3. VERIFICATION METHODS FOR NWP MODEL OUTPUT

Numerical Weather Prediction (NWP) model output forecasts result from computer integrations of the equations governing the physical processes that occur in the atmosphere. NWP output is never a perfect forecast of real atmosphere variables for several reasons: our knowledge of the mathematical formulations of the physics is incomplete; we have to make assumptions and short-cuts in the mathematical formulation of models due to computer-time limitations; our knowledge of initial conditions of the variables is incomplete because we must sample them at a finite number of points in space; there are computer round-off errors in the integrations; and it is widely believed from predictability theory that the maximum period in which detailed predictions of atmospheric flow will have any accuracy is about two weeks. However, impressive results have been achieved, and will continue to be achieved, since the first computer integrations of a simple one-level barotropic model in the early 1950's.

Along with the advances in NWP forecasts has grown a body of techniques for monitoring the skill of those forecasts. In order to assess a model's performance we need measures of its simulation of the basic atmospheric variables (wind, temperature, humidity) and the physical consistency of its simulation of energy conversion and transport processes such as cyclone development and the budgets of heat and momentum. No single "score" is adequate, hence we need several different measures of model performance. Any measure of atmospheric model skill depends on the time and space dimensions used to calculate it, and these depend upon the interests of the user. For example, a forecaster is primarily interested in a model's skill at predicting baroclinic pressure systems for his regional area of responsibility and for projection times generally less than 3 days, while research scientists are interested in model performance for all time and space scales. We will limit ourselves in this document to a review of model performance measures of interest mainly to *operationally-oriented users*.

3.1 DEFINITIONS

Following are some definitions which facilitate the discussion:

3.1.1 Miscellaneous terms:

Prog: an abbreviation of the word *prognosis*. Used by forecasters when referring to a forecast output map. For example, *48-h surface prog* means the forecast surface map valid at 48 hours after model run time.

Error: the difference between forecast and observed variables. The error of variable X at grid point i,j is

$$E_{i,j} = (X^{f}_{i,j} - X^{o}_{i,j})$$

where superscript "f" means a *forecast* variable and "o" means an *observed* variable. There are three main methods used to generate forecasts: by NWP model, by climatology, and by persistence of the observed flow.

Summation: Many of the skill measures involve summation over a set of horizontal grid points. For brevity we will adopt a convention in this section for summation over all grid points of interest:

$$Sum = \sum_{i=1}^{N} \sum_{j=1}^{M} X_{i,j}$$

where N,M are the numbers of grid points over the (horizontal) x and y directions, respectively, for the *user's area of interest*. Since we may be interested in a portion of the entire grid, we specify that $1 \le N \le I$ and $1 \le M \le J$, and I,J are the *total* numbers of grid points in the x and y directions respectively.

Averages: variables which are averaged over some direction will be enclosed in brackets with subscript(s) to indicate which direction the averages are for. For example

$$\langle X \rangle_{x,y,t}$$

means the average of X over x and y in space, and time (t). *Time averages* are just a simple average for data at equal time intervals. However, most data is located at a set of grid points on some map projection in space. For *spatial averages* we must use *weights* which are ratio of the area surrounding a grid point to the area of the entire region of interest on the real surface of the earth. For example, the horizontal average of X is

$$\langle X \rangle_{x,y} = \left(\sum_{i,j} w_{i,j} X_{i,j} \right) / (NM)$$

where $w_{i,j}$ is the ratio of the area surrounding grid point i,j on the map projection to its actual area on the earth, N is the number of grid points over x, and M is the number of grid points over y. On a polar stereographic grid true at any latitude the normalized weight of a grid point is

$$w_{i,j} = \frac{[1 + \sin(\phi_{i,j})]^2}{\sum [1 + \sin(\phi_{i,j})]^2}$$

where $\phi_{i,j}$ is the latitude of grid point i,j. In the above formulation a flat square area tangent to the earth is assumed at each grid point. A different weight factor would apply if the grid distance were not equal in the x and y directions.

3.1.2 Spatial Representation of Variables in NWP Models:

(1) Horizontal:

Grid point: model variables are predicted at a series of points over the area of interest. Grid points are numbered $1 \le i \le I$ for the x direction and $1 \le j \le J$ for the y direction. Generally the *x direction* runs from west to east and the *y direction* from south to north, although in most models the x and y directions are rotated from this orientation by a factor which is a function of latitude.

