilson *
Environment and Climate Change Canada, Montreal, Canada (Retired)

Barbara Casati
Recherche en Prévision Numérique, Environment and Climate Change Canada, Montreal, Canada

1. INTRODUCTION

Environment and Climate Change Canada (formerly Environment Canada) has been involved in the development and operational use of statistical methods for post-processing the output of numerical weather prediction (NWP) models for more than 50 years. Throughout that period we have been motivated by the knowledge that: 1) Statistical post-processing is inexpensive in computing terms (at least compared to modelling); 2) There is the potential to significantly improve on the quality of surface-based direct model output (DMO) forecasts for specific locations; and 3) Statistical methods have the flexibility to format the forecast output in ways that are meaningful and useful in the preparation of forecasts.

Over the years, we have been heavily influenced by research and development activities in the US, Our early efforts followed the Perfect Prog formulation (PPM) (Klein and Lewis 1970), while the landmark introduction of Model Output Statistics (MOS), (Glahn and Lowry 1972) encouraged us to pursue that development direction as soon as operational computing power would allow it. We also picked up on the Generalized Equivalent Markov method for very short range forecasts (Miller 1981), (Miller and Best 1981), and more recently, we have been converting our main operational MOS system to forecast at grid points.

There are two significant ways in which statistical post-processing diverged from the experiences in the U.S. First, we have freely experimented with a variety of statistical methods, finding that alternatives to linear regression (MLR) could work very well in applications to the more episodic or categorical weather variables such as precipitation, cloud amount, and precipitation type. Thus we have MLR, multiple discriminant analysis (MDA), classification and regression trees (CART), and, for ensemble post processing, Bayesian model averaging (BMA) in our operational toolbox of statistical methods.

Second, we developed and implemented an updating version of MOS (or “dynamic MOS” as it is sometimes called) before it was done in the US. That was successful both for re-establishing MOS forecast products and as a toolbox for development. It has run .for 20 years now.

While most of the statistical post-processing products developed in Canada were designed and intended for implementation into national operational forecasting at the Canadian Meteorological Centre (CMC), the 6 large regional centres were interested in running post-processed forecasts as well. Therefore, some methods were developed to help the regional centres with specific forecasting problems.

Organizationally, the main contributors to the operational statistical interpretation suite of products in Canada were in two groups, both staffed to the level of about 3 scientists typically. The group in CMC, led at first by Nathan Yacowar, then later by Richard Verret, began early, in the late 1960s and continues to the present day, but is now subsumed into a larger development division at CMC with a much broader mandate. The other group began in 1979 in Meteorological Research Division (MRD) and was headed by Laurence (Laurie) Wilson. Within research, this group was mandated to carry out more research-oriented projects in statistical interpretation, and to facilitate the implementation of the resulting new products into operations either at CMC or in the regional offices as appropriate. The “special products” described below were developed by the MRD group. The MRD group was based in ECCC

Corresponding author address: Laurence Wilson, 448 Claremont Ave, Westmount QC, H3Y 2N2; e-mail: laurence.wilson@sympatico.ca headquarters in Toronto until the mid 1990s then moved to Montreal and was absorbed into the modelling research group Recherche en Prévision Numerique (RPN).

Up until 1978, when the MRD group was formed, all the national statistical post-processing development was implemented into CMC operations. The two groups, one in Montreal, one in Toronto, operated concurrently until 1995, when the MRD group was moved to Montreal to join RPN.
The relationship between the two groups was at times competitive, which was seen as a positive force by the management of the day, and usually cooperative. The MRD group, with a research mandate, would explore alternative statistical methodology such as MDA and CART, or concentrate on different predictands, or produce products for regional application, while the CMC group would concentrate on developing and upgrading products for the most important weather elements, more closely following the methods of the US statistical post-processing system. In those years, statistical post-processing was generally viewed as a valuable addition to the weather forecast services.

The paper is organized as follows. We begin with the story of the development of the national statistical forecast guidance system, from the beginning until the present day. This is followed by a description of some of the special products developed for operational use. Section 4 describes the current operational system and current development activities, and finally section 5 presents a perspective on current trends and issues related to statistical post-processing.

As a notation convention, all the statistical products are identified by their formulation (classical, perfect prog or model output statistics) and then by their statistical method (MLR, MDA, CART, analogue).

2. A BRIEF HISTORY OF OPERATIONAL CENTRALIZED POST-PROCESSING IN CANADA

The earliest statistical guidance products intended for use by forecasters across the country were for daily max/min temperature. These were PPM-MLR products and became operational at CMC in the late 1960s (Yacowar, 1973).

Figure 1 shows a sample verification result for the PPM max/min temperature forecasts for days 1 to 3, for the 1200 UTC model run. The forecasts are for approximately 12 h, 36 h and 60 h. The verification measure is the reduction of variance, a skill score based on the mean squared error. The reference climatology is a two-month average. These results show the usual seasonal cycle, with best quality during the winter time. A slight upward trend is visible over the 9 year period, which suggests that the model change from a filtered equations baroclinic model to a primitive equations spectral model in 1976 has improved the quality of the model output. Accuracy levels are apparently not high by today's standards!

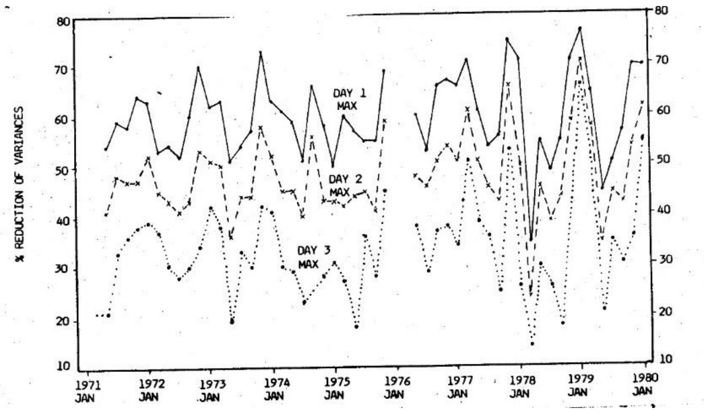


Figure 1. Percentage reduction of variance for PPM-MLR Day 1, Day 2 and Day 3 maximum temperature forecasts produced from model output at 1200 UTC on Day 1. Each point represents an average taken over a two-month period. The gap in the curves corresponds to the change from a filtered equations baroclinic model to a primitive equations spectral model. (From Wilson and Yacowar (1980)).

Early probability of precipitation (POP) guidance products were based on the analogue method (Yacowar, 1975), using a 22 year low resolution analysis dataset to obtain the analogue cases. The analogue-based precipitation probability forecasts ran until 1980.

With operational PPM temperature and PPM-analogue forecasts running at CMC, the MRD group started working on another predictand, probability of precipitation amount (POPA). The result was PPM-MDA 12 h precipitation amount forecasts in four categories, <0.2 mm, 0.2 to <2 mm, 2 mm to <10 mm, and 10 mm or more. Tests of the MDA technique revealed that it was capable of producing sharper probability forecasts (greater tendency to predict higher and lower probabilities) than the regression estimation of event probability (REEP – Miller 1964) method, but the forecasts were less reliable as probabilities. An example of these results is shown for the extreme category in Figure 2.

All new products proposed for implementation at CMC must be passed by an implementation committee which makes its decision on the basis of demonstrated improvement over an existing operational product. Thus, the PPM POPA forecasts were compared with existing analogue-based POPA forecasts. The results of such a comparison are shown in Figure 3.

Figure 3 shows that both the REEP and MDA forecasts are superior to the analogue forecasts. REEP forecasts are slightly superior to the MDA forecasts after the first period, reflecting their greater reliability. The sharpness of the MDA forecasts, especially for the extreme category led to their choice as guidance for forecasters. Some overforecasting of the probability of an extreme event serves as a warning or “heads up” to forecasters, which was seen as a benefit to the users.

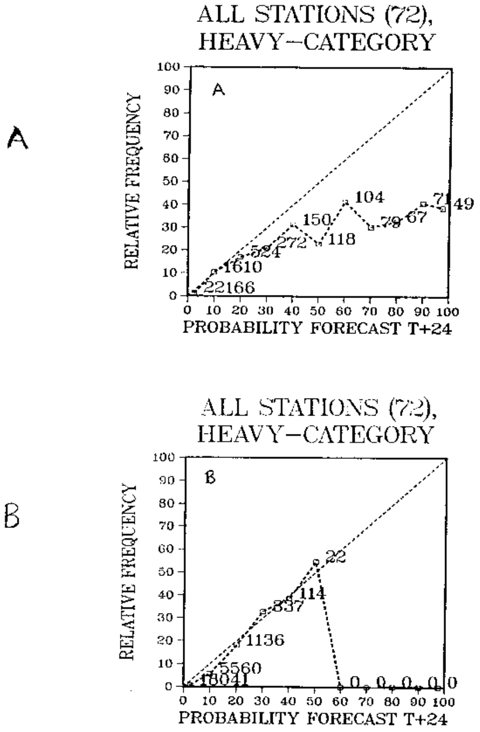


Figure 2. Reliability table verification of (A) MDA and (B) REEP probability forecasts of >10 mm precipitation in 12 h, 24h projection. (From Wilson and Stanski 1983)

During the early 1980s, the available computing power, and more importantly, the ability to store and manage large quantities of data was increasing to the point that we could now consider bringing the acknowledged superior MOS method into practice in Canada. In addition, the 1980s were a period of relatively stable models (The spectral model on a 381 km grid). Both groups competed in the ensuing race, leading to Canada’s first MOS product, surface wind direction and speed, developed by Bill Burrows in MRD, implemented in CMC in 1986, and verified by Wilson et al (1986). Although this was the first MOS product to run operationally from CMC, MOS forecasts for temperature and POP, developed by the CMC group went into operation later the same year. (Yacowar et al 1985).

