U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
OFFICE OF SYSTEMS DEVELOPMENT
TECHNIQUES DEVELOPMENT LABORATORY

TDL OFFICE NOTE 91-3

ON MOS AND PERFECT PROG FOR INTERPRETIVE GUIDANCE

Harry R. Glahn

July 1991

ON MOS AND PERFECT PROG FOR INTERPRETIVE GUIDANCE

Harry R. Glahn

1. INTRODUCTION

A rather extensive interpretive weather element guidance system has been in place in the National Weather Service (NWS) for about 20 years. Its beginnings can be traced back to work at the Massachusetts Institute of Technology and the Travelers Research Center (TRC). A number of people, many working on Federal contracts, experimented with various statistical methods and brought those methods into meteorology. Regression and discriminant analysis were used at TRC soon after those methods appeared in the statistical literature. Of those involved, Drs. George P. Wadsworth, Robert M. White, Joseph G. Bryan, and Robert G. Miller deserve special mention. The papers of Bryan (1944) and Miller (1964) on regression stand at the foundation of the NWS's interpretive system.

Much of the experimental work at TRC centered around very short range forecasting, especially of aviation-related weather variables. The predictors in this work were largely surface observations, and since the relationships developed between predictands and predictors had the lag (forecast projection) built in, it falls under what has come to be called the "classical" method.

At about the same time, Dr. William H. Klein of the Extended Range Forecast Division of the National Meteorological Center (NMC), recognizing the role that numerical (dynamic) models of the atmosphere were to play, was, with others, developing methods to use the output of those models (Klein et al., 1959). The models were simple in those days and produced, essentially, only geopotential heights at one, or at most a few, pressure levels. Klein et al. (1969) developed nearly concurrent regression relationships between pressure level heights and surface maximum and minimum (max/min) temperature, then applied those relationships to the output of the numerical models to yield temperature forecasts valid about the same time as the model output valid time. Because the model was assumed to be "perfect" for this purpose, the method was eventually, according to Klein (1989), dubbed "perfect prognosis" or "perfect prog" by Veigas (1966). Forecasts from perfect prog relationships were used to forecast max/min temperature at NMC as early as 1965 (Klein et al., 1969, p. 1) and temperature forecasts were transmitted to field stations by teletype starting in 1968 (Weather Bureau, 1968a).

Another method of making use of the output of numerical weather prediction (NWP) models was to actually derive the statistical relationships between the desired predictand and the output of the model. This necessitated a sample of the model and, therefore, a fairly extensive effort just to test the concept. The name coined for this method was Model Output Statistics (MOS), for obvious reasons.

During the early years of the Techniques Development Laboratory (TDL), both the perfect prog and MOS approaches were actively pursued (Klein, 1970). Eventually, it became evident, at that rather early stage of numerical model development, that MOS was the preferred approach, at least for most

applications (Klein and Glahn, 1974).[1] It is around MOS that the NWS short range interpretive guidance system is built. The first MOS forecasts were disseminated to field sites in 1968 (Weather Bureau, 1968b).

Because MOS development requires (1) a sample of model data to be collected over a period when the model does not undergo major changes and (2) the model to still be basically the same in operation as it was during the data collection period, the desirability of switching the guidance system to perfect prog has been discussed many times. Most recently, it has been proposed that a system based on the perfect prog concept be developed and implemented (1) to support advanced capabilities in product preparation now being developed and (2) to provide a mechanism for testing new models, and versions of them. The latter is being proposed in order to test the models in terms of aviation-critical weather elements, elements that the models are unlikely to forecast directly, at least with the required skill.

This office note arises out of an action item assigned to TDL by a subgroup of the Advanced Weather Interactive Processing System (AWIPS) Requirements Task Team (ARTT) when it met in Boulder, Colo., in December 1990. TDL was asked to address the issue of using perfect prog for the above stated purposes.

As the reader might expect, since TDL has relied almost exclusively on MOS for producing short range (up to 60 hours) guidance, this note will find more virtue with MOS than perfect prog. The reader will have to judge whether this is unwarranted bias on the part of the author or well-founded conclusions.

2. DEFINITIONS

A number of statistical "models" can be used to relate predictands to predictors including scatter diagrams, contingency tables, decision trees, regression, discriminant analysis, canonical correlation, logit, map typing, analogues, and self-adaptive techniques. Many of these models are reviewed by Glahn (1965, 1985), and considerable experimentation was carried out with them in the early days of TDL and even earlier than that in the Short Range Forecast Development Section of the Weather Bureau's Office of Meteorological Research. While each model is interesting and has its own set of strengths and weaknesses, it became clear that linear regression, together with all the variations and innovations that can be attached to it, is undoubtedly the most appropriate for large interpretive systems. Linear regression is used almost exclusively in the U.S. MOS and perfect prog work and is used heavily in other countries. Discriminant analysis is also used, especially in Canada, when probability forecasts are desired, and may in certain instances provide as good or better results, depending on what characteristics are desired for the forecasts. However, discriminant analysis consists of two parts: (1) the development of a set of discriminant equations (not unlike regression

---

[1]Quoted from Klein and Glahn (1974), p. 1218, "Until last year our most notable application of the perfect prog method was the preparation of automated forecasts of maximum and minimum surface temperatures.... However, in August of 1973, the operational system for making these forecasts was shifted to MOS because of its greater accuracy and convenience...." Table 1 in that reference shows MOS to be better than perfect prog by about 0.5°F mean absolute error for a 3-mo test in the spring of 1972.

equations and for the two-group case, equivalent--the coefficients are proportional--to regression) and (2) the use of those equations to determine probabilities, since the discriminant functions have no bound on variance and do not produce probabilities directly. For efficiency in both development and operation, since discriminant analysis does not provide a clear advantage, TDL uses regression for all interpretive work. Therefore, any examples of predictand-predictor relationships given in this paper will be in terms of regression equations.

Essentially, any of the statistical models can be applied to any of the three methods described below--classical, MOS, and perfect prog.

A. Classical

As mentioned in the Introduction, early uses of statistics for meteorological predictive purposes were exemplified by the fact that the "lag" (or projection--the time between the input data and the forecast valid time) was built into the relationship between the predictand (what is being predicted) and the predictors (the variables being used to make the prediction). Since the use of statistics for this purpose predated NWP models, no other viable approach was available. Scatter diagrams (sometimes called graphical regression) were used extensively in this way in the 1940's and 50's. This classical approach [probably so named by Klein (1969)] was formalized mathematically by the group at TRC.

Professor Hans Panofsky once remarked (private communication) that whenever anything (in science) is called "classical" it usually means it is wrong. (This statement was actually made in the present context.) That is probably true here. Certainly, when one is considering projections beyond a few hours, today's numerical models must be used in order for the product to be useful. While numerical models have not been very helpful within the range of up to 1 hour, the reasons are largely lack of appropriate, fine scale input data; incomplete understanding of the physical processes of particular importance on the scales of a few minutes and a few kilometers; and insufficient computer power to adequately test concepts and theories. Models useful for projections of less than an hour will some day be available and will be used. Even today, if one considers simple advection and extrapolation to be "models," then statistical guidance techniques should not ignore models for projections under an hour (e.g., the approach of a line of thunderstorms identified by radar can be dealt with by advection or extrapolation for a few tens of minutes, and while the forecast will be far from perfect, will be better than not using these simple processes). The definition of "classical" is usually not thought to encompass these simple "models," except that a calculation of instantaneous advection or past movement of a variable could be a predictor. That is, simple calculations from current or past data are fair game, but trajectories and future positions of air parcels involve "models" which would not require the lag to be built in, and are, therefore, outside the scope of classical.