Spectral: variables are represented as a series expansion where coefficients are multiplied by a known mathematical function, of which the most common functions are *spherical harmonics*. These are functions of latitude and longitude that can be evaluated at any horizontal point. Latitudinal dependence is governed by Legendre polynomials of degree n and order M, while longitudinal dependence is governed by sine/cosine trigonometric functions of order M. An illustration of this method can be found in Burrows (1976).

(2) Vertical:

Coordinate direction perpendicular to horizontal coordinates. Most models use a "sigma" coordinate defined as $r = p/p_o$ where p = pressure, p_o = surface pressure. In most archived data sets, data generated at sigma levels has been interpolated to fixed pressure levels (isobaric coordinates).

3.1.3 Terms Applying to Space:

Zonal: A direction parallel to latitude circles. Positive is east, negative is west.

Meridional: A direction parallel to parallels of longitude. Positive is north, negative is south.

3.1.4 Terms Applying to Time:

Climate: the average of a variable over a long time. The period is defined by the user's purposes and data availability, but it should be of several years duration. Because the period used for averaging can vary, "climate" can be known only imperfectly. This results in a source of error in interpretation of measures that incorporate climate, e.g. the anomaly correlation (Section 3.2.5).

Systematic Error: the time-averaged error:

$$\langle E \rangle_t = \langle X^f_{i,j} - X^o_{i,j} \rangle_t$$

Transient Error: the remainder left when the systematic error of a variable is subtracted from the instantaneous error:

$$E'_{i,j} = E_{i,j} - \langle E_{i,j} \rangle_t$$

3.2 OBJECTIVE MEASURES OF NWP MODEL FORECAST SKILL

All of the following measures can be calculated directly at any set of grid points. Spectral model data can be output at any desired set of grid points in space, although a common practice is to output spectral model data at regularly-spaced longitudes and "Gaussian latitudes" (latitudes of the zeroes of a Legendre polynomial of degree n and order M). Results can be averaged over time and space, and so can be global and/or regional in space, and have systematic and transient components.

3.2.1 Bias, or Mean Error (ME):

This is defined as

$$\langle E \rangle_\chi$$

where subscript χ can be any, some, or all of x,y,σ,t, depending on the user's purposes. A very useful form for forecasters of bias for NWP model outputs is *systematic bias* $\chi = t$. For example, if a model's systematic bias in the surface pressure field is known for a forecast region, a forecaster can correct for it when using the prog. This would be useful for making surface wind forecasts, for example. An example of systematic bias for the Canadian Meteorological Centre's (CMC) operational NWP model is shown in Figure 3.1. In the regions enclosed by solid lines the model's forecast of mean sea level pressure are too high, while the dashed lines enclose regions where the mean sea level pressure forecasts are too low.