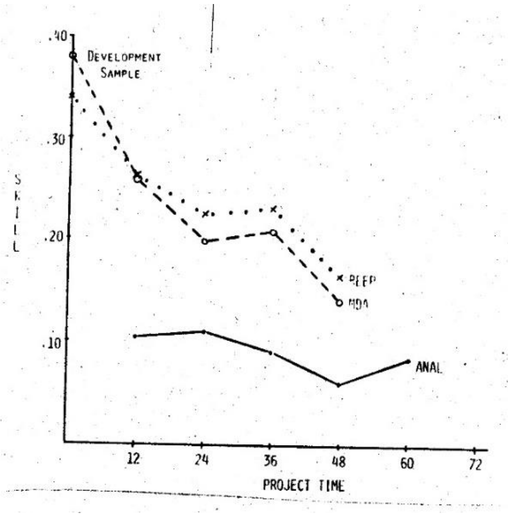
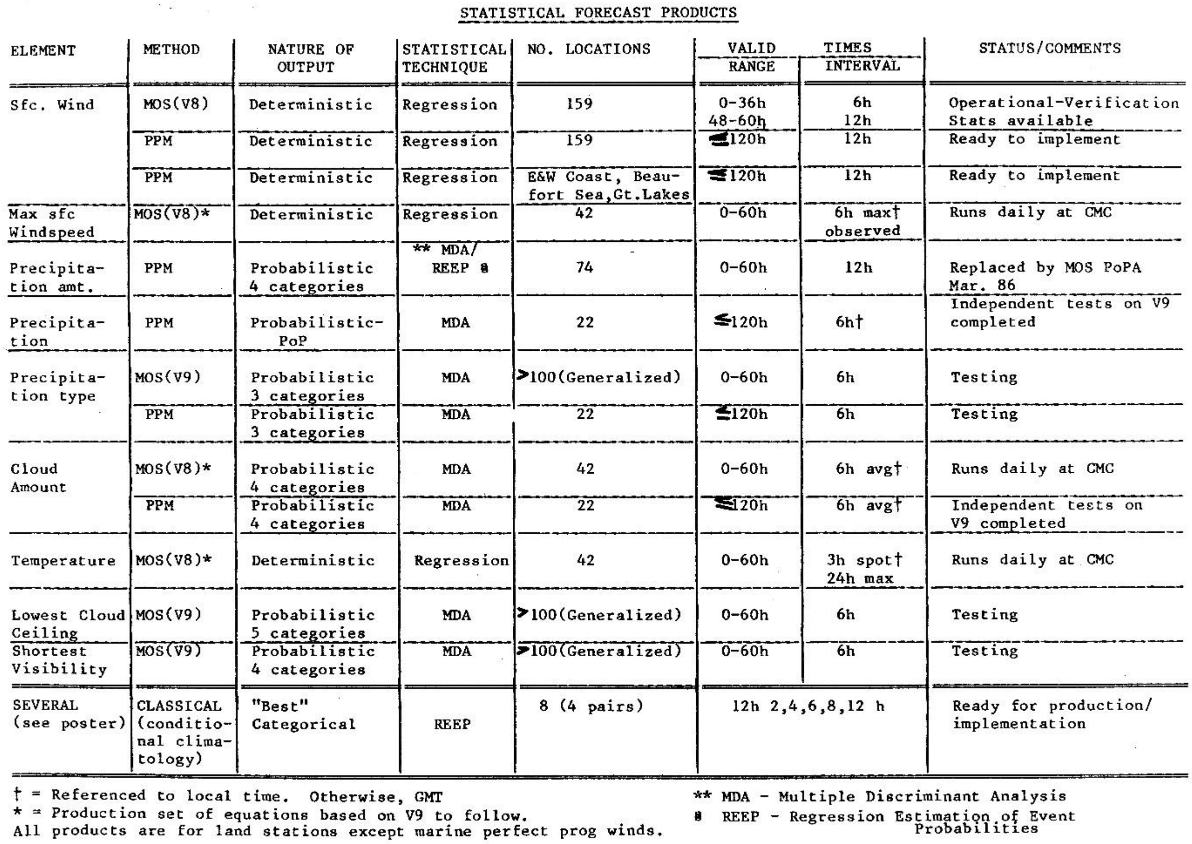


Figure 3. Independent sample verification of REEP, MDA and analogue 4-category POPA forecasts, as a function of projection time, using the rank probability skill score (RPSS). (Wilson and Yacowar 1980)

Activity and productivity in statistical development reached a peak in both groups in the mid to late 1980s, with many new products produced, most of which were MOS. Table 1 gives an idea of the level of activity in statistical interpretation at one point in 1987. Both groups published papers in the refereed literature (Brunet et al, 1988; Wilson and Sarrazin, 1989). However, model changes were becoming more frequent in the late 80s, and, new models were being added to the old. In Canada, the regional finite element model began operations in the late 80s, producing forecast out to 60 h, but the Spectral model was still in use.

Just as MOS products were coming on line, the addition of a second model effectively doubled the development and data storage needs for a MOS system. Together with frequent model changes it became clear that MOS development would not be sustainable, which led to several more studies of PPM vs MOS, to try to justify reverting to PPM, and also to consider ways of reducing the overhead of MOS.

Table 1. A list of statistical products under development in MRD in 1987.



STATISTICAL FORECAST PRODUCTS

ELEMENT	METHOD	NATURE OF OUTPUT	STATISTICAL TECHNIQUE	NO. LOCATIONS	VALID TIMES RANGE	VALID TIMES INTERVAL	STATUS/COMMENTS
Sfc. Wind	MOS(V8)	Deterministic	Regression	159	0-36h 48-60h	6h 12h	Operational-Verification Stats available
	PPM	Deterministic	Regression	159	≤120h	12h	Ready to implement
	PPM	Deterministic	Regression	E&W Coast, Beaufort Sea, Gt. Lakes	≤120h	12h	Ready to implement
Max sfc Windspeed	MOS(V8)*	Deterministic	Regression	42	0-60h	6h max† observed	Runs daily at CMC
Precipitation amt.	PPM	Probabilistic 4 categories	** MDA/ REEP ꞵ	74	0-60h	12h	Replaced by MOS PoPA Mar. 86
Precipitation	PPM	Probabilistic-PoP	MDA	22	≤120h	6h†	Independent tests on V9 completed
Precipitation type	MOS(V9)	Probabilistic 3 categories	MDA	>100(Generalized)	0-60h	6h	Testing
	PPM	Probabilistic 3 categories	MDA	22	≤120h	6h	Testing
Cloud Amount	MOS(V8)*	Probabilistic 4 categories	MDA	42	0-60h	6h avg†	Runs daily at CMC
	PPM	Probabilistic 4 categories	MDA	22	≤120h	6h avg†	Independent tests on V9 completed
Temperature	MOS(V8)*	Deterministic	Regression	42	0-60h	3h spot† 24h max	Runs daily at CMC
Lowest Cloud Ceiling	MOS(V9)	Probabilistic 5 categories	MDA	>100(Generalized)	0-60h	6h	Testing
Shortest Visibility	MOS(V9)	Probabilistic 4 categories	MDA	>100(Generalized)	0-60h	6h	Testing
SEVERAL (see poster)	CLASSICAL (conditional climatology)	"Best" Categorical	REEP	8 (4 pairs)	12h 2,4,6,8,12 h		Ready for production/ implementation

† = Referenced to local time. Otherwise, GMT
* = Production set of equations based on V9 to follow.
All products are for land stations except marine perfect prog winds.

** MDA - Multiple Discriminant Analysis
ꞵ REEP - Regression Estimation of Event Probabilities

An earlier study by Kruizinga and Böttger (1984) showed that the difference between PPM and MOS forecasts was less than earlier thought, for example Klein and Glahn (1974) and this result was supported by Brunet et al (1988), which featured a redevelopment of PPM equations using a longer and more accurate predictor dataset. One result of this CMC based study is that MOS would not be used for medium range categorical forecasts because the forecasts were not sharp enough. Increasing error variance in the model-derived predictors for the medium range projections caused the statistical equations to become “weaker” and forecast more towards the development sample mean. Such behaviour might have led to better verification scores based on quadratic scoring rules, but once again the relative sharpness of the PPM forecasts was preferred as an alert to forecasters. A study by Dallavalle (1988) showed that medium range PPM forecasts were superior to MOS forecasts run on a different version of a model from the one on which they were developed. Additional MOS-PPM comparison studies in the mid 1980s for surface wind (Wilson 1985) and for POP (Sarrazin and Wilson 1987), (Wilson and Sarrazin 1987) found that PPM systems could be competitive in quality with MOS, though different in terms of specific forecast attributes.

After proving that MOS wind forecasts (Burrows 1984) lost quality when run on a new model without redevelopment (Macdonald 1986), the MRD group experimented with retuning PPM wind forecasts to the model, using a smaller sample from the latest model (Sarrazin and Wilson 1989). This work was successful: The tuned PPM forecasts improved on the older MOS wind speed forecasts by about 1.5 knots and the direction forecasts by 7 to 10 degrees. This was accomplished not only by the tuning, but also by the careful selection of the predictors for the perfect prog equations. Operationally, the advantage was that the PPM equations could always be run on a new model until enough data became available to do the tuning. And the problem of supplying forecasts from multiple operational models is simplified because the same PPM equations are used for all models. Tuning makes the equations model-dependent, and the tuning should be redone after each model change. In fact, the tuned wind equations were implemented in 1989 and ran in operations at CMC until replaced by updateable MOS in 2000. Their quality surely must have decreased as the model changed several times during this period, but no complaints were received from forecasters.

By 1989, all the MOS forecasts implemented in the mid-1980s were replaced with PPM forecasts.

Government austerity in the early 1990s ensured that there would be little new development in either CMC or MRD in the national statistical interpretation system, and the reinstated PPM forecasts continued to run at CMC. Although the pace of statistical interpretation R&D slowed in the 1990s, there were significant developments. One special project that occupied the MRD group for a significant time in the early 1990s was the statistical ozone/UV index forecasts, which even drew the attention of the private sector in the US. This effort is described in the next section as a “special product”.

By the mid 1990s, the quality of the PPM-based forecasts was low enough especially in the short range that something for sure had to be done to address the issue of frequently changing models. In MRD, two ideas were tried, the Kalman Filter (KF - Simonsen 1991), a recursive procedure where recent errors are used both to correct the forecast and to update the coefficients of the (linear) model used to generate the corrected forecasts, and updateable model output statistics (UMOS) (Ross 1987, 1989). Some of our studies of the Kalman Filter are described in Vallée et al (1996). However, in the end, UMOS was selected mainly because it could be applied to all elements (the KF couldn’t easily be applied to categorical variables), but also because it would be easier to manage in operations and is more flexible, allowing for the use of many predictor variables without complications. UMOS work was started in 1996, and implemented first for temperature in 2000, and soon after for 6 h POP and wind direction and speed. Cloud amount forecasts based on MDA were added in 2003, and MLR- based marine wind forecasts for buoys were added in 2007.