B. Perfect Prog

The concept of perfect prog is quite simple, and its early and successful use was correspondingly simple. All that is required (at least until one considers today's circumstances of sophisticated, high resolution models and the specific requirements for weather element guidance) is a sample of historical data consisting of observations of the desired predictand(s) and

observations of the predictor variables for which forecasts will be available from the implementation NWP model. In the first applications, the predictand was max/min temperature covering a 24-h period (very unspecific in time) and upper air heights (heights of constant pressure surfaces) at one, or at most a few, standard levels. Such heights were all the models were capable of producing in those days. Perfect prog implies that there is (essentially) no time lag built into the equation. At least, whatever it is, it applies to each projection for which a forecast is to be made.

Even though there is some problem of "concurrency" with this predictand (covering a 24-h period), relationships were found between the max (and the min) temperature at particular stations and heights (or thicknesses--differences between heights) at gridpoints. These relationships (a different one for each station and for max and for min) were applied to the model available at the time--starting with the barotropic. A predictor can also be the previous day's max (or min) temperature. That is, this observation would be the corresponding temperature 24 hours before. This is a "lagged" predictor, and being the same variable as is being forecast, the specification/predictive relationships can be applied iteratively. That is, for tomorrow's temperature, use tomorrow's forecast heights and today's observed temperature; for day after tomorrow's temperature, use day after tomorrow's forecast heights and tomorrow's predicted temperature, etc. Note that the definition of perfect prog is taken from the use of the model data and does not preclude the use of observations as predictors (Klein et al., 1969).

C. Model Output Statistics

The concept of MOS is also quite simple. The sample of data needed for MOS differs from that needed for perfect prog in that the predictor variables (those to be used from a model) must be from a model--not observations or analyses from observations. Different relationships (e.g., regression equations) are developed for different projections. That is, a temperature to be predicted 24 hours from initial data time will have model predictors valid at or near the predictand valid time. In addition, the current observation--the one that will be available when the equation is actually used for prediction--can be used. This observation is extremely important for very short range prediction (on the order of hours) and of little or no importance for medium range prediction (on the order of days). Because there is a different relationship for different projections, the lag of the observation is built in appropriately, and the equations are not applied in an iterative fashion as they are (or can be) with perfect prog. Note that, as with perfect prog, the definition is taken from the use of model data and does not preclude the use of observations as predictors (Glahn, 1970).

3. WEATHER VARIABLES FOR WHICH GUIDANCE IS NECESSARY

In the early days of interpretive guidance, concentration was on max/min temperature, as stated earlier, and probability of precipitation. This is very understandable; temperature and precipitation are unquestionably of prime importance in public weather products. Also, centrally-run models, with a twice per day cycle, are not as useful for the shorter-range aviation products, which are issued three or four times per day with, sometimes frequent (alas), amendments. Another intuitive, if not explicitly formulated, reason was that the aviation-related weather elements of ceiling height, visibility, and cloud layers are much more difficult to deal with statistically because of

having highly nonnormal distributions and not having simple, physical relationships with model predictors. Precipitation occurrence posed some similar problems, but not nearly to the degree of the aviation-related elements mentioned.

In the past, and even today, very short range forecast products are prepared manually and issued to the user. If guidance is available--fine; if not--the forecaster can manage. In fact, because of the relatively low skill level, which is related to the twice per day cycle, some very short range guidance is not heavily relied on.

In the modernized Weather Service, most products, other than the very short-fused warnings and watches, will be prepared by computer from a digital database (National Weather Service, 1987, p. 5-8). This digital database will hold the "official" forecasts for specific points, specified elements, and specified projections. Formatting software will use those numbers to mold the final official products--with final editing by the forecaster possible as needed.

Although the advantages of the digital database/product preparation (DD/PP) concept will undoubtedly prevail, there are drawbacks--a principal one being that these digits have to be put there by the forecaster. When one considers the enormous number of values needed to adequately describe the future weather in time and space and to keep that description relatively current, the inescapable conclusion is reached that there must be very good guidance values for essentially all elements in the products to be automatically prepared, else the forecaster won't have the time to prepare the database of digital values.

Another conclusion that can just as readily be drawn is that there must be a very efficient and user friendly system to allow the forecaster to interact with the existing values--be they guidance or the current official values--and to modify/update them as desired.

So, the NWS cannot be satisfied with guidance for well behaved elements such as temperature and wind, but must also deal with the problem children, and in sufficient detail and accuracy to make the guidance useful.

4. RELEVANT EXPERIENCE

Many countries have interpretive guidance for at least some weather elements. In the World Meteorological Organization (WMO) (1991) document Numerical Weather Prediction Progress Report for 1990, a total of 30 countries reported NWP involvement of which 11 reported MOS activities, 7 reported perfect prog activities, and another four reported interpretive work but didn't label it as MOS or perfect prog. Of those countries reporting MOS or perfect prog work, four reported both--Canada, United States, France, and the Democratic Peoples Republic of Korea. This may not be a very accurate assessment, because some countries may not report interpretive work and some may report only changes in the operational system, rather than the full scope of such activities. For instance, we know The Netherlands has an interpretive system, but didn't report such activities.

Both the U.S. and Canadian systems are well documented in English and in publications readily available in the United States. These include summary

papers by Carter et al. (1989), Wilson (1985), and Yacowar and Verret (1991), each of which contains many references to specific applications. In addition, lecture notes and invited papers for a WMO Training Workshop on the Interpretation of NWP Products in Terms of Local Weather Phenomena and Their Verification are contained in Glahn et al. (1991). This latter document is intended to be a tutorial for developing an interpretive system--including basic statistical concepts and forecast verification--and the current status of interpretive weather forecasting.

As an indication of how the U.S. system is regarded, I quote L. Wilson (1985) from Canada when delivering an invited lead paper at the American Meteorological Society's Ninth Conference on Probability and Statistics in Atmospheric Sciences at Virginia Beach in 1985, "The U.S. operational MOS system, the most complete set of statistical forecast products in the world, has been enormously successful, and has been carefully watched in many other countries."

5. CHARACTERISTICS OF MOS

A. Advantages

The advantages of MOS boil down to one major one--greater accuracy than any other approach, given, of course, an adequate sample and a stable model. This is hardly refutable either on theoretical or experiential grounds. Statistical relationships to be used on a sample of data should be developed from another sample drawn from the same population. Only if the predictors in both the developmental and operational samples are from a (the same) forecast model can this be true (both samples being of observations doesn't, of course, fulfill the goal of prediction). Early experiments in TDL showed MOS decidedly superior in accuracy. In Finland, a perfect prog system for temperature based on the ECMWF (European Center for Medium Range Forecasts) model was replaced by a MOS scheme after "Several case studies...have shown...that a new NWP interpretation scheme was in need" (Nurmi and Kilpinen, 1991).