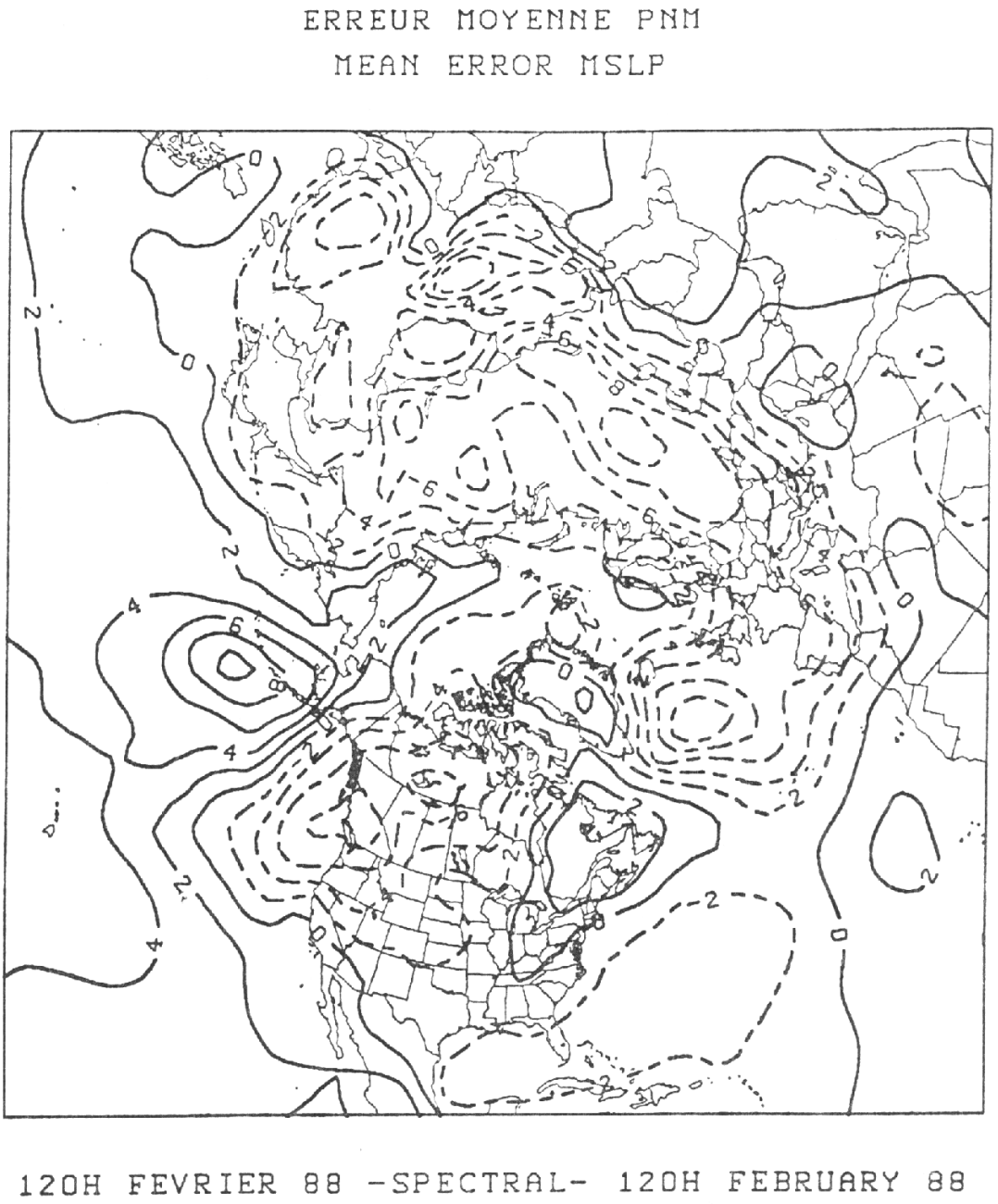


Figure 3.1. An example of systematic bias in an operational NWP model.

3.2.2 Mean Absolute Error (MAE):

This is defined as:

$$\langle |E| \rangle_{\chi}$$

where subscript χ is any, some, or all of x,y,σ,t. Generally speaking, MAE is used in conjunction with ME to give an estimate of the confidence with which we can correct products by their bias. For example, if the MAE of a product

is not too close to the ME then we correct the product by the bias at our peril. If MAE and ME are relatively close, then we can correct by the bias with confidence.

3.2.3 Mean Square Error Verification Measures (MSE, RMSE, RMSE(CMC), RMSGE, Skill Scores):

Mean square error measures are one of the most basic and widely used methods of verification of NWP model output forecasts. *Mean Square Error (MSE)* is defined as follows:

$$\mathbf{MSE} = \langle \langle \mathbf{E}^2 \rangle_{x,y} \rangle_t$$

Root Mean Square Error ***(RMSE) is simply the square root of MSE***, and is a measure of the *amplitude* of the error. MSE and RMSE can be calculated over any or all spatial directions and time. *For example, systematic horizontal RMSE at some vertical level is defined as follows:*

$$\mathbf{RMSE} = \langle \langle \mathbf{E}^2 \rangle_{x,y}^{0.5} \rangle_t$$

Note that both MSE and RMSE are equal to zero only for perfect agreement everywhere between forecasts and verifying observations, otherwise they are greater than zero. Further discussions of the use of RMSE for model verification appears in Section 3.2.5, since RMSE is widely used together with the anomaly correlation. The anomaly correlation is a measure of phase error in NWP forecasts.

A variation of RMSE known as **RMSE(CMC)** ***is used by CMC:***

$$\mathbf{RMSE(CMC)} = \langle [\langle \mathbf{E}^2 \rangle_{x,y} - \langle \mathbf{E} \rangle_{x,y}^2]^{1/2} \rangle_t$$

where the second term is ME^2. The advantage of this formulation is that it is a measure of the variability of the forecast error in a particular region once the bias has been removed, however the price paid is that it can hide the actual RMSE of a prog.