The Canadian UMOS system, which is still in operation at CMC, is described in Wilson and Vallée (2002, 2003). Consistent with earlier studies, it was found that the higher resolution variables available as predictors in UMOS were useful only in the short range predictions. Beyond 48 h, and except for temperature, it was difficult to improve on our PPM system, which used relatively low resolution analysis data as predictors. It should be noted that PPM equations which use analysed variables as predictors are not model-independent; they carry the climatology of the model which was in use when the analysis was done, especially in data-sparse areas. This might be different from the current model climatology. Therefore there would be advantages to redeveloping PPM equations using analyses based on the current model as a trial field. Better still, if there is a reanalysis dataset available, this should benefit PPM equations even more.

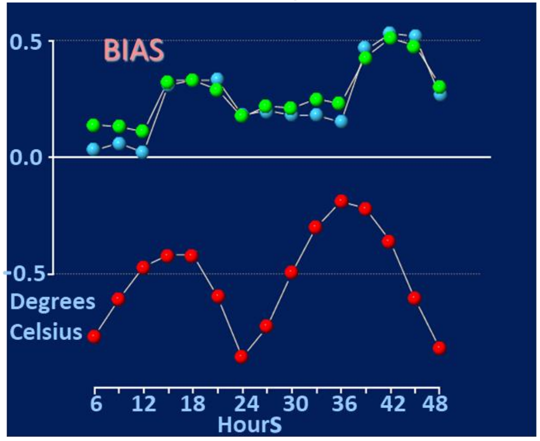


Figure 4. Example of the performance of UMOS for temperature: Bias for DMO (red), "old" model UMOS (blue) and UMOS (new + old) (green).

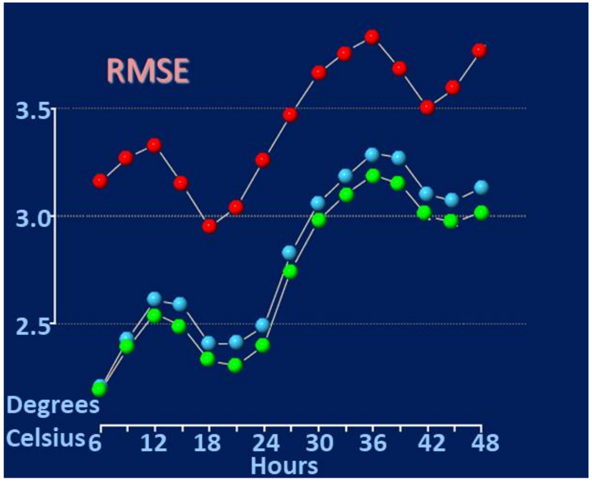


Figure 5. Example of the performance of UMOS for 3 h spot temperature, RMSE in degrees. DMO (Red), "old model" UMOS (blue) and "new + old" model UMOS (green)

Figures 4 and 5 demonstrate the benefits of UMOS. Figure 4 shows the bias for 3h spot temperature forecasts during a development period where the model had been recently changed. UMOS"old" (the blue curve) gives the result for UMOS before the model change. The green curve is the result after some blending of data from the new model. For comparison, the direct model output (DMO) result is shown in red. In this case, the bias has been mostly removed both in the old and UMOS blended versions. The fact that the bias characteristics don't change much with the addition of new model data suggests that the bias didn't change much when the model was changed.

Figure 5 clearly shows that the addition of the new model data (green curve) has improved the RMSE compared to the old model, and both are much better than the DMO. Furthermore, these results are available to the modelers very soon after the model change since the UMOS blending of old and new data starts only a month after the implementation of the new model. It provides early confirmation of the model improvements.

Following a rewrite of the UMOS code, the system was turned over fully to CMC, though MRD (now RPN) scientists assist CMC with issues that arise from time to time. UMOS runs operationally as a basis for the automated text (SCRIBE) forecasts. As new models were added to the operational mix, UMOS was extended to provide forecasts based on these models, mainly for temperature, dew point and wind. The categorical predictands for the medium range continued to be predicted using PPM, following the methods described in Brunet et al (1988). One notable addition after 2010 was the use of UMOS for regional prediction of air quality parameters surface ozone, nitrogen dioxide .and particulates pm 2.5.

In the late 2000's Dr. William Burrows, who was a stalwart of the MRD group, and the architect of Canada's first operational MOS product, moved to Edmonton, but continued his post-processing research work from there. Examples of this later work include a rule-based model to forecast fog occurrence (visibility < ½ mile) and stratus ceilings less than 500 ft. over all of North America and adjacent oceans. This model was verified at CMC and outperformed all other guidance there. It is now widely used by forecasters in all regions, and by the

Canadian Coast Guard and National Defense. In 2019 it was implemented at CMC. The method is described in Burrows and Toth (2011).

Dr. Burrows more recent work includes application of CART techniques to forecasting of blizzard conditions in the Arctic. An example of the forecast appears in “The Canadian Arctic Weather Science Project” in BAMS, 2019.

In summary, the national statistical post-processing system consisted of PPM max/min temperature until 1986 and analogue-POP until 1980. PPM-REEP POP and POPA forecasts ran until 1986. A full set of MOS products was developed for the regional model for projections out to 48 h (wind, temperature, dew point temperature, POP, cloud amount and POPA, and precipitation type). Most of these were implemented in 1986-87, but were converted back to redeveloped PPM forecasts by 1989. Using redeveloped PPM equations, a full set of products was developed for the medium range (days 3 to 5) using the global model and ran from 1989 until the present. Under the UMOS system, MOS 3 h spot temperature, dew point temperature, wind direction and speed, POPA, POP and cloud amount forecasts were developed on the regional model for short ranges, and were implemented in stages from 2000 on. The temperature, dew point and wind forecasts still run, while the categorical variable forecasts were stopped in 2018 in favour of smoothed direct model output. UMOS products were developed for the global model run out to 144 h for temperature and dew point temperature, and still run. MOS forecasts were not redeveloped for the medium range global model, following the results of Brunet et al (1988). Table 2 is a mostly complete list of statistical products implemented into operations in Canada.

3. SPECIAL PRODUCTS

While development of the national statistical interpretation system proceeded more or less continuously from the 1960s to the present day, there were occasional forays into the development of special products, some of which were national and ran at CMC, while others were used as forecast guidance in specific regions. Several of these products are described in this section, in chronological order of development.

3.1 The suite of MOS/PPM forecasts for the 1988 Calgary Winter Olympic Games.

These products are probably the best example we have seen of MOS successfully used for downscaling coarse-resolution model output (381 km). The Olympic setup for Calgary consisted of a total of 3 locations, several venues in Calgary, the Nordic skiing at Canmore and the downhill skiing at Mt. Allen. The last two venues are about 75 and 50 km W of Calgary. All the sites were instrumented with numerous sensors during the three winters before the games. This data served as the training sample for a set of MOS forecasts covering all the weather elements of interest to the Olympic committee, including wind forecasts, visibility, precipitation occurrence. Temperature forecasts were needed especially if near, or above 0. The NWP model at the time was V9 of the spectral model running on a 381 km grid. Thus all the venues would be within one grid box of the model and would be seen as essentially one point. This was therefore a chance to use the detailed local observation sets to condition a local forecast on the broad scale forecast as predicted by the model.

Table 2. A list of statistical forecast products run in operations in Canada, mostly at CMC in Montreal, along with the approximate period over which they ran operationally. National products which formed the National statistical post-processing system are indicated in red.

1968-1975	PPM-MLR max/min temp PPM-Analogue POP	At CMC – N. Yacowar Temperature ran until 1986, POP until 1980
1980-1986	PPM-MDA POPA	First National MDA-based forecast tool

1982-1989		Weather element product development software, ECMWF, completed 1982 and used for 7 years by the member states of ECMWF.
1983-1984	PPM-MLR marine winds	Completed 1984, not implemented at CMC.
1986-1989	MOS-MLR winds	Competed 1986. **Canada's first MOS product**, ran operationally at CMC until 1989
1986	MOS-MLR, MDA Several elements	Special for the CASP I field project, 1986. Used and evaluated by Atlantic Region meteorologists during the CASP field experiment.
1986-1989	MOS-MLR spot temperature and POP	Implemented at CMC 1986, ran until 1989
1986-1991	Classical-REEP VSRF Several elements	Designed for aviation: "SHORT", completed 1986, revised several times until 1991. Tested in four MSC regions.
1987-1988	MOS-MLR and MDA Several elements PPM as well.	MOS forecasts for the Calgary winter Olympics, 1988. Based on special datasets at Olympic sites, Also regional forecasts using regional model output
1989	MOS-MDA POPT	Completed 1989. Not implemented due to model change.
1988-	PPM POPA, cloud amount	Redeveloped PPM suite for use on Global model for medium range. Still run at CMC
1989-2000	Tuned PPM-MLR winds	Completed 1989, ran operationally at CMC until September, 2000
1991-	PPM-CART Great Lakes snowsquall forecasts	Completed 1991. Used operationally by Ontario Region.
1992-1993	First UV Index forecast	Completed May, 1992. Replaced in 1993.
1993-1997	PPM-MLR UV Index	Completed Spring, 1993. Basis for Canadian UV index, ran operationally at CMC until 1997, when replaced with upgraded version.
1997-	PPM-CART-MLR UV Index	Takes account of clouds, completed 1997. Currently operational at CMC.
1997-	CART-fuzzy inference (CANFIS) ground-based O3	Completed 1997. Currently operational in four Canadian regions.
1998-	UMOS-MLR 3h spot temperature, wind	Completed 1998. Implemented into full operations at CMC, September 2000.
2002-	UMOS-MLR 3h spot temperature, wind on Global model	out to 6 days.. Implemented at CMC 2003