The degree to which MOS is better than perfect prog can be considerable and can easily make the difference between whether a forecaster is willing to accept guidance values or not. Good comparisons of accuracy between the two which are relevant today are hard to find. One early comparison has already been noted in the Introduction for short range max/min temperature (Klein and Glahn, 1974). A more recent comparison of max/min temperature is given by Klein (1982) and indicates MOS to be better than perfect prog by 0.2°F in 1974 and by nearly 0.5°F for each of the years 1975 through 1979, the average being over 126 stations, two cycles, four projections, and 12 month periods. Another study (Carter et al., 1989) compared MOS based on the LFM with both "traditional" perfect prog and "modified" perfect prog systems applied to the NGM for the test (independent) data period May to September 1987. Results indicated that, for four projections combined, the two perfect prog systems were about equal in accuracy, and that MOS was about 0.7°F better in terms of mean absolute error.

In comparing the operational LFM forecasts with perfect prog forecasts based on the NGM, Erickson (1988) found that perfect prog NGM guidance was better for surface wind forecasts, that LFM MOS was better for probability of precipitation (PoP) forecasts, and that both were about equal in skill for cloud amount forecasts. Forecasts verified were for 204 stations and four

projections for the cool season of October 1986 through March 1987. Overall, when we consider the four weather elements--max/min temperature, PoP, surface wind, and opaque cloud amount--the LFM MOS produced better forecasts than the NGM perfect prog, even though the LFM is considered to be less skillful than the NGM.

Very recent results have shown that MOS forecasts made for these same four weather elements with equations developed on three seasons of data are better than LFM-based MOS forecasts or NGM-based perfect prog forecasts, even though the NGM has not been without change during the developmental and test periods.

Dallavalle (1988) compared the relative skill of MOS and perfect prog max/min temperature forecasts for medium range projections of approximately 3 to 6 days. The results indicated that the perfect prog system provided more skillful guidance than did MOS for all projections. However, the MOS equations had been developed on data prior to the implementation of major enhancements to the medium range forecast model. These changes in 1987 involved increasing the horizontal resolution, improving the surface physics, and including moisture in all 18 vertical levels of the model (Sela, 1988). In essence, then, the MOS equations were applied to a dynamical model that differed substantially from the one used to define the equations.

B. Disadvantages

The disadvantages of MOS also boil down to one major one--the need for a historical developmental sample from the implementation model for projections as far out in time as guidance forecasts are needed. Is this serious? It certainly can be and will always be thought to be by model developers who would like to see the fruits of their labor used immediately--that is, (1) a change made to an operational model (because of forecast quality or efficiency of the forecast "system" arguments) and the interpretive statistics still be applicable or (2) a new model supplant an existing one and interpretive statistics be immediately available for it.

How long a sample is needed? There is no firm answer, and details of the development will adapt insofar as possible to the sample available. It is generally thought that one season of data is rather short, two are certainly useable, and three or more are decidedly better, a season being a 6-mo warm or cool period (e.g., April through September). With larger samples, more "regions" will be used (a region is an area or group of stations for which data are pooled for a "regionalized" relationship) and more seasons will be used (e.g., four, 3-mo seasons, rather than two, 6-mo seasons). Smaller regions and shorter seasons lead to better forecasts, provided the sample size can support them.

A minor disadvantage is the relatively large number of relationships that are needed (e.g., regression equations), due to a different relationship being determined for each projection (for each variable and each initial data input time). This can be thought of as a disadvantage because a (pure) perfect prog system would not require different relationships for different projections. Development of more equations certainly takes more time and work; however, the developmental process for one projection is the same as for any other, and many of the decisions made for one projection do not have to be reevaluated for other projections. For instance, one would likely use the same regions, same seasons, and basically the same potential predictors (at different

projections, of course) for all predictand projections. This would not be a disadvantage were some "modified" perfect prog system to also require separate equations for different projections.

6. CHARACTERISTICS OF PERFECT PROG

A. Advantages

The advantages of perfect prog, when we are comparing to MOS, mirror the disadvantages of MOS--what is an advantage for one will be a disadvantage for the other. The advantages are two: (1) the relationships can be applied to any model, and (2) only a single relationship need be developed between a predictand (e.g., max temperature or occurrence of precipitation) and predictors. These are truly advantages; the question is, "Can reasonable accuracy be achieved and still retain them?" Note that these advantages presuppose the definition of "pure" perfect prog in Section 2.B, which follows the defining use of perfect prog in the early 1960's. Adjustments to the definition can be made for an actual implementation, but should be made with the realization that they <u>are</u> adjustments which will undoubtedly erode the advantages as stated here.

B. Disadvantages

In the final analysis, the disadvantages to perfect prog all have to do with accuracy. However, the considerations in trying to achieve accuracy may, at times, seem more associated with implementation problems than with accuracy. Yet, those problems exist only because we must be concerned with accuracy. The major disadvantages are discussed below, not necessarily in order of importance, under headings which only introduce and not define the specific problem.

<u>Autocorrelation of Predictand</u>

For forecast projections of a few minutes or even a few hours, the current observation is of prime importance. Especially for the aviation-related weather elements, it would be foolhardy to make a 1-h, say, forecast without considering the current observation at the forecast location, provided it is available. If it is not available, some substitute should be found--a recent one in time for the same location or one implied from surrounding observations at the same or previous times.

So, for a 1-h forecast, let's include the current observation, along with observations of variables that will be predicted by a model, in a regression equation. This is a reasonable procedure and can be used in operation. Now, what about a 2-h forecast? We can't use the same equation, because the influence on the forecast by the current observation should diminish at 2 hours over 1 hour. We can, of course, develop another equation differing from the first only in the coefficients of the predictors such that the autocorrelation of the observation is properly accounted for. This new equation can be used in operation. Now, what about a 3-h forecast? See the problem?

In short, we are faced with developing a different relationship for each projection out to a few hours. We can do this, but we no longer have a (pure) perfect prog process, because the lag is specifically accounted for and

different relationships are required for different projections. This possible modification of the concept has scuttled some of the appeal of perfect prog--one relationship for all time.

I want to emphasize again, using the observation of the same weather element that is being forecast for a 1-h forecast is not optional. It would be far better to use a regression based only on the initial observation (a markov process) than to ignore the observation and use everything that any existing dynamic model can produce, unless one is not concerned with accuracy or skill. (Only if the model is so simple that the actual observation is very conservatively carried forward in time, may this statement not hold.)

For projections > 24 hours, the current observation is of such reduced importance that it could reasonably be ignored. If used (properly), it may furnish some information over and above model predictors, but not sufficient to mandate its use.[2] For that reason, the use of a single relationship that does not directly account for autocorrelation of the predictand is more practical for projections of > 24 hours than for projections < 24 hours.

Vertical Resolution of Data

Another problem with perfect prog, that is not limited to projections of < 24 hours but tends to be more important there, relates to the vertical resolution of upper air observations and the difference between actual elevation and model elevation of the ground at predictand points (stations).