Example: consider 4 points which are consistently underforecast by 5 mb:

X^f_i	X^o_i
1000	1015
990	1005
985	1000
990	1005

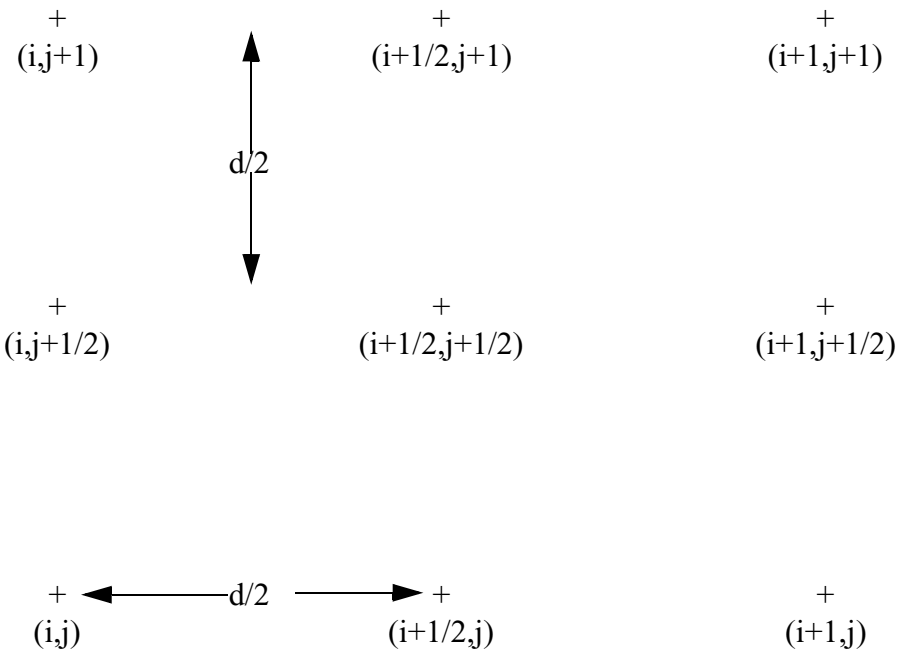
BIAS = [-15-15-15-15]/4 = -15

RMSE = $\{[15^2+15^2+15^2+15^2]/4\}^{1/2}$ = 15

RMSE(CMC) = $\{[15^2+15^2+15^2+15^2]/4 - 15^2]\}^{1/2}$ = 0

The difference between RMSE and RMSE(CMC) is that RMSE(CMC) is the error standard deviation about the *forecast* mean of a field X after it has been corrected by the bias (ME), while RMSE is the error standard deviation about the *actual* mean of a field X.

Another variation of RMSE is the ***Root Mean Square Gradient Error (RMSGE),*** which measures the average magnitude of the error per grid length. Assume we have the following grid points spaced a distance *d* apart:



Where $X^f_{i,j}$ represents the forecast value of X at grid point i,j and $X^o_{i,j}$ represents the observed value of X at grid point i,j, the following calculations are then made:

Forecast Difference in x direction: $G_x^f = X^f_{i+1,j} - X^f_{i,j}$
Forecast Difference in y direction: $G_y^f = X^f_{i,j+1} - X^f_{i,j}$
Observed Difference in x direction: $G_x^o = X^o_{i+1,j} - X^o_{i,j}$
Observed Difference in y direction: $G_y^o = X^o_{i,j+1} - X^o_{i,j}$
Forecast Difference Error in x direction: $= G_x^f - G_x^o$
Forecast Difference Error in y direction: $= G_y^f - G_y^o$

$$\mathbf{RMSGE} = \frac{1}{d}\left\langle\left[\left\langle(G_x^f - G_x^o)^2 + (G_y^f - G_y^o)^2\right\rangle_{x,y}\right]^{1/2}\right\rangle_t$$

Note that G_x is centered at (i+1/2,j) while G_y is centered at (i,j+1/2) in these formulations.

The RMSGE is also used in a **skill score (SS)** ***defined as:***

$$\mathbf{SS} = \frac{\mathbf{RMSGE}}{\mathbf{RMSG}}$$

where RMSG is the Root Mean Square of the *observed* gradient. RMSG is calculated as:

$$\mathbf{RMSG} = \frac{1}{d}\left\langle G_x^{o2} + G_y^{o2}\right\rangle_{x,y}^{1/2}$$

A perfect score would be 0.0. *In general, a score of less than 0.6 is deemed to measure a good forecast while a score*

of greater than 0.6 is deemed to indicate a poor forecast. This form of skill score is similar to the scatter index defined above for wind forecasts. The error is expressed as a fraction of the average magnitude of the parameter being verified