2003-	MOS-CART day 1-2 lightning probability forecasts	Uses lightning detection network as observation. Revised in 2008, operational for all regions, doesn’t run from CMC.
2004-2007	Bayesian model averaging – eps temperatures	Uses independent model sources of temperature forecasts from deterministic and eps to produce optimal weighted pdf
2007-2018	UMOS-MDA POPA, Cloud amount	Regional model version, implemented 2007. Stopped 2018 for smoothed DMO
2008-	Rule-based fog forecasting system	Developed in Edmonton, runs as desired in all regions, implemented in CMC 2019
2008-	UMOS-MLR for AQ	Prediction of O3, NO2 and PM2.5 from high resolution model. Operational from 2010
2018-	CART arctic blizzard forecasts	Under development in Edmonton

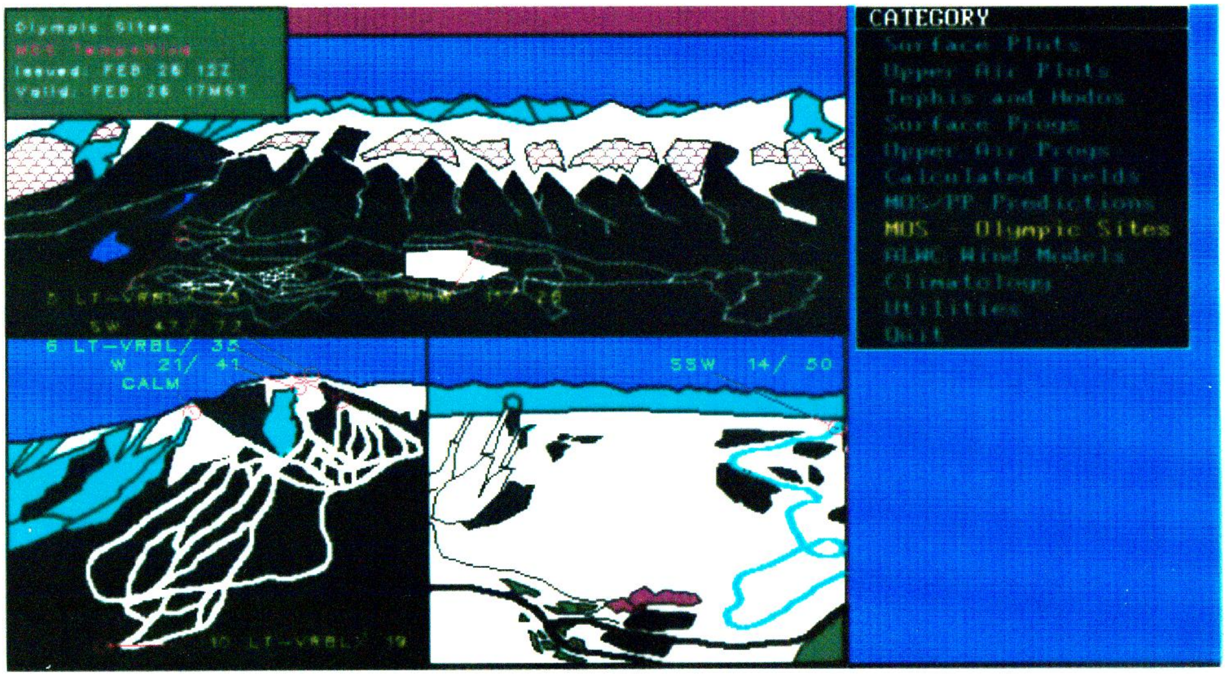


Figure 6. An example of the display for MOS forecasts for the Olympic venues Canmore (Top L), Mt Allen downhill skiing (bottom L), and Calgary sites (bottom middle). The box at the upper right is the list of forecast types available.

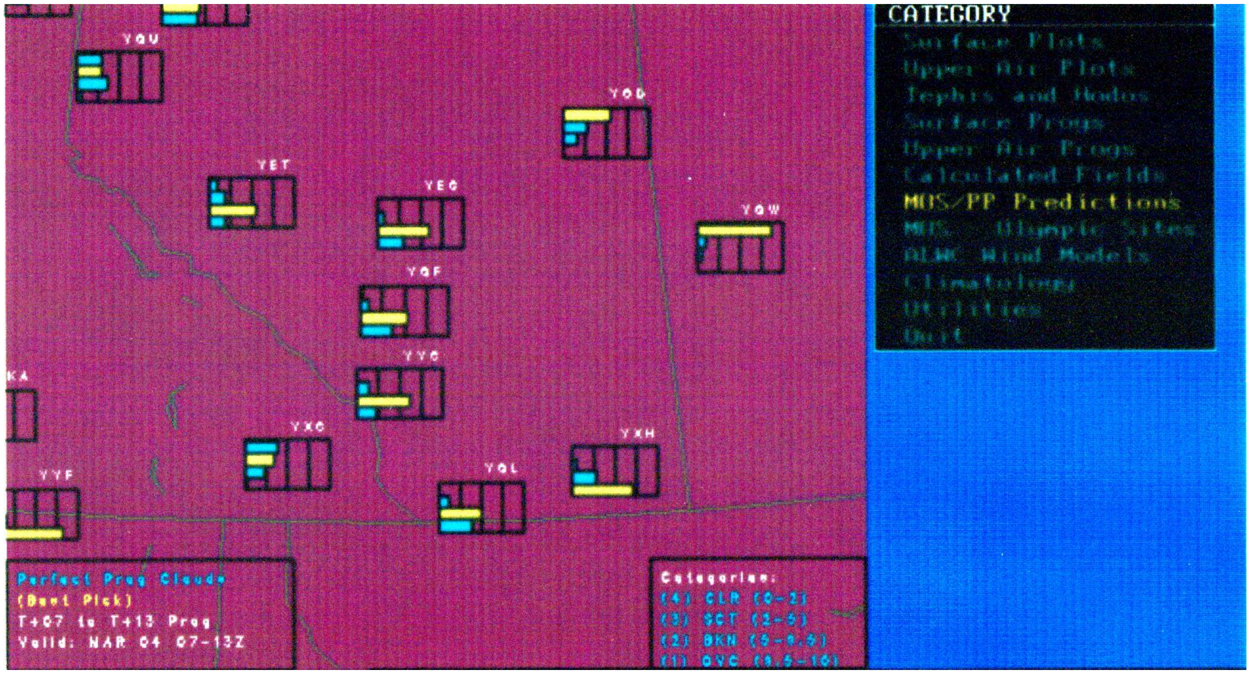


Figure 7. An example of the regional forecasts made available for the 1988 Calgary winter Olympics These are PPM forecasts of cloud amount in 4 categories. Bar graphs for each station give the category probabilities; the yellow bar is the suggested forecast category.

It was also recognized that the MOS products would need to be presented in a useful format if we wanted them used in the operational context. And so a display format like that shown in Figure 6 was devised and the forecasts would be downloaded to an onsite workstation and displayed in this format. The example in Figure 6 is for MOS winds. Three values are shown for the Mt. Allen site, corresponding to winds at the top of the run, the middle and the bottom of the runs. In this way, differences in the local winds, conditioned on the large scale circulation predicted by the model, could be displayed.

The Olympic statistical forecasts also included station specific forecasts of several variables over the broader Western region, derived for all the stations in the area. MLR was used for temperature and wind, while MDA was used for clouds and precipitation.

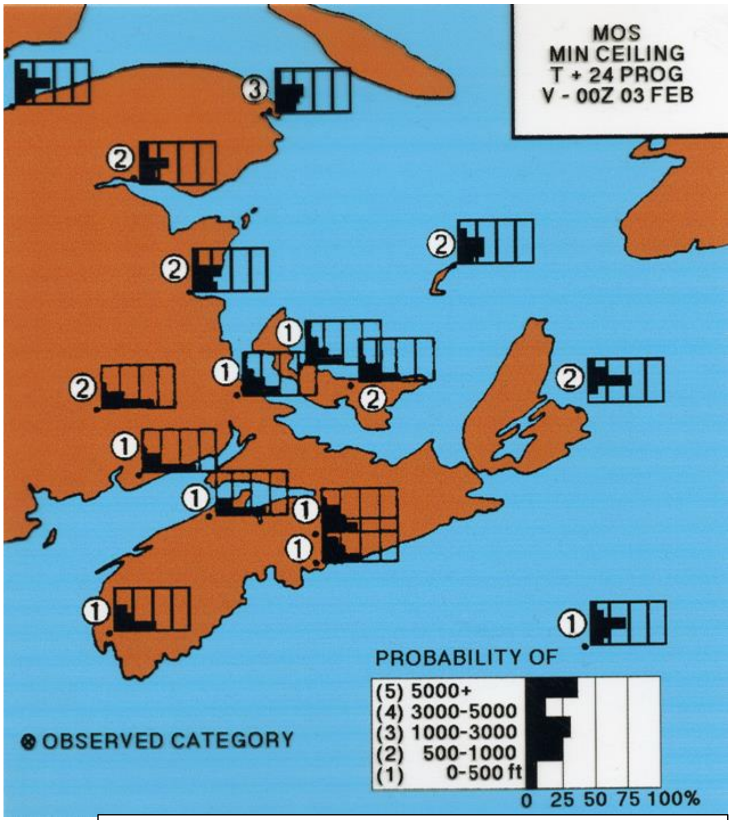


Figure 8. Five category MOS-MDA probability of minimum ceiling over 6 h period, for Atlantic Canada. Verifying category shown in circles.

The display of these forecasts was also specially designed for both the Olympics forecasts and also for a field research experiment called Canadian

Atlantic Storms Experiment, which took place around the same time. Once again, the forecasters appreciated using a graphical representation rather than tables of numbers.

Figure 7 shows an example of PPM cloud forecasts for the region of Calgary. The forecast is in four categories, clear, scattered, broken (cloudy) and overcast. The bar graphs represent the probabilities of each category at each station, with the yellow bar suggesting the "best" category to predict. Note that the "best" category isn't always the one with the highest forecast probability. The category selection procedure takes into account the climatological frequencies of the categories.

Figure 8 is an example of the guidance available to the planners of the CASP field experiment, in this case MOS forecasts of minimum ceiling over a 6 h valid period, in categories. The number next to the station is the verifying category. These are MOS forecasts based on the Spectral model version 8, displayed for Atlantic region meteorologists in a user-friendly fashion.