It is the general practice to interpolate upper air variables to station locations (the locations where the predictand observations are) before determining the specification/predictive relationships. That is, predictors are "at or above" the predictand. However, this is not mandatory; early uses of perfect prog employed the "field" concept (Klein et al., 1959). Predictors were at gridpoints, while predictands were at stations. A particular predictor variable could be used from more than one gridpoint. This procedure is viable, and use of a variable at more than one gridpoint allows the "interpolation" to the station location to be built into the equation. However, when one considers, for today's environment, the large number of model variables at one point, to multiply this by the number of gridpoints to be used in regression and/or selection of predictors increases the size and complexity of the procedure. Also, this process builds in distance and directional relationships specific to a station and makes development of regional equations (pooling of data for several stations) difficult if not impossible.[3] Finally, to use gridpoint data in perfect prog requires analyses of upper air data to gridpoints (which is, of course, possible) and implies that that same grid will be available from the implementation model. With the model as yet unspecified, this is a problem. Implementation would require that the implementation model be interpolated to the developmental grid if the grids

---

[2]The 24-h dividing line is, of course, somewhat arbitrary; if anything, it would be better to put it at a greater projection rather then lesser.

[3]A predictor can, though, have a definition which would incorporate an "offset," provided that offset were the same for each station. For instance, the temperature 100 km to the west of the station could be a predictor. For this discussion, this is still "interpolation to the station."

were not identical.[4] So, for several reasons, interpolation to the predictand point is preferred. For ease of discussion, this interpolation is assumed in the rest of this paper, although the arguments and conclusions in no way depend on it.

So, how do we interpolate upper air observations to a specific point? A satisfactory, and possibly preferred, way is to objectively analyze the observations to a particular grid (the specific one does not much matter, just so it can capture the essential information in the observations) and then interpolate from this regular grid to the necessary predictand locations. This can work well for heights, temperatures, wind, and moisture at mandatory (for reporting) constant pressure levels, and, in fact, some such analyses already exist and can be used. There is more of a problem if one tries to use fine scale vertical detail. This would likely require a vertical interpolation of some sort followed by the horizontal objective analysis. That is, in order to make an analysis at 950 mb, say, either values are needed at 950 mb at each upper air location or a much more sophisticated objective analysis procedure would be needed--one that probably doesn't exist in a form that could be used for this purpose without considerable tuning.

Suppose analyses were to be made, should they be made at constant pressure levels--which is most convenient from the way the observations are available--at constant height above ground, or what? Since a particular pressure level may be "above ground" in one model and "below ground" in another model (model terrain does not agree well with actual terrain at gridpoints and even if it did, interpolation to station locations would not produce the same values for those models unless the grids were identical), it's difficult to see how data at constant pressure levels could capture actual fine scale, low level detail that could be successfully related to a model's forecasts of that detail. This suggests that the analyses be in a coordinate system other than pressure, perhaps distance above ground or a sigma (terrain following) system. Whatever is done, it won't conform to each and every future model, so adjustments will have to be made at implementation, probably vertical interpolation of model output.

While this vertical definition/resolution problem may not be insurmountable, it certainly complicates the picture and rules out "simple" systems that can be easily developed for nationwide implementation on models of choice.

Diurnal Trend of Predictand

Another consideration, which applies equally to very short range projections as well as to longer ones, is the diurnal trend of the predictand. If the predictands are, for instance, temperature (or wind, or dew point, or ceiling height, etc.) at 3-hourly intervals (or 6-hourly, or 1-hourly, etc.), then the relationship of those predictands to model predictors will vary with time of day, at least with existing models.[5] If a different relationship is necessary

---

[4]Variables from a spectral model, however, can be evaluated at the developmental gridpoints just as easily as at any other set of points. While global models may well be spectral, mesoscale models will likely not be.

[5]Some day, models may exist that have sufficient detail in their various parameters that the diurnal cycle is adequately captured. When that time

for each predictand time of day, then some of the appeal of perfect prog is lost.

Upper air observations are available in sufficient quantity and quality to furnish a basis for developing predictand-predictor relationships only at 0000 and 1200 UTC. If we want a perfect prog forecast valid at 1200 UTC, then a relationship can be developed between the predictand at 1200 UTC and upper air data (or analyses of them). Suppose we want a forecast at 0600 UTC; what upper air data do we use? Possibilities are (1) that we use both 0000 and 1200 UTC upper data, (2) that we interpolate between 0000 and 1200 UTC to 0600 UTC, and (3) that we use 6-h forecasts from some model valid at 0600 UTC. This model wouldn't be the implementation model (because the archive wouldn't exist), and more than one model could be involved. The model, for this purpose, would be used as an interpolation device to get "observations" valid at 0600 UTC. The latter two possibilities are probably better than the first. If interpolation is done, some nonlinear method may be advisable, which increases the complexity somewhat over linear interpolation. No matter what procedure is used, the quality of the forecasts when the relationships are applied to a (any) model will likely be less at 0600 UTC than forecasts valid at 0000 and 1200 UTC. The result will be an oscillation of skill over a 12-h, or maybe a 24-h, period.

Given that intermediate values are to be obtained by interpolation or from an existing model, an alternative to multiple relationships is to include one or more time-of-day predictors in the regression (e.g., one or two harmonics of the 24-h clock). This will undoubtedly reduce the accuracy somewhat, but perhaps not significantly. An important point to note here, however, is that the question is not just whether the diurnal cycle of the predictand can be modeled by a couple of harmonics and, therefore, constants in the regression equation will account for it, but whether the relationship of the predictand to the predictors changes with time of day, and, therefore, additive constants in the equation won't be sufficient to capture the cycle. That is, should the coefficients of the model predictors change with time of day? Very innovative use of computed predictors (see Glahn et al., 1991, Chapters VII and X through XIII) would probably be necessary to use one relationship for all times of the day.

Variance of Forecasts

The most important accuracy problem, once we are beyond the vital importance of the current observation, is the inability of a perfect prog system to lessen the variance of its forecasts with time. This is of most, and in fact critical, importance for probability forecasts. For instance, the occurrence of precipitation, dealt with as a binary predictand, will produce a probability forecast (see Glahn et al., 1991, Chapter XII). The regression output can vary from less than zero to greater than one. Probabilities should, of course, be bounded by the 0 to 1 range; regression does not insure this,[6] but

---

comes, then we are in the age of direct model output and interpretation will be largely unnecessary. We are a long way from that now and will still be in 2001, Hal notwithstanding.

[6]Discriminant analysis in a reasonable implementation will keep the values within the 0 to 1 range, but is not necessarily better, or worse, than

there is no practical problem. Values < 0 can be increased to 0; values > 100 can be decreased to 100.[7]

One would like the objective interpretive procedure to produce the full range of values, and for no rain to usually occur with near 0% forecasts and rain to usually occur with near 100% forecasts. Unfortunately, the current state of the science is such that this accuracy is not possible, even for very short range forecasts (consider, for instance, a vicissitudinous summer shower). But, on the other hand, there are situations when 24-h numerical model output will indicate so clearly that rain (no rain) will occur, that 100% (0%) can be stated with confidence.[8]

One of the noncontroversial aspects of a probability (forecast) is that the event for which the probability applies should occur with (about) the relative frequency of the stated probability.[9] "Reliability" is the term coined by Sanders (1958) to denote the correspondence of the relative frequency of the event to the forecast.[10] In order to determine the reliability, we have to collect data for many events and may even aggregate over points (station observations) as well as time (e.g., 6-mo periods). While regression does not guarantee reliability even on the developmental sample, the developer can devise predictors capable of achieving reliability within acceptable limits. Many studies have shown this to be true within a very few percent for all forecast values between 0 and 100% (e.g., Glahn and Lowry, 1969; Bocchieri, 1974; Murphy, 1985; Wilks, 1990; and Glahn et al., 1991, Chapter XII, Fig. XII-6).