Another skill score can be defined based on MSE to give the fractional improvement in model forecast accuracy with respect to some standard procedure of forecast generation. For example, Murphy and Epstein (1988) define a skill score with respect to climatology:

$$SS(M, C) = 1 - \frac{MSE(M)}{MSE(C)}$$

where on the right-hand side, MSE(M) is the MSE calculated by using model forecasts while MSE(C) is the MSE calculated by using climatology as the forecast. SS(M,C) = 1 when MSE(M) = 0 (perfect model forecast), SS(M,C) = 0 when MSE(M) = MSE(C) (no skill of model forecast over climatology), and SS(M,C) is positive (negative) when the MSE of the model forecast (M) is less (greater) than the MSE of the reference forecast (climatology, C). Of course, SS(M,C) suffers from the usual problem of any scoring method which invokes climatology in its definition, namely that C is imperfectly known. This makes comparison of results involving different climatologies precarious, however comparison of different forecast procedures is possible if a common climatology is used. This skill score is the model verification analogue to the reduction of variance since the denominator is the variance of the forecast field about the climatological (mean) field.

3.2.4 Standard Deviation Error (SDE):

Standard deviation is a basic statistical measure of the degree of variation of data about the mean. If we calculate the standard deviation of the errors in NWP forecasts, the trend of the standard deviation over a period of several years provides a measure of whether models have improved as new physics, numerical techniques, and better computers are implemented. The lower the standard deviation of the forecast error, the more reliable the forecast when compared over a period of several years. For example, Bengtsson (1985) states that European four-day forecasts for the 500 mb geopotential in January 1984 are as accurate as one-day forecasts for the November-April winters of 1951-1954, since both have a standard deviation error of about 75 m. A forecast with low standard deviation error can be relied on not to produce too many "surprises" from day to day.

3.2.5 Anomaly Correlations (AC, zAC):

RMSE is a measure of the *amplitude* of model error. As a companion measure to RMSE, in recent years "correlation measures" have been used as a measure of spatial *phase* difference between two sets of data. For meteorological purposes the two data sets could be forecasts generated by an NWP model and observations from the real atmosphere, or forecasts generated by two NWP models. One of the data sets is usually considered to be a control model (i.e., the "truth"). The correlation is usually between "anomaly" patterns, in which a "climate" value of a variable is subtracted from its instantaneous value at each grid point, thus the correlation between two sources generating the variables has become known as the *anomaly correlation (AC)*. The most common variable used is geopotential height, although any other variable could be used.

There are two variations of AC which have been used by researchers in accordance with their purposes. The most widely known form is used for comparison of operational NWP model forecasts and observations of the real atmosphere when studying *model forecast error*. Here AC is defined as (see, e.g., Murphy and Epstein, 1989):

$$AC = \left\langle \frac{\langle (X^f - \langle X^o \rangle_t)(X^o - \langle X^o \rangle_t) \rangle_{x,y}}{(\langle (X^f - \langle X^o \rangle_t)^2 \rangle_{x,y} \langle (X^o - \langle X^o \rangle_t)^2 \rangle_{x,y})^{1/2}} \right\rangle_t$$

where superscripts "f" and "o" mean model forecast and real atmosphere observation, respectively. To calculate AC we first obtain the real atmosphere climatological value of X at each grid point, $(X^o)_t$. Then we calculate two anomalies at each grid point and time: one is the difference between the NWP forecast of X and the real atmosphere climate,

and the other is the difference between the observed value of X and the real atmosphere climate. AC is then computed in the same manner as a conventional correlation, except that the climate of the real atmosphere is used in the NWP anomaly term, rather than the NWP model climate. In this sense it is not really a true correlation coefficient between the distributions of NWP forecasts and verifying observations.

This definition has often been interpreted as a skill score, but in fact, as Murphy and Epstein point out, it can be decomposed into other measures of skill which indicate that the square of AC can be interpreted as a measure of *potential* rather than skill. However, as with the S1 score (section 3.2.6), long term trends in the AC, and even trends between periods of a couple of weeks duration, can be interpreted as meaningful indicators of a *change* in skill.