The Olympic MOS and PPM forecast guidance package, with its special displays may seem like a lot of work for a short operational period – it was run only for about a month bracketing the two-week Olympic games – but that goes along with the nature of the Olympics. And, we learned a lot about MOS and PPM equation development issues, and user interface issues during this work. The forecasts were used during the Olympics, but fate conspired against us in a statistical sense: The weather that occurred during the Olympics was unlike anything that occurred during the three previous winters of special data collection. Bearing that in mind, and also the shortness of the operational period, it was decided that a full retrospective verification would not be done.

We still believe the point forecasts available to the 1988 Olympics for the Olympic venues were probably better than the pointwise guidance available to the Vancouver 2010 games. By the end of the 2000's decade development resources dedicated to statistical post-processing had decreased in favour of high-resolution modelling. The emphasis in 2010 was showcasing the 2.5 km regional model in the mountainous terrain of BC, along with its higher resolution 1km version.

3.2 The Classical-REEP (CR) very short range forecast guidance "SHORT".

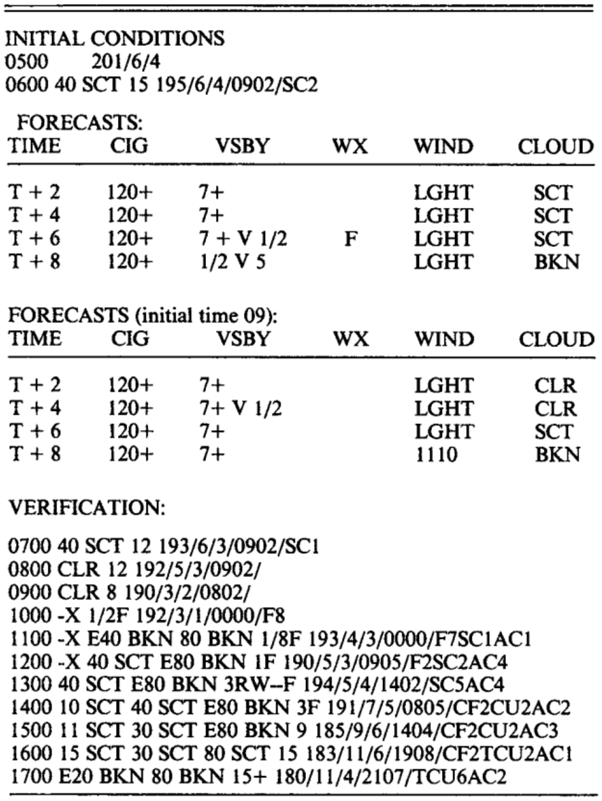
The CR method is a variant of Miller's (1981) generalized equivalent markov (GEM) method. Like GEM, we use REEP (Miller 1964) to define the statistical relationship, but we develop separate REEP equations for each forecast projection rather than following the markov process of using the powers of the transition matrices to define successive forecast projections. In the CR technique the forecast projection is determined by the time lag between predictors and predictand ("classical" formulation). Our CR method is described in Wilson and Sarrazin (1989).

The method was designed as an aid for the preparation of aviation TAF forecasts, in the first 12 hours following observation. Forecasts were created for 2, 4, 6, and 8 h ahead using all the elements of the hourly surface observations that would be of interest to aviation, divided into 90 exclusive and exhaustive categories. The REEP prediction equations were built, then the resulting probability forecasts were subjected to a "best category" categorization procedure designed to maximize the threat score (Miller and Best 1981). Development sample size was typically 28 years, but varied by location.

Table 3 shows an example of the forecasts and the verifying observations for London Ontario, Oct 5 1988. It can be seen that the CR forecasts have correctly identified the formation of fog several hours ahead of time, but were slow with the onset time.

Named "SHORT" following the name of the interface program written for forecasters, the method was used most extensively in the Ontario Weather Center in Toronto, which has responsibility for TAFs for all of Ontario. Requests were received from other regions too. New equations were produced on demand for specific stations. Stations were paired according to similarities of climatology to enhance the stability of the equations for less common events. The program would run on the minicomputers of the day, or any PC now, and could be initialized by input of any two consecutive hourly observations. It seemed remarkable that, by using only sequences of observations, one could correctly predict the onset of fog, or other significant changes. We even saw an example where a thunderstorm was correctly predicted several hours in advance. We believe that by using all the elements of the observation at once, there were patterns in the hourly evolution of all the components taken together that signalled the approach of significant changes. Those sudden changes were enhanced by the categorization procedure.

Table 3. Classical-REEP forecasts at 2 h intervals for London Ontario, October 6, 1988. Two initial times are shown, 06 UTC and 09 UTC



INITIAL CONDITIONS
0500 201/6/4
0600 40 SCT 15 195/6/4/0902/SC2

FORECASTS:

TIME	CIG	VSBY	WX	WIND	CLOUD
T + 2	120+	7+		LGHT	SCT
T + 4	120+	7+		LGHT	SCT
T + 6	120+	7 + V 1/2	F	LGHT	SCT
T + 8	120+	1/2 V 5		LGHT	BKN

FORECASTS (initial time 09):

TIME	CIG	VSBY	WX	WIND	CLOUD
T + 2	120+	7+		LGHT	CLR
T + 4	120+	7+ V 1/2		LGHT	CLR
T + 6	120+	7+		LGHT	SCT
T + 8	120+	7+		1110	BKN

VERIFICATION:

0700 40 SCT 12 193/6/3/0902/SC1
0800 CLR 12 192/5/3/0902/
0900 CLR 8 190/3/2/0802/
1000 -X 1/2F 192/3/1/0000/F8
1100 -X E40 BKN 80 BKN 1/8F 193/4/3/0000/F7SC1AC1
1200 -X 40 SCT E80 BKN 1F 190/5/3/0905/F2SC2AC4
1300 40 SCT E80 BKN 3RW--F 194/5/4/1402/SC5AC4
1400 10 SCT 40 SCT E80 BKN 3F 191/7/5/0805/CF2CU2AC2
1500 11 SCT 30 SCT E80 BKN 9 185/9/6/1404/CF2CU2AC3
1600 15 SCT 30 SCT 80 SCT 15 183/11/6/1908/CF2TCU2AC1
1700 E20 BKN 80 BKN 15+ 180/11/4/2107/TCU6AC2

Around the turn of the millennium, a project called "TAFTOOLS" was launched, a collaborative effort between the MRD group and the CMC group. The plan was to use SHORT as the very short range component of a more comprehensive statistical forecast system, blending the SHORT forecasts with MOS forecasts over the 6 to 12 h range to produce a more rapid response statistical forecast system, and to optimally use the predictive information in the most recent observation. TAFTOOLS tests are described in Bourgouin et al 2002 and Montpetit et al 2002. Unfortunately, the planned system never made it into full operation.

3.3 The Perfect Prog ozone/UV index forecasts.

A reaction on the part of the Minister of the Environment to a US forecast of significant ozone thinning in the North Polar Region during the spring of 1992 led to a flurry of activity to bring into operations a UV index, "To warn Canadians of the hazards of enhanced UV radiation caused by ozone thinning". In collaboration with colleagues in Research Branch the MRD group set about to meet the Minister's wish to start a UV index service to Canadians by the summer of 1992. This project diverted the resources of MRD for a couple of years or so, and resulted in an operational national forecast product at CMC.

With only two months lead time before the summer season in 1992, there was not enough time to carry out any research. The procedure mounted in 1992 to meet the minister's demand was therefore scientifically shaky at best, but relied mostly on real time ozone observations and worked sufficiently well to meet the need. For 1993, a statistical forecast product for stratospheric ozone was developed (PPM-MLR), and implemented at CMC. This product was upgraded to account for cloud cover in 1997 and still runs at CMC in 2020 as the basis of the UV index. The development of the index is described in Burrows et al (1994), and the development of both the UV index and the operational observation program is described in Vallée et al (1994).

Figure 9 shows a verification of the UV forecasts in Dobson units (DU) for the summer of 1994. For this verification, we used observation data from two independent sources, Brewer spectrophotometer measurements, analysed for the Northern Hemisphere, and TOMS satellite data. Since the RMS difference between the two datasets was 9.9 DU and the error with respect to the datasets was of the order of 13 DU, we concluded that the statistical technique, along with its daily correction based on the previous day's data worked well enough, and it would not likely be worthwhile to try to move to a MOS formulation. (Wilson and Vallée 1995).

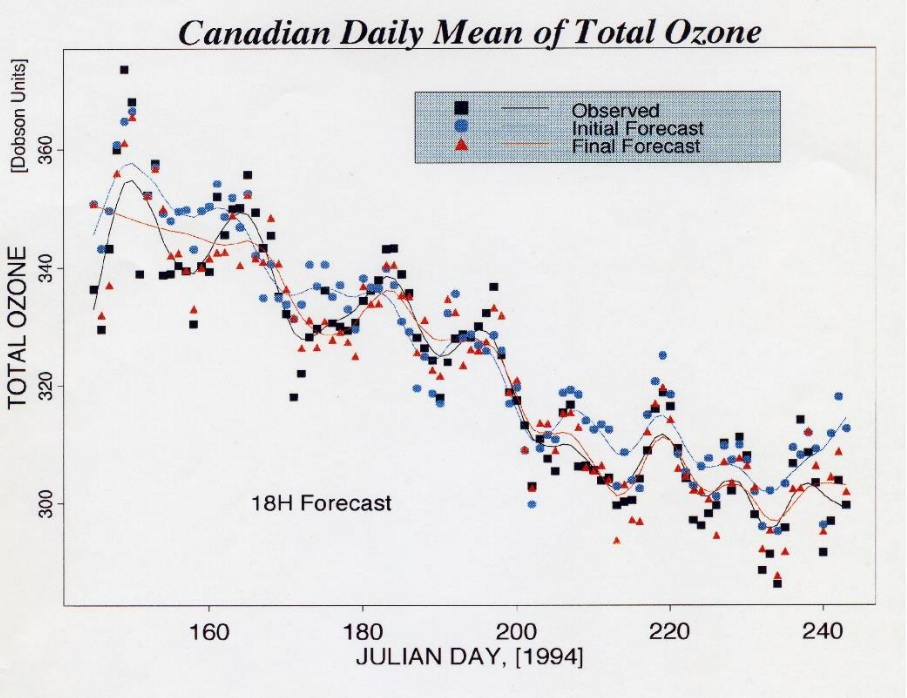


Figure 9. Verification of PPM-MLR total ozone forecasts for Canada for summer, 1994. Curves are local regression fits to the data points. (From Wilson and Vallée 1995)

3.4 Short range MOS-CART Lightning forecasts.

The extension of the US lightning detection network northward to cover most of Canada provided a real boost to statistical interpretation because, for the first time, we were able to positively identify the location and timing of thunderstorms. While the detection efficiency isn't perfect, (poor detection of cloud to cloud strikes and about 80% detection rate for cloud to ground strikes), it is far better than the surface observation network which seriously undersamples small-scale features such as thunderstorms. The MRD group began working with the lightning data as soon as enough was collected to provide an adequate training sample. The first step was to carry out a national climatology of thunderstorms, which revealed some previously unknown spatial features (Burrows et al 2002). Then, we turned our attention to MOS prediction using the then-operational global environmental multiscale (GEM) model data.