With MOS, good reliability is achieved <u>for all projections</u> because a different relationship is developed for each projection. High probabilities are not forecast for long projections, because the accuracy of the numerical models does not warrant it. The same is true for very low probabilities, but to a much lesser extent. The relative frequency of precipitation over a 12-h period is considerably less than 50% at most stations, and regression estimates tend toward the mean as the skill goes to zero. With a predictand mean of 30% (overall relative frequency of precipitation), a departure from it of

---

regression in other regards, except that it is more complicated to develop and implement.

[7]In this two-category situation, not only is there no problem, but some would argue, perhaps rightly so, that a -10 is a better indicator (from this objective procedure, even though it can't be considered a probability) of no precipitation than 0, and 110 is a better indicator of precipitation than 100.

[8]Perhaps exactly 100% and 0% should not be used, because they indicate absolute certainty that rain will and will not occur, respectively. However, for practical purposes, and especially if we round the forecasts to the nearest 10%, using these values is reasonable.

[9]We do have to be careful that the "event" is precisely defined. For precipitation occurrence, this is defined in the NWS by $\geq$ 0.01 inch liquid equivalent at a point in some specified period of time, say, 12 hours.

[10]Reliability also goes by other names; even Sanders (1963) adopted Bross's term "validity" (Bross, 1953).

20% can give a 10% forecast (rather close to 0%) or a 50% forecast (not very close to 100%). But, with MOS, whatever the range of forecast probabilities, those forecasts are quite reliable.

On the other hand, perfect prog will give the same range of probabilities at 120 hours as at 12 hours. The only reason this might not be true of a pure perfect prog relationship is that the numerical model itself would have bias characteristics or go toward the mean circulation in such a way that the (range of the) forecasts would be affected. A 100% forecast of precipitation for the period 12 to 24 hours after model cycle time is reasonable and desirable; a 100% forecast for the period 108 to 120 hours after cycle time is not justified with today's models.

The verification score the NWS uses to judge precipitation probability accuracy is one-half the score P defined by Brier (1950)--the so called Brier score. The Brier score is the mean square error, where error is defined as the difference between the probability forecast and either 1 or 0 depending, respectively, on whether the event occurred or didn't occur. The regression relationship actually minimizes this score on the developmental sample--no other linear relationship will do better. The Brier score is highly influenced by large errors, and much has been written about its "components" and how lack of reliability affects it.[11] Suffice it to say, that this score will be poor (large) if forecasts are made over the whole 0 to 100% range and such forecasts are not warranted (aren't reliable).

An argument is sometimes made that a perfect prog forecast is a conditional forecast and is the correct forecast given that the model is correct. This may be an interesting piece of information, but it's a practical impossibility to use that information to calculate the unconditional forecast. (The unconditional forecast is still based on the model, but does not depend on its being correct.) To do that, the user of this information would have to know not only the probability of occurrence of that particular set of values of the model predictors in the regression equation but a whole ensemble of conditional probabilities and associated unconditional probabilities of model predictors. Alternatively, the forecaster can use his/her judgment to arrive at the unconditional probability, but surely that process would be easier and more exact if the objective unconditional (MOS) forecast were available rather than the objective conditional (perfect prog) forecast.

The two-category event discussed above is simpler in some respects than a multi-category event. Regression estimates outside the 0 to 1 range are more of a problem with more than two categories because adjustment of one category to the proper range should occasion an opposite adjustment to another category or categories, and a reasonable algorithm to use for adjustment may differ with predictand.[12] However, this is usually not a serious problem, since the estimates don't stray far outside the proper range.[13] It is more of a

---

[11]The first discussion of this in the meteorological literature was probably by Sanders (1958).

[12]For the two-category situation, the algorithm is obviously just to add to (subtract from) one category what is subtracted from (added to) the other.

[13]If equations have been developed on a very short sample, instability may

problem for perfect prog than for MOS because the departure of the forecast probabilities from the 0 to 1 range decreases with projection for MOS but does not for perfect prog. In fact, model biases (to be discussed later) could cause severe problems in this regard for perfect prog.

To some extent, the same basic argument used for reliability of probability estimates can be used for estimates of a quasi-continuous predictand such as temperature. However, it's not quite so clear that one wants an "unbiased" temperature forecast as it is that one wants an unbiased probability forecast. That is, in hot, dry conditions with a persistent upper air pattern, does one want a 3-day, say, forecast to trend toward climatology or not? If it does not, then the average observed temperature for those occasions (over a long period of time) when very high temperatures are forecast will be lower than the forecast average. The temperature will have been "overforecast," and the verification score will suffer. Although a continual, day-after-day regression toward the mean may not be desirable in a drought situation, one has to consider that the model may produce nearly the same set of predictor values in non-drought situations in which a (near) record forecast temperature would be inappropriate at the 3-day range. Interpretive systems are not yet smart enough to sort out the very persistent situations, in which the model may be quite accurate for several days in a row at long projections, from the non-persistent situations, in which the model forecast skill is considerably reduced. On the whole, it seems the forecaster could more easily "beef up" the MOS forecasts if necessary in unusual, persistent conditions than to always deal with conditional (perfect prog) temperature forecasts. Adjustments would undoubtedly, on the whole, be larger and more frequent for perfect prog guidance than for MOS.

Near-Perfect Predictors

Another problem with perfect prog is how to incorporate "near perfect" predictors. For instance, suppose that one assumes the implementation model will produce a "surface" temperature and this will be an important predictor for an objective, specific-time temperature prediction. Therefore, for the predictand temperature, a predictor would be the temperature at that same location and time. This is no problem for MOS, because the predictor in the developmental sample is from a model and contains the appropriate errors. However, for perfect prog, in the purest sense, the predictor and predictand in the developmental sample would be the same and the relationship would be perfect. No other predictors would (could) be put into the equation. In operation, the objective estimation would be exactly the model temperature at all projections. This is not a tolerable situation. What to do?

---

show up as values being outside the 0 to 1 range. Also, if the predictand is conditional, then the results of applying the equation should be expected to be reasonable only in situations in which the "condition" is reasonable. That is, if we have a regression equation to produce the probability of each of three categories of liquid precipitation (steady rain, drizzle, and showers), then the results are appropriate only if the probability of liquid precipitation is not low. This is because the relationship is developed only on cases of liquid precipitation (as it must be), and if the atmospheric conditions are not right for liquid precipitation, then the input to the equation would likely be outside the domain of the developmental data--a no-no for the application of a regression equation.