As previously mentioned, AC has been used as a comparison measure in *NWP model experiments* involving *two models*. These are called *sensitivity experiments*, where one model integration (the "truth") is compared against another (the experiment) rather than against the observed atmosphere. Daley and Chervin (1985) mention three broad classes of sensitivity experiments. One experiment is comparison of results from *two versions of the same model* where one version differs from the other in some aspect of the model structure, such as resolution or parameterizations of physical processes. A second type of experiment is with differing boundary conditions. Both of these comparison experiments are usually done when changes to an existing model are being considered. A third type of experiment is comparison with some model-generated "truth" of analyses and forecasts produced by integrations of *one model version when initial states differ.* Examples of this are observational system simulation experiments (OSSE), where one of the simulations is initialized with the existing observational network while the other simulation is initialized with a planned or potential observational network. The objective in all three experiments is to show that one model integration is significantly different from another one. For model sensitivity experiments AC takes the form:

$$AC = \left\langle \frac{\langle (X^E - \langle X^E \rangle_t)(X^T - \langle X^T \rangle_t) \rangle_{x,y}}{(\langle (X^E - \langle X^E \rangle_t)^2 \rangle_{x,y} \langle (X^T - \langle X^T \rangle_t)^2 \rangle_{x,y})^{1/2}} \right\rangle_t$$

where the "truth (T) climate" and the "experimental model (E) climate" are different. AC values for each model experiment are then compared. Daley and Chervin give methods for testing the significance of differences in the AC values.

Illustrations of the nature of the AC measure follow. Greatest AC magnitude occurs when the model (experiment) climate anomaly and the "truth" climate anomaly occur in phase or exactly opposite in phase, while small or zero scores occur when the anomalies are out of phase. We can see from a simple example how the AC score is a measure of phase error between the two:

> Suppose that the model-generated and truth anomalies can be represented as sine waves where the truth anomaly lags or leads the model anomaly by some angle λ:
>
> $$X^f - \langle X^o \rangle_t = A \sin\theta \quad \text{and} \quad X^o - \langle X^o \rangle_t = B \sin(\theta + \lambda)$$
>
> where A and B are constants. For simplicity we consider only the x direction, and define all x grid points to lie in the interval 0 to 2p. The anomaly correlation is then

$$AC = \frac{\int_0^{2\pi} A\sin(\theta)\, B\sin(\theta+\lambda)\, d\theta}{\left\{\int_0^{2\pi} A^2\sin^2(\theta)\, d\theta \int_0^{2\pi} B^2\sin^2(\theta+\lambda)\, d\theta\right\}^{1/2}}$$

Using the identity $\sin(\theta+\lambda) = \sin\theta\cos\lambda + \cos\theta\sin\lambda$

we find that AC = 1, $\lambda = 0$ AC = 0, $\lambda = \pm\pi/2$

$(AC = \pm 0.318), \lambda = \pm(\pi/4), \pm 3\pi/4$ $AC = -1, \lambda = -\pi$

Thus AC is bounded by 1 and -1. (This is true in general for AC). When the anomalies are exactly in phase $\lambda = 0$ we have a high value of 1, if the anomalies are exactly opposite in phase $\lambda = -\pi$ we have a low value of -1. The value drops quickly to ± 0.318 if the anomalies are $\pm\pi/4$ or $\mp 3\pi/4$ (± 45 or ∓ 135 degrees) out of phase and to 0 if the anomalies are $\pi/2$ (90 degrees) out of phase.

Studies of model forecast error have shown that an AC value of about .60 is around the lower limit of useful model forecasts (see, e.g. Branstator, 1986). Murphy and Epstein point out that an AC value of .6 translates into a true skill score of about .2, that is, 20% of the way to a perfect forecast.

Since the AC score is bounded by -1 and +1 and the correlation between the two distributions (observed and forecast anomalies) is not zero, the distribution of AC's is not normal. However, the *Fisher z-transformation of AC, known as zAC*, will be unbounded and have a nearly normal distribution. zAC is defined as

$$zAC = (1/2)\ln\frac{(1+AC)}{(1-AC)}$$

The near normality of zAC is desirable since it means that a large part of the distribution of zAC is represented by its mean and standard deviation, and the standard deviation of zAC is independent of its mean value. Branstator (1986) notes that because the latter is true we can use zAC to compare variability of model skill between periods when mean skill is quite different.

RMSE and AC (or zAC) can be used to document progress over several years in a model's skill at prediction. Fig.3.2 shows these scores for the NMC 72-h 500 mb northern hemisphere forecasts for 11 winters from 1974 to 1985. Until 1980 the NMC model employed a grid-point horizontal representation, then a switch to a spectral horizontal representation was done in August 1980. Improvement in the forecasts is characterized by lower RMSE's and higher AC's and zAC's. We see a steady improvement in model skill in the 1970's as model physics was improved and more vertical layers were added, a significant jump in skill in the 1980/81 winter after the switch to spectral occurred, followed by no significant improvement. The RMSE scores actually degraded in 1983/84 and 1984/85. Branstator (1986) attributed this to natural interannual variability. Since the anomaly correlation scores did not change as much as the RMSE's compared to 1981/82, it appears that the model was forecasting the atmospheric wave disturbances with about the same phase error during the period, but amplitudes were not forecast as well after 1981/82.