Thunderstorm occurrence is an inherently categorical phenomenon, which meant that it would be suitable to use the Classification and Regression Trees (CART) statistical method. (Steinberg and Colla 1995). CART is a decision tree-based method which is said to mimic the thought processes of forecasters; at least we were told by forecasters that they liked the concept. Burrows et al (2005) describes the development and testing of the MOS-CART forecasts.

Figure 10 shows a typical sample result from the forecast method. It is a 24 h forecast, valid for the 3 h period 21 to 24 h ahead of data time. The forecast probabilities of lightning occurrence are contoured at 25% and 75% and the actual occurrence of lightning during the valid period is superimposed on the map. The availability of observation data at higher resolution than the model output gives rise to challenges in the interpretation of the verification results. For example, Figure 10 shows that the broad structure of the line of thunderstorms along the east coast has been fairly well predicted, but that the detailed structure of the forecasts is not so well matched to the details of the observations.

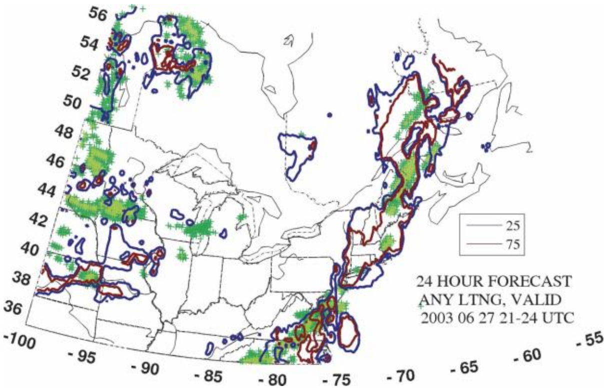


Figure 10. Lightning probability forecast for 21 to 24 h. Contours are 25% probability (blue) and 75% probability (red) of lightning during the valid period. Observations are shown as green crosses (any lightning) and yellow crosses (intense lightning). (From Burrows et al 2005)

Since the observations are at higher spatial resolution than the model forecasts, the lightning forecasts can be used as a diagnostic tool for the model. Casati and Wilson (2009) used a wavelet-based procedure to partition the Brier Skill Score into scale components for verification of the lightning forecasts. Results (Figure 11) indicate that the GEM model has skill in forecasting only the larger scales, greater than about 400 km.

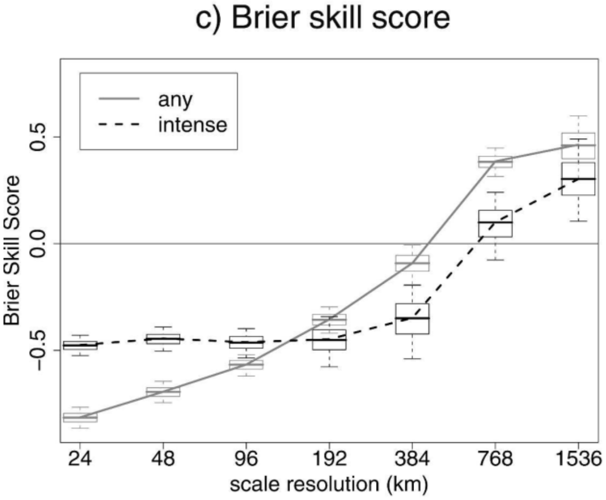


Figure 11. Brier Skill Score verification of the MOS-CART lightning forecasts according to scale. Bootstrapped confidence intervals are shown by the box and whisker plots. (From Casati and Wilson 2009)

As for the lightning forecast technique, the equations were redeveloped for all of Canada in 2008, and have run operationally since then, supplying forecasts to the ECCC website. They are communicated to forecasters in all regions, to the Canadian Forest Service, BC Hydro, and Vaisala. See Burrows (2008) for a description of the newest version.

3.5 Post-processing ensemble forecasts

3.5.1 Bayesian Model Averaging. (BMA)

The advent of ensemble systems in 1992 and in Canada in 1996 has provided great opportunities to statistical interpretation in general, although Canadian efforts suffer from a lack of resources. Both the CMC and MRD statistics groups have turned some of their attention to post-processing ensemble forecasts. From about 2003 to 2007, the two groups have collaborated on Bayesian Model Averaging (BMA), a method designed to combine and weight the forecasts from several sources to produce a corrected probability distribution function (pdf). Of the many post-processing methods reported in the literature, BMA was chosen for its potential to combine independent forecasts from different sources to refine a forecast probability distribution of the event, weighting the different sources according to their recent performance statistics. Both ensemble sources and deterministic sources may be included. With NWP centers now running several different models and also ensemble systems, methods such as BMA are needed to help forecasters assimilate all the available information.

Our experiments with BMA for temperature are described in Wilson et al (2007). The BMA was set up to use the last 40 days of forecasts to determine the error levels of all the candidate models, including the discretely different components of the ensemble system and the deterministic model. The main result from the study is that the BMA was able to reduce the spread of the probability distribution function by about 20% overall, while scoring just as well as the ensemble mean dressed with its RMS error based on the same training sample.

Figure 12 shows an example of a BMA- optimized 48 h forecast temperature distribution for Montreal. The forecast question is whether, at the forecast time, the cold front would have passed or not. The bias-corrected original forecasts show a highly bimodal distribution, according to whether the front will have passed or not (crosses). Based on 40 days of recent forecast history, the analysis has decided that all the members of the cold-valued mode are not reliable forecasts and that some of the warmer-valued members are more reliable and have been given higher weights. Thus the final blended distribution is aligned with the warmer option, that is before the frontal passage.

Analyses such as BMA can thus be used to blend model information from several different sources, and can be relied on to assign higher weights to those models which are more accurate in each situation. One caution about BMA is in order: only distinctly different (systematically different) models should be included. Since most ensemble systems are actually derived from a single model with stochastic perturbations to the initial conditions and/or the model parameters, all forecasts from such an ensemble are to be considered equally likely, and, for the purpose of BMA, should therefore be treated as a single member in the BMA analysis, to be combined with the deterministic model, the unperturbed control forecast, different resolution models from the same center, and even model forecasts from other centers, if operationally available.

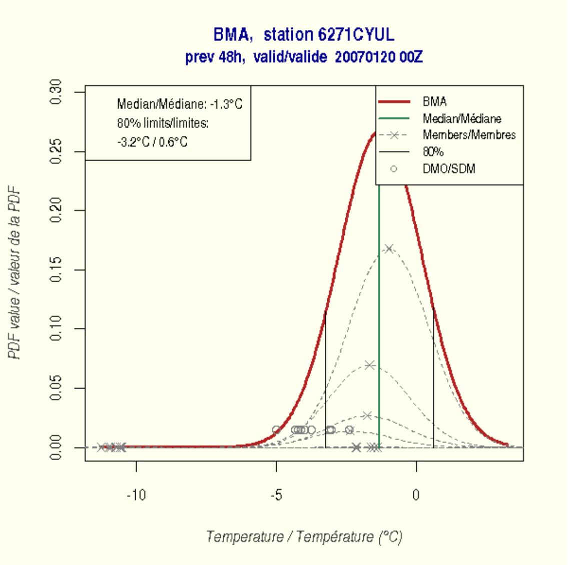


Figure 12. An example of a 48 h BMA – optimized temperature forecast for Montreal. The circles represent the original ensemble forecasts and the crosses the bias-corrected forecasts.

BMA isn't the only way of blending model information; other methods could be used. It is, however a sensible way of blending information from diverse sources, taking into account recent quality of performance of each model, to assist the forecaster in optimizing his forecast.

Following the Canadian development effort, the BMA was put into test implementation at CMC, but was not implemented into full operation.

3.5.2 Other ensemble processing

Using the UMOS tools, a new perfect prog temperature forecast has been developed by Bertrand Denis for the global ensemble system. The training sample came from 31 years of NCEP reanalysis data; the development and tests are described in Denis and Verret (2004) (available from https://ams.confex.com/ams/pdfpapers/67887.pdf)

These forecasts became operational in 2006.

4. CURRENT STATUS AND DEVELOPMENT ACTIVITIES

In 2020, UMOS still is running, as follows:
-Regional deterministic prediction system: 0 to 84 h, every 3 h, for temperature, dew point, wind direction and speed, marine wind direction and speed.
-Global deterministic prediction system: 0 to 144 h, every 3 h, for temperature and dew point temperature.
-High resolution deterministic prediction system: 0 to 48 h, every 1 h, for temperature, dew point, wind direction and speed over land (not currently operational)
-Regional air quality deterministic prediction system: 0 to 72 h every 1 h, for O_3, NO_2 and PM2.5 (implemented in 2010)

The UMOS probabilistic forecasts, POP, POPA and 4-category cloud opacity were stopped in 2018 in favour of simpler post-processed model output. The new post-processing is in the form of smoothing of the gridded model output precipitation forecasts using neighbourhood representations of the probability of precipitation at gridpoints. Cloud forecasts were subjected to simple smoothing over a 30 km neighbourhood. The main advantage is that the simpler post-processing reduces the management overhead of the operational forecasts, but the downscaling and model bias correction abilities of MOS have been lost. It would be interesting to see whether improvements could be made using UMOS with the smoothed model variables as input predictors, related to station data as the predictand.