If the model temperature is to be used as a predictor (observed temperature in perfect prog development), then it must be somehow degraded! Possibilities are (1) some random error with a mean of zero be added to it; (2) an observed value offset in space be used; (3) an observed value offset in time be used; and (4) an actual model forecast for a very short projection, such as 6 hours, be used. All of these are possibilities, the latter two being probably the best and have been used by TDL (Erickson, 1988; Dallavalle, 1988), but all seem artificial and arbitrary. (The model here is just for degradation, and the same model would not have to be used in operation.)

The "perfect predictor" malady doesn't affect predictands that have a poor (not direct) relationship with variables from a model. For instance, ceiling height or visibility would not likely be a direct model output and therefore would not be a predictor in a regression equation. However, most models produce precipitation amount and this, sometimes used only as precipitation occurrence, is a very important predictor for precipitation occurrence. Observed precipitation can't be used as a predictor for the predictand precipitation occurrence without some "tinkering."

Model Biases

Nothing much has been said yet about model biases and their effect on the forecasts. One might suppose that overall model biases would be eliminated (either by direct dynamic modelling, or statistical error feedback correction) before implementation, but experience has not shown this to be the case. For instance, the first set of perfect prog max/min temperature equations applied to an early version of the Nested Grid Model (Hoke et al., 1989) produced very inaccurate guidance because the primary predictor, the 1000-850 mb thickness, exhibited an extreme cold bias (Jensenius, 1988).[14] That problem had to be solved by removing the 1000-850 mb thickness predictor from the equation. That is, the most important predictor could not be used, and another development was necessary--another scratch on the myth that perfect prog relationships can be developed once and for all. Persson (1991) from Sweden states, "It is a common fact that NWP models exhibit systematic error in the forecasts of the near surface weather elements. The 2 m temperatures, for example, are often systematically biased, though the magnitude of this bias varies with geographical location and time of the season." Nurmi and Kilpinen (1991) show a cold bias of direct output of the High Resolution Limited Area Model (HIRLAM) developed jointly by Denmark, Finland, Iceland, Norway, and Sweden which varies between 1.5°C for 6-h forecasts to 0.6°C for 48-h forecasts for January 1991. Note that systematic bias is accounted for by MOS even though it may vary with projection.

Predictability of Model Variables

The different levels of skill with which a particular model predicts its parameters is of considerable importance in the interpretation of that model. While this topic is related to the previous three topics discussed, the emphasis here is different. It could well be that a model is quite good at predicting temperature at a level above the boundary layer (say 850 mb) but quite poor at predicting temperature a few tens of meters above the ground. This poor predictability might be characterized by model bias discussed above,

[14]This problem with the NGM has now been largely eliminated.

by an inability to follow the diurnal trend, by apparently random errors due to numerical techniques employed in the model, or even by a reversal of the diurnal trend.[15] Another model might have entirely different predictability characteristics.

It is important that predictors be used that reflect the skill of the model. MOS does this automatically through a predictor selection technique associated with regression (or discriminant analysis). For perfect prog, we have to guess, in effect, as to what predictors to include and exclude. What may be a good predictor for one model may be a poor one for another model. This leads to conservatism in perfect prog predictor selection; that is, we have to exclude variables that may be the least predictable by NWP models. Unfortunately, this tends to exclude low level variables and those related to fine scale detail.

Thresholding for Categorical Forecasts

Many times, probability forecasts are processed to yield a "categorical" forecast. For instance, the probability of each of four categories of cloud amount may be used in an algorithm to produce a "best category" forecast. The particular algorithm may have one or more parameters that are tuned to the developmental sample.[16] If the developmental sample (observations) for perfect prog is used to derive/tune the algorithm, it may or may not be appropriate for projections beyond, say, 24 hours. To some extent, the problem exists with MOS, but there a model sample is available and the algorithm can be tuned to projection, if necessary.

7. DISCUSSION

If one understands and believes the information in Sections 5 and 6, he/she should be convinced that pure perfect prog is a long way from a workable solution. Reasonable accuracy cannot be obtained for all forecast projections without some, perhaps major, adjustments to the concept. At a minimum, different relationships would have to be developed for each projection < 12, or perhaps < 24, hours in which the observation was a predictor. Quite likely, different relationships would have to be developed for each predictand time-of-day, depending on the predictand variable. Difficult decisions would have to be made about what vertical resolution and variables to use from upper air observations in order to incorporate useful, fine scale detail that would, with a high probability, be available from the implementation model and be predicted with skill. Effects of model bias will be of concern each time the model is modified (as they are with MOS). Something must be done to make probability estimates reliable and, quite likely, to make distributions of forecasts match observed distributions reasonably well.

---

[15]The diurnal trend of wind in the NGM's lowest layer exhibited, at least at one stage of the NGM's development, a maximum at night (National Weather Service, 1986). While this may be the correct trend for this height above the ground, it is at the lowest level predicted and is not what one would expect for the surface wind.

[16]See, for example, Bryan and Enger (1967) and Bermowitz and Best (1979).

One technique that has been used in Canada is to postprocess the perfect prog forecasts (Yacowar and Verret, 1991). This can be done in various ways, depending on the predictand. An algorithm can be used that "goes in the right direction" without any dependence on the particular model used in implementation. Or one can collect a sample of perfect prog forecasts and matching observations and determine a statistical (MOS!) adjustment procedure. Note that this requires an implementation model sample. Since an adjustment to precipitation probability would undoubtedly depend on the forecast itself, a sample large enough to give an estimate of bias over the range of forecast values would be necessary. Given that models do not perform equally over different parts of a country, limiting somewhat the area over which data can be aggregated for this purpose, it's not hard to see why a significant part of a 6-mo season would be necessary to determine a reasonably good adjustment. Also note that this adjustment will not necessarily apply well next season if the model is changed or replaced, although the trauma might not be as great as if MOS were used. For a continuous variable like temperature, adjustments might reasonably, but with some danger, be based on a 2- or 3-mo sample.

Another method of postprocessing that is coming into vogue is some sort of self-adjustment (e.g., Yacowar and Verret, 1991) such as the Kalman filter (Persson, 1991; Simonsen, 1991). Such schemes correct for bias in the forecasts. There is always the question of how rapidly such a system should move toward less bias; that is, how long should a "bias" be exhibited before an appreciable correction is made? Parameters in the correction method can control this, and the degree of correction can even depend on how much data the existing system has been trained on. Once the various controls and options--that would necessarily vary with weather variable, projection, and location--were built in, it's not obvious that such a system, together with the necessary perfect prog (or some such) front end that would be required, would be less cumbersome than a rapidly-updatable, more conventional, MOS system. Also, such a system only corrects for bias--it doesn't change coefficients on individual model predictors. Yacowar and Verret (1991) indicate that their postprocessing corrections applied to cloud amount and PoP are based on 3 months of data and for temperature, 5 months of data.

The theory around which the Kalman filter is built (and implied in other similar schemes) may not hold for synoptic-scale meteorological processes. Regression does not imply much concerning time dependency of departures from the mean. If one uses one season of data to develop a MOS system to use on the next season, then the seasons should be similar, but nothing is implied about autocorrelation on the order of days. There is also the question of forecaster response to a continually changing interpretive system. He/she may come to expect a cold bias of temperature in very persistent, hot and dry conditions and will correct accordingly. To have the objective system correcting automatically throughout this persistent period might lead to confusion. How would the forecaster know how to correct, if at all? The object is not necessarily to get the "best objective" estimate, but to get a very good one that the forecaster can improve upon in special situations.