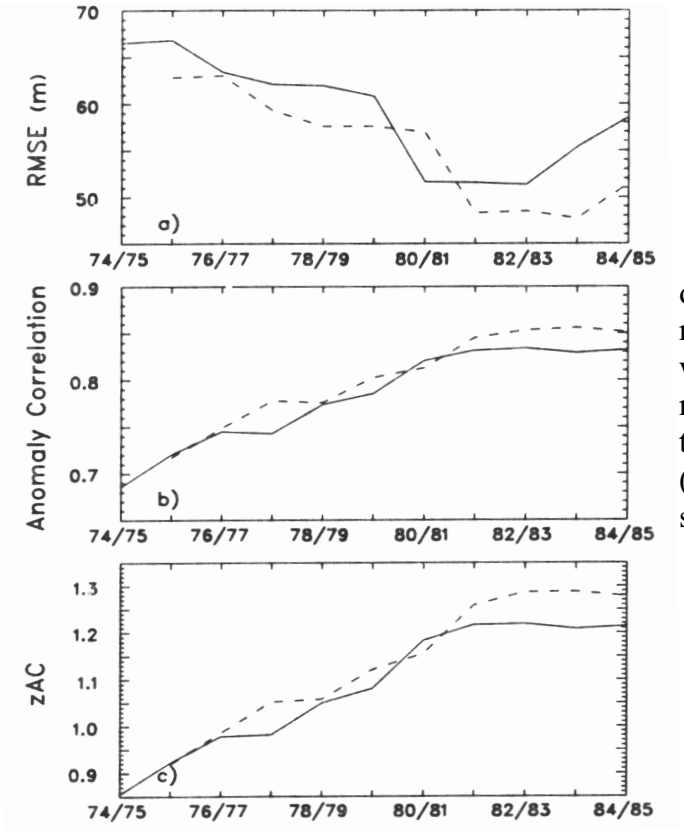


Figure 3.2 (a) Solid line: mean RMSE score attained by 72-h NMC forecasts during 11 winters. Dashed line: level of mean skill that must be attained in each winter for it to be a significant improvement over the previous winter's skill at the 5% level. (b) As in (a) except for AC. (c) As in (a) except for zAC. From Branstator (1986)

Anomaly correlation scores are useful for demonstrating the deterioration of model skill that occurs as model projection time increases as well as changes in model skill over several years. Figure 3.3 shows average AC scores for northern hemisphere 500 mb forecasts as a function of model projection time in days for the European Center for Medium Range Weather Forecasts (ECMWF) model in the winters of 1983-84, and for January 1964-69 runs of an early general circulation model developed by Miyakoda et al (1972). We see that the .60 value of AC is reached by

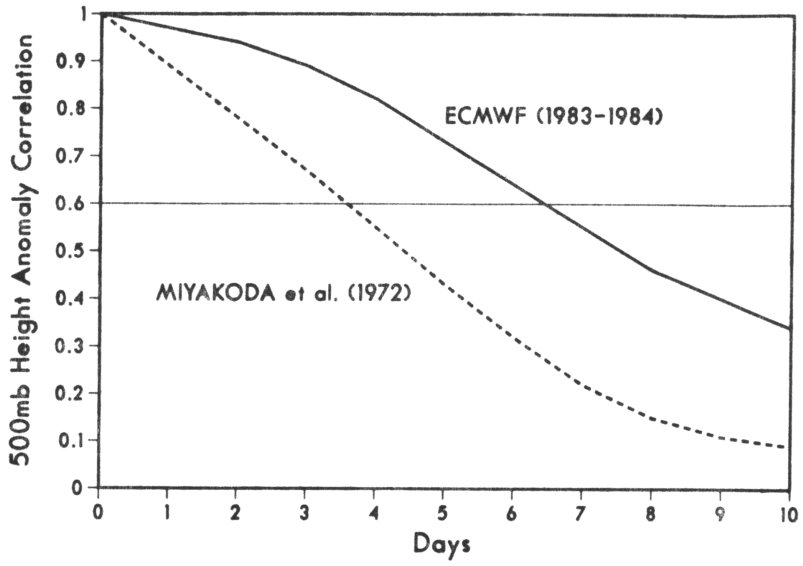


Figure 3.3. Mean 500mb height anomaly correlations as functions of forecast range. The solid line denotes average results from ECMWF operational forecasts for the winters of 1983 and 1984, and the dashed curve shows the mean of twelve forecats from January cases chosen from the years 1964-69. From Bengtsson (1985).

the ECMWF 1983-84 model after about 6.5 days, while this value was reached in only about 3.5 days in the 1964-69 model. This tells us that the projection period in which model forecasts have skill (in a phase-error sense) has doubled over a twenty year period. We can attribute this to periodic implementation of improvements in model resolution and physics as better computers became available and research continued.