It is planned to convert the whole statistical interpretation system over to gridded format, as has been done in the US. RPN and CMC are collaborating to do studies aimed at figuring out the best way to do this. One way that has been suggested is to carry out UMOS post-processing, then use kriging to spatially extrapolate the corrections over the whole domain, with reference to a model-based background field. The UMOS forecasts at stations would be preserved in this way, while corrections in data-sparse areas would tend more towards the model trial field. This approach is claimed to allow the benefit of the spatial coherence in the model output to be included in the post-processed forecast, while retaining the ability of the forecasts to represent local effects where there is data to support them. Disadvantages are that biases and errors in the model output are transmitted through to the post-processed output. As well, the blending of high resolution pointwise forecasts with lower resolution forecast information from the model in data sparse areas will lead to spurious spatial variations in the forecast values due to, for example, differences between the lower resolution model climatology and the pointwise climatology at stations, and, effects of differences between the smooth model physical boundary conditions and the real topography represented in the vicinity of the station.

Another idea is to grid the observations, then do the post-processing at the grid points. This would have the effect of taking away any ability of the statistical interpretation to respond to local effects, unless a highly localized analysis method were to be used. Worse, if the gridding (analysis) method uses model output as a trial field, then the ability of the forecast to account for any biases between the model climatology and the station climatology (of the training sample) is lost. This approach might be simple to apply, but would be limited to correcting the model error bias component with respect to the analysis.

One might ask the more general question of why it is important or necessary to post-process surface variables at all locations in Canada (or at all grid points). The station network mirrors at least approximately the population distribution across the country. If no one lives anywhere near a grid point, as in much of the Arctic, then why go to the trouble of trying to fix the model output at that point? In data sparse (and people-sparse) regions, surely we can accept that the existing uncorrected model output is the best estimate we can get. There is no guarantee that any attempts to improve it will correct it in the right direction, and also, since there is no data there anyway, we will never know.

At CMC, quite a lot of effort is going into a system called PROGNOS, which is intended to eventually replace UMOS. This project was described at last year's AMS annual meeting by Saad et al (2019) (Available at https://ams.confex.com/ams/2019Annual/webprogram/Paper354421.html). This work is intended to renew and extend the rapid updating capabilities of UMOS, leading to reduced system maintenance cost and increased flexibility. In addition to the statistical techniques MLR and MDA, the new system will include AI techniques such as neural nets. It will be a modular system and it will be easier to add new predictors and predictands than it currently is with UMOS. An experimental version is running on the regional deterministic model and on the regional air quality model.

In conclusion, statistical post-processing is still quite well-supported in CMC, although the level of effort in RPN has dropped since about 2008.

5. ISSUES AND DISCUSSION

5.1 Is statistical interpretation still useful?

The question that often comes to mind is, "Is it still worth the effort to do statistical post-processing to improve model forecasts?" Early on, say 35 years ago, it had already been shown that statistical post processing does clearly improve on direct model output, by removing model error bias (MOS), by downscaling coarse resolution model output to local scales, and by translating model output variables into weather elements as observed at stations. As some have said, post-processing "keeps the model honest and grounded in reality according to in situ observations.” It was also generally believed that, as models improved, the “room for improvement” available to post-processing efforts would necessarily decrease, and eventually post-processing would be unnecessary.

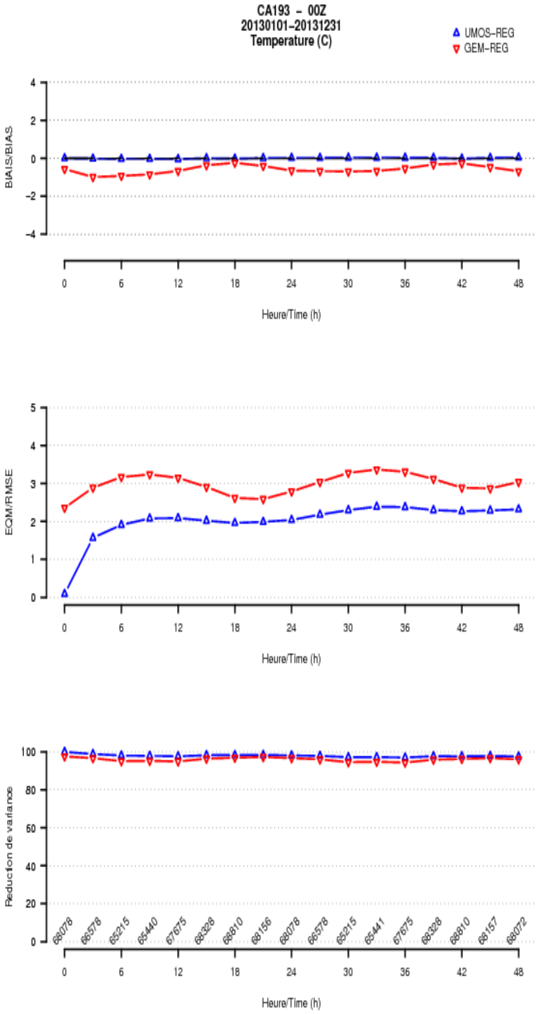


Figure 13. Comparison of direct model output (red) and UMOS (blue) temperature forecasts for 2013, in terms of bias (top), RMSE (middle) and reduction of variance (bottom). Sample sizes are given along the abscissa

Well, it can be said with confidence that that point has not been reached in the last 30 years. Consider Figure 13, showing bias, root mean square error and reduction of variance for a large sample of Canadian temperature forecasts during 2013.

Figure 13 shows that the UMOS forecasts are unbiased, while the DMO forecasts have a slight cold bias; RMSE is 0.5 degrees lower for the UMOS forecasts, and more than that in the shortest ranges; and the reduction of variance is slightly better for UMOS. Note that the reduction of variance is artificially high for all techniques because the reference climatology is annual and averaged over all stations.

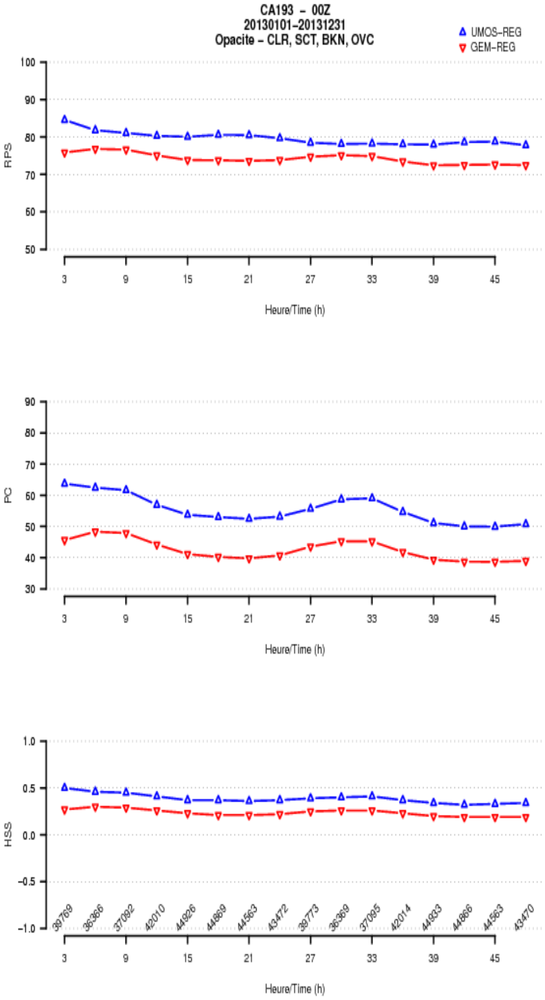


Figure 14. Comparison of direct model output (red) and UMOS (blue) categorical forecasts of cloud opacity in 4 categories for 2013, in terms of rank probability score (positive orientation), percent correct and heidke skill score. Sample sizes are given along the abscissa.

Figure 14 shows the comparison for the 4-category UMOS-MDA cloud opacity forecasts. Here also, the differences are noticeable, with percent correct about 15% higher for UMOS than the DMO. The RPS is typically about .05 higher for the UMOS forecasts, as is the HSS.

These differences are large enough to be statistically significant, given the large sample sizes, even allowing for some serial and spatial correlation in the dataset. It is interesting to compare Figure 13 with Figures 4 and 5. The two evaluations involve data that is 10 years apart, but the performance advantage compared to the DMO is about the same, 0.5 degrees improvement in RMSE. This supports the belief that the quality of the DMO is not approaching that of the post-processed model output, at least not yet. We conclude that statistical interpretation has not decreased in value or importance.

5.2 Disconnection from the weather at points.

The trend towards statistical post-processing for grids rather than at station locations is worrisome. In the US, gridded MOS has been in operation for some time, while in Canada, CMC is rapidly working towards this goal. Expressing statistical guidance forecasts on grids may not be a bad idea as long as the statistical relationships are built with respect to in situ observations as the predictand variable. Only in that case, where the predictand variable is not pre-processed (by smoothing, analysis etc) can the full downscaling potential of MOS be realized. Processing of the MOS predictor variables, however, is often a good idea, for example to smooth out spatial and temporal scales that are not well-predicted by the model. Point, or at least locally-defined observations are always the most relevant as predictands since we all experience weather at points, and not, for example as an average over some arbitrary grid box.

Models, of course are better at predicting weather variables which are averaged over an area, since the averaging eliminates the smallest unresolvable and least accurate spatial scale components from the variables. Thus it may be tempting to carry out statistical interpretation with respect to analysed data as “observations“. MOS done in this way will account for the model’s forecast mean error, but cannot downscale the forecast. Bias between the analysis (which is nudged towards the model’s climatology via the use of the model as a trial field) and the point observations is not considered. Thus, MOS equations developed this way may score better than those developed with in situ point data, but will not be more accurate in a real sense. On the other hand, the use of analyses, or better still, reanalyses which use a high resolution constant model for the background field is a good way to build PPM equations with respect to point in situ observations. PPM prediction using such equations is more transparent to users in the sense that the forecast output will contain mostly model error information (both bias and variable error), since the effects of average representativeness (scale) differences are accounted for in the PPM equations.