Any viable, operational system that has to deal with changing models as its primary input must be able to update the forecast relationships quickly and easily. For instance, if MOS were used to develop a temperature prediction system on two seasons of data, the developer must be able to add next season's data as soon as the season ends to develop and implement better forecast equations. Quite likely, the predictors would be identical, but the

coefficients would change. As necessary, subsets of the sample could be used. For instance, when the third season became available, if it were judged that the changes in the model were such that the first season of data should be excluded, it could be with little effort. Lengths of seasons and, to a lesser extent, sizes of regions could be adjusted as more or less data were available. Such an updatable scheme (Ross, 1987) has been in operation for temperature (World Meteorological Organization, 1991, p. 237) in the United Kingdom for a number of years.

In a like manner, if some modified perfect prog procedure were used, a data collection/adjustment system would have to be in place so that tuning of the perfect prog output could be done on at least a seasonal basis. A real-time error feedback scheme is attractive in concept and should eventually be possible, if deemed desirable. The necessary engineering for implementing such a system for all weather elements has not been done, would be considerable effort, and much testing--including forecaster response--would be needed.

A method has been proposed by F. Lewis and tested by Lewis and J. P. Dallavalle and others (private communication) in which a short sample of a new model is used to "calibrate" regression equations (they could be MOS or perfect prog) developed on a larger sample of data. Preliminary indications are that (1) the method is better than perfect prog, (2) the method is better than MOS based on one season of data, and (3) the method is not as good as MOS based on two seasons of data.

Although perfect prog seems at first blush to be simple to develop and implement, in actuality, to achieve quality guidance, it presents many more problems than MOS. If one considers some form of model-dependent postprocessing as necessary to achieve acceptable accuracy, then even the archiving step cannot be eliminated, although it would be of the statistical forecasts and not necessarily the model output. The statement has been made many times that two numerical models are more alike than either is like the atmosphere. As models get better, this will probably be more true in the future than in the past. If this is true, why is not MOS developed on one model and applied to a modified model or even a different one a better solution than "perfect" statistical relationships applied to imperfect models?[17] The methods and machinery necessary for developing and implementing MOS are better understood and more pieces are in place in the NWS than for developing and implementing a viable perfect prog system.

TDL has been able to develop a perfect prog system that when applied to the NGM gave good results, but only after considerable experimentation and knowledge of the NGM--the implementation model (Dallavalle, 1988; Carter et al., 1989).

We must remember that in the modernized Weather Service, it is not sufficient to produce guidance "in the right ball park" especially for projections beyond, say, 36 hours, but we must produce guidance that is so good that professional forecasters will be willing to declare it a hands-off, over the fence, home run on many occasions. So let's not lose sight of our major objective while searching for the holy grail.

---

[17]However, limited TDL experience has not shown this to be true.

8. USE OF INTERPRETIVE FORECASTS TO JUDGE ACCURACY OF NUMERICAL MODELS

Interpretive forecasts of max/min temperature and PoP have been verified over about a 20-year period. The trend is for gradual improvement. Most of that improvement is from improvement in the numerical models. However, even with aggregation over 6-mo seasons and nearly 100 stations to steady the statistics, the variability year to year is considerable (see World Meteorological Organization, 1991, pp. 292-293). One could not judge, based on a yearly score, that the model was improved or not from the previous year.

There are three basic problems associated with determining the relative quality of two numerical models based on the verification of weather elements (such as temperature and ceiling height) interpreted from those models: (1) the sample size necessary to make a determination of differences or sameness of statistics computed from that sample, (2) the quality of the interpretive statistics, and (3) the appropriateness of the scores (or measures) computed. Let's tackle these in reverse order.

Scores for temperature and PoP present little problem. For PoP, the Brier score or a climatology-based skill score (Hughes, 1965) and reliability diagrams (Murphy and Daan, 1985) are appropriate. For temperature, the mean absolute error, mean square error, or percent of errors greater than some threshold are usually appropriate. The solution is not nearly so clear for aviation-related variables such as height and amount of cloud layers and visibility. Even wind presents problems. Do we verify in terms of vector difference or in terms of direction and speed, or, in a particular application, in terms of cross-runway error. One must be convinced the scores computed are appropriate to the use of the forecast.

The quality of the interpretive statistics can be a major problem. For instance, suppose one model has a finer vertical resolution than a competing model. How does one insure that each gets a fair shake at being interpreted? It's probably not reasonable to use one perfect prog relationship for both, especially for very short range forecasts where vertical resolution of model output should be most important. Certainly one has to make sure that the interpretation is not so pathetically bad that neither of the forecasts is reasonable.

And finally, the sample size. It is almost imperative that the samples, one from each model, be matched. This makes for a much more robust test, and the paired t-test is a good candidate for providing a basis for judgment (see Glahn et al., 1991, Chapter IX). Depending on the resources required to run the models, the major sampling problem may be that not all kinds of situations and seasons will be covered. You could make 3-h forecasts every hour for 3 summer months and might be able to judge quite well whether one model was better than another on situations exemplified by those 3 summer months, but the judgment would not necessarily hold for winter months. Even another 3 summer months might give a different picture.

On the whole, the biggest problem is (2)--the interpretive statistics. The decision reached would include the whole forecast system--including both the numerical model and the statistical model. If that's the intent--to judge the combination--that's one thing. If it's the intent to judge the numerical model alone, that's quite another. It's doubtful enough effort would be put on the statistical component of the system to be able to adequately judge the

relative quality of two similar numerical models. The quality of numerical models is better judged on the basis of verification of the actual model output parameters rather than on variables interpreted from that model.

9. SUMMARY

The strengths and weaknesses of MOS and perfect prog approaches to providing interpretive guidance have been reviewed, as well as the possibility of comparing the quality of different numerical models based on interpretive statistics. In my judgment, the strengths (weaknesses) of MOS (perfect prog) considerably outweigh the strengths (weaknesses) of perfect prog (MOS). More developmental decisions have to be made with perfect prog than with MOS, leaving more room for judgmental error. Both MOS and perfect prog require meteorological decisions relating to forms of predictands and predictors, calculation of predictors to achieve "linearity," etc., but in addition, perfect prog requires pragmatic, meteorological and statistical decisions that can be neither right nor wrong, but only "reasonable" for an unknown model. Perfect prog makes a forecast that many times has to be modified before it can be considered a good "objective" estimate; MOS gives a useable value directly. Therefore, MOS does not need a postprocessing step,[18] while perfect prog undoubtedly would. It would require more software to develop and implement a perfect prog system that would probably produce mediocre guidance than to develop and implement a MOS system that would produce quality guidance. As long as there are changes made to models, one can never walk away from the interpretive system and say, "It's done." It will always have to be watched over and problems solved as they arise. I can't imagine that perfect prog equations developed in this decade will be best for use in the next. Maintenance and further development of a MOS system should require no more resources than a perfect prog system, and the quality of the guidance would undoubtedly be higher.

ACKNOWLEDGMENTS

I appreciate the assistance of Messrs. J. Paul Dallavalle and John J. Jensenius, Jr., who provided information and then read versions of this paper and offered helpful suggestions.

REFERENCES

Bermowitz, R. J., and D. L. Best, 1979: An objective method for maximizing threat score. Preprints Sixth Conference on Probability and Statistics in Atmospheric Sciences, Banff, Amer. Meteor. Soc. 103-107.

Bocchieri, J. R., 1974: A comparison between the single station and generalized operator techniques for automated prediction of precipitation probability. NOAA Technical Memorandum NWS TDL-53, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 20 pp.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. Mon. Wea. Rev., 78, 1-3.

[18]Determining categorical forecasts from probability forecasts is here considered part of the forecast system and not postprocessing.

Bross, I. D. J., 1953: Design for Decision. The Macmillan Co., New York, pp. 47-52.

Bryan, J. G., 1944: Special techniques in multiple regression. Unpublished manuscript, Massachusetts Institute of Technology.

_____, and I. Enger, 1967: Use of probability forecasts to maximize various skill scores. J. Appl. Meteor., 6, 762-769.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. Wea. Forecasting, 4, 401-412.

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, Amer. Meteor. Soc., 572-579.

Erickson, M. C., 1988: Development and evaluation of perfect prog guidance based on output from the nested grid model. Preprints Eighth Conference on Numerical Weather Prediction, Baltimore, Amer. Meteor. Soc., 565-571.

Glahn, H. R., 1965: Objective weather forecasting by statistical methods. The Statistician, 15, 111-142.

_____, 1970: A method for predicting surface winds. ESSA Technical Memorandum WBTM TDL 29, Environmental Science Services Administration, U.S. Department of Commerce, 18 pp.

_____, 1985: Statistical weather forecasting. Probability, Statistics, and Decision Making in the Atmospheric Sciences (A. H. Murphy and R. W. Katz, Eds.), Westview Press, Boulder, 289-335.

_____, and D. A. Lowry, 1969: An operational method for objectively forecasting probability of precipitation. ESSA Technical Memorandum WBTM TDL 27, Environmental Science Services Administration, U.S. Department of Commerce, 24 pp.

_____, A. H. Murphy, L. A. Wilson, and J. S. Jensenius, Jr., 1991: Lectures and papers presented at the WMO training workshop on the interpretation of NWP products in terms of local weather phenomena and their verification, Wageningen, The Netherlands, World Meteorological Organization, Geneva, 340 pp.

Hoke, J. E., N. A. Phillips, G. J. DiMego, J. J. Tucillo, and J. G. Sela, 1989: The regional analysis and forecast system of the National Meteorological Center. Wea. Forecasting, 4, 323-334.

Hughes, L. A., 1965: On the probability forecasting of the occurrence of precipitation. Technical Note 20-CR-3, Weather Bureau, Environmental Science Services Administration, U.S. Department of Commerce, p. 13.

Jensenius, J. S., Jr., 1988: Statistical characteristics of the National Meteorological Center's regional and global weather prediction models. Preprints Eighth Conference on Numerical Weather Prediction, Baltimore, Amer. Meteor. Soc., 550-557.

Klein, W. H., 1969: The computer's role in weather forecasting. Weatherwise, 22, 195-218.

_____, 1970: The forecast research program of the Techniques Development Laboratory. Bull. Amer. Meteor. Soc., 51, 133-142.

_____, 1982: Statistical weather forecasting on different time scales. Bull. Amer. Meteor. Soc., 63, 170-177.

_____, 1989: Objective guidance for weather forecasts. Preprints 11th Conference on Probability and Statistics in Atmospheric Sciences and 12th Conference on Weather Analysis and Forecasting, Monterey, Amer. Meteor. Soc., J28-J32.

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

_____, F. Lewis, and G. P. Casely, 1969: Computer forecasts of maximum and minimum surface temperatures. ESSA Technical Memorandum WBTM TDL-26, Environmental Science Services Administration, U.S. Department of Commerce, 27 pp.

_____, and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. Bull. Amer. Meteor. Soc., 55, 1217-1227.

Miller, R. G., 1964: Regression estimation of event probabilities. Technical Report No. 1, Contract Cwb-10704, The Travelers Research Center, Hartford, Conn., 153 pp.

Murphy, A. H., 1985: Probabilistic weather forecasting. Probability, Statistics, and Decision Making in the Atmospheric Sciences (A. H. Murphy and R. W. Katz, Eds.), Westview Press, Boulder, 337-360.

_____, and H. Daan, 1985: Forecast evaluation. Probability, Statistics, and Decision Making in the Atmospheric Sciences (A. H. Murphy and R. W. Katz, Eds.), Westview Press, Boulder, 379-437.

National Weather Service, 1986: Introduction of radiation and a diurnal cycle into the Nested Grid Model. Technical Procedures Bulletin No. 363, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 24 pp.

_____, 1987: AWIPS-90 Operations Concept. National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 65 pp.

Nurmi, P., and J. Kilpinen, 1991: NWP products, their interpretation and verification of local weather forecasts in Finnish Meteorological Institute. In Glahn et al. (1991), pp. XX-13 to XX-18.

Persson, A. O., 1991: Kalman filtering - A new approach to adaptive statistical interpretation of numerical meteorological forecasts. In Glahn et al. (1991), pp. XX-27 to XX-32.

Ross, G. H., 1987: An updatable model output statistics scheme. Extended Abstracts of Papers Presented at the WMO Workshop on Significant Weather Elements Prediction and Objective Interpretation Methods, Toulouse, PSMP Report Series No. 25, World Meteorological Organization, Geneva, 45-48.

Sanders, F., 1958: The evaluation of subjective probability forecasts. Scientific Report No. 5, Massachusetts Institute of Technology, Cambridge, Mass., 60 pp.

_____, 1963: On subjective probability forecasting. J. Appl. Meteor., 2, 191-201.

Sela, J. G., 1988: The new T80 NMC operational spectral model. Preprints Eighth Conference on Numerical Weather Prediction, Baltimore, Amer. Meteor. Soc., 312-313.

Simonsen, C., 1991: Self adaptive model output statistics based on Kalman filtering. In Glahn et al. (1991), pp. XX-33 to XX-37.

Veigas, K. W., 1966: The development of a statistical-physical hurricane prediction model. Final Report, Contract Cwb-10966, The Travelers Research Center, Hartford, Conn. 19 pp.

Weather Bureau, 1968a: Experimental computer forecasts of maximum and minimum temperature. Technical Procedures Bulletin No. 18, Environmental Sciences Services Administration, U.S. Department of Commerce, 9 pp.

_____, 1968b: Operational forecasts with the Sub-synoptic Advection Model (SAM). Technical Procedures Bulletin No. 16, Environmental Sciences Services Administration, U.S. Department of Commerce, 19 pp.

Wilks, D. S., 1990: On the combination of forecast probabilities for consecutive precipitation periods. Wea. Forecasting, 5, 640-650.

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

Yacowar, N. and R. Verret, 1991: Updating weather element forecasts through postprocessing techniques. In Glahn et al. (1991), pp. XX-49 to XX-56.

World Meteorological Organization, 1991: Numerical Weather Prediction Progress Report for 1990. NWPP Report Series No. 17, Technical Document WMO/TD No. 403, Geneva, 328 pp.