We fear that the connection between model forecasts and real weather as represented by station data is being lost in the move towards gridded forecasts. And, along with it we lose the ability to connect model output with the weather experienced by individuals.

5.3 Attitude of the modeling community.

For most of the last 40 years or so, the modeling community has barely tolerated statistical interpretation research and development. The prevailing attitude seems to have been that the model will solve all forecasting problems, if not now, then “soon”. Models will be good enough that any attempt to improve on their output, except maybe by simple interpolation or smoothing will be futile. We have heard this attitude expressed in many parts of the world. In the Canadian context, it has led to reduction in resources dedicated to statistical interpretation, especially after 2000. It is clear that model output has improved considerably, but as the model output has improved, so have the products of statistical interpretation. As shown above, there is still room for improvement, and the magnitude of improvement is significant.

In practical terms, there has always been a trade-off between the extra effort required to run a statistical interpretation system operationally, and the cost of using direct model without correction. Recently, as models have improved, there are indications, at least in Canada, that the extra effort is sometimes not considered worth the improvement obtained, and MOS or PPM products have been replaced with smoothed direct model output forecasts. That is, there is an increasing tendency to deem the model output “good enough”, lowering standards for the sake of expediency. This too is a worrying trend.

5.4 New data sources?

To end on a positive note, the last 15 years or so have seen a rapid increase in the accuracy and decrease in the cost of measuring meteorological variables, especially temperature. Coupled with GIS technology, the real possibility exists of harvesting this data for eventual use in statistical interpretation. It could reduce the chronic problem of data sparsity, especially in more populated areas. We don’t know how such a system could be set up, but efforts to collect some of this type of data are being made in support of road weather forecast services in Finland and other countries. It would be a real boost to be able to use some of this “non-standard” data in real-time statistical interpretation of model output.

6. ACKNOWLEDGEMENTS

The authors wish to thank all those who contributed information to this summary, and for convincing us that statistical interpretation is alive and well in ECCC. They are: Dr. William Burrows, Jacques Montpetit, Marcel Vallée, Stéphane Beauregard, Gérard Croteau, and Marc Klasa.

7. REFERENCES.

Bourgouin, P., J. Montpetit, R. Verret and L.J. Wilson, 2002 : TAFTOOLS : Development of objective TAF guidance for Canada Part One – Introduction and development of the very short-range module. Preprints, *Sixteenth Conference on Probability and Statistics in Atmospheric Sciences*, Orlando, Florida, Amer. Meteor. Soc., 94-98.

Brunet, N., R. Verret and N. Yacowar, 1988 : An objective comparison of Model Output Statistics and “Perfect Prog” systems producing numerical weather element forecasts. *Wea. Forecasting* **3**, 273-283.

Burrows, W.R. and G. Toth, 2011: Automated fog and stratus forecasts from the Canadian RDPS operational NWP model. Extended Abstracts, *24th Conference on Weather and Forecasting*, Seattle, Washington, Amer. Meteor. Soc., 33 pp. DOI: 10.13140/2.1.4852.0648.

Burrows, W.R., 1984: Development and testing of a MOS wind forecasting system for Canada. Meteorological Services Research Branch Internal Report No. MSRB 84-3, Atmospheric Environment Service, Canada.

Burrows, W.R., 2008: Dynamical-statistical models for lightning prediction to 48-hr. Extended Abstracts, *Third Conference on Meteorological Applications of Lightning Data*, New Orleans, Louisiana, 23 pp.

Burrows, W.R., P. King, P. J. Lewis, B. Kochtubajda, B. Snyder, and V. Turcotte, 2002: Lightning occurrence patterns over Canada and adjacent United States from lightning detection network observations. *Atmos.–Ocean,* **40,** 59–80.

Burrows, W.R., C. Price and L.J. Wilson, 2005: Warm season lightning probability prediction for Canada and the northern United States. *Wea. Forecasting*, **20**, 971-988.

Burrows, W.R., M. Vallée, D. Wardle, J.B. Kerr, L.J. Wilson and D. Tarasick, 1994: The Canadian operational procedure for forecasting total ozone and UV radiation. *Meteorological Applications*, **1**, 247-265.

Casati, B. and L.J. Wilson, 2009: A new spatial-scale decomposition of the Brier score: Application to the verification of lightning probability forecasts. *Mon. Wea. Rev*., **135**, 3052-3069.

Dallavalle, J.P., 1988: An evaluation of techniques used by the National Weather Service to produce objective maximum/minimum temperature forecasts. Preprints, *Eighth Conference on Numerical Weather Prediction*, Baltimore, Maryland, American Meteorological Society, 572-579.

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.

Klein, W.H. and F. Lewis, 1970: Computer forecasts of maximum and minimum temperature. *J. Appl. Meteor*., **9**, 350-359.

Klein, W.H. and H.R. Glahn, 1974: Forecasting local weather by means of Model Output Statistics. *Bull. Amer. Meteor. Soc*., **55**, 1217-1227.

Kruizinga, S. and H. Boettger, 1984: Report on the application and use of ECMWF products in member states, including verification results. European Center for Medium Range Weather Forecasts, 173pp.

Macdonald, K.A., 1986: An evaluation of operational MOS surface wind forecasts. *Atmospheric Research Report No. 86-3*, Environment Canada, 33 pp.

Miller, R.G., 1981: GEM: A statistical weather forecasting procedure. NOAA Tech. Report NWS-28, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 103 pp.

Miller, R.G. and D.L. Best, 1981: A Beta classification model. TDL office note 81-8, National Weather Service, NOAA, U.S. Department of Commerce, 14 pp.

Miller, R.G., 1964: Regression Estimation of Event Probabilities, Tech. Rep. No. 1, Contract cwb 10704, The Travelers Research Center Inc., Hartford, Conn., 153 pp.

Montpetit, J., P. Bourgouin, L.J. Wilson and R. Verret, 2002: TAFTOOLS: Development of objective TAF guidance for Canada Part 2 – Development of the short range forecast module and results. Preprints, *Sixteenth Conference on Probability and Statistics in Atmospheric Sciences*, Orlando, Florida, Amer. Meteor. Soc., 99-101.

Ross, G.H., 1987: An updateable model output statistics scheme. Programme on Short- and Medium Range Weather Prediction, PSMP Report Series No. 25, World Meteorological Organization, 25-28.

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

Sarrazin, R. and L.J. Wilson, 1987: Comparison of MOS and perfect prog probability of precipitation forecasts using the signal detection theory model. *Preprints, Tenth Conference on Probability and Statistics in Atmospheric Sciences*, Edmonton, Alberta, American Meteorological Society

Sarrazin, R. and L.J. Wilson, 1989: A tuned perfect prog wind forecast system for Canada. In Preprints, *Eleventh Conference on Probability and Statistics in Atmospheric Sciences*, Monterey, California, American Meteorological Society.

Simonsen, C., 1991: Self-adaptive output statistics based on Kalman filtering. Programme on Short- and Medium Range Weather Prediction Research, PSMP Report Series No. 34, WMO TD No. 421, XX33-XX37.

Steinberg, D. and P. Colla, 1995: CART: Tree-structured non-parametric data analysis. Salford Systems, 336 pp.

Vallée, M., D. Wardle, L.J. Wilson, W.R. Burrows, J.B. Kerr and D.W. Tarasick, 1994: The Canadian operational daily observation and forecasting program for total ozone and UV Index. CARD Research Report No. 94-007, Environment Canada, 52 pp.

Vallée, M., L. J. Wilson and P. Bourgouin, 1996: New statistical methods for the interpretation of NWP output at the Canadian Meteorological Center. Preprints, *Thirteenth Conference on Probability and Statistics in Atmospheric Sciences*, San Francisco, Amer. Meteor. Soc., 37-44.

Wilson, L.J. and N. Yacowar, 1980: Statistical Weather Element Prediction in the Canadian Weather Service. Proc., *WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting*, Nice, 401-406.

Wilson, L.J. and H. Stanski, 1983: Assessment of operational REEP/MDA probability of precipitation forecasts. In Preprints, *Eighth Conference on Probability and Statistics in Atmospheric Sciences*, Hot Springs Arkansas, American Meteorological Society, 193-199.

Wilson, L.J. and M. Vallée, 1995: Verification of the operational ozone-UV index forecasts for the summer of 1994. CARD Research Report No. 95-006, Environment Canada, 11 pp.

Wilson, L.J. and M. Vallée, 2002: The Canadian updateable Model Output Statistics (UMOS) system: Design and development tests. *Wea. Forecasting*, **17**, 206-222.

Wilson, L.J. and M. Vallée, 2003: The Canadian updateable Model Output Statistics (UMOS) system: Validation against Perfect Prog. *Wea. Forecasting*, **18**, 288-302.

Wilson, L.J. and R. Sarrazin, 1987: Comparison of MOS and perfect prog probability of precipitation forecasts using the signal detection theory model. Proc. *WMO Workshop on Significant Weather Elements Prediction and Objective Interpretation Methods*, Toulouse, 57-60.

Wilson, L.J. and R. Sarrazin, 1989: A classical-REEP short-range forecast procedure. *Wea. Forecasting* **4**, 502-516.

Wilson, L.J., 1985: A new comparison of MOS and perfect prog methods applied to surface wind forecasting. Preprints, *Ninth Conference on Probability and Statistics in Atmospheric Sciences*, Virginia Beach, Virginia, American Meteorological Society, 44-47.

Wilson, L.J., S. Beauregard, A.E. Raftery and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Mon. Wea. Rev*., **135**, 1364-1385.

Wilson, L.J., W.R. Burrows, and K. Macdonald, 1986: An assessment of operational MOS surface wind forecasts. In Preprints, *Eleventh Conference on Weather Analysis and Forecasting*, Kansas City, Missouri, American Meteorological Society, 11-16.

Yacowar, N., 1973: Computer produced forecasts of maximum and minimum temperatures. Canadian Meteorological Research Reports CMRR1/73

Yacowar, N., 1975: Probability forecasts using finely-tuned analogues. In Preprints, *Fourth Conference on Probability and Statistics in Atmospheric Sciences*, Tallahassee, Florida, pp. 49-50.

Yacowar, N., G. Richard, N. Brunet and H. Yang, 1985: Development of a MOS system to forecast the probability of precipitation. CMC Information, 2, No. 4, 24-46.