Model skill will vary from day to day and from week to week as the characteristics of the large-scale flow vary. Figure 3.4 shows northern hemisphere average zAC scores for each day of November 1983 for the ECMWF model. Note how the scores drop off with model projection time (i.e. D+7 scores are lower than D+3 scores). There are day to day fluctuations of the scores as well as fluctuations on a scale of about 2 weeks. The D+7 AC scores were below the .6 lower limit for the beginning and ending parts of the month, but were above .6 for the middle part of the month. The latter result indicates that model skill will be better for some atmospheric flow regimes than for others. A current topic of interest in forecast research is to determine if model skill can be *predicted* for various large-scale flow regimes, and by so doing, to provide the forecaster with a daily estimate of confidence in the progs.

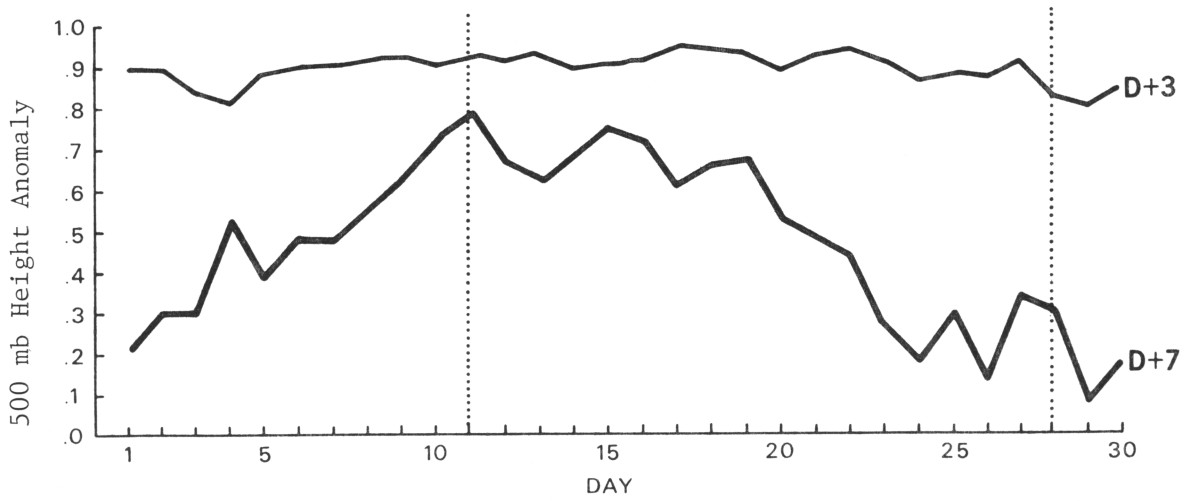


Figure 3.4. Anomaly correlation of 500mb height scores for the Northern Hemisphere of ECMWF three day (D+3) and seven-day (D+7) forecasts for each day of November 1983. From Bengtsson (1985).

3.2.6 S1 SCORE:

This score was proposed by Teweles and Wobus (1954) as a means of verifying the USA's National Meteorological Centre (NMC) model forecasts. It has been used internally ever since at both NMC and CMC, although it is rarely seen in published literature discussing NWP model skill in recent years. The S1 score is a function of the pressure difference (pressure gradient) between pairs of points selected within the main area of interest. The points can be either grid points or observing stations. S1 is defined below:

$$S1 = 100 \frac{\langle |E_g| \rangle_{x,y}}{\langle |G_l| \rangle_{x,y}}$$

where: E_g is the *forecast* pressure difference (or gradient) minus the *observed* pressure difference between pairs of selected stations. Note that the absolute value is to be taken after the pressure gradients have been calculated.

G_l is the observed or forecast pressure difference whichever is larger.

Some of the main characteristics of the S1 score follow:

1) *The full range of the S1 score is from 0 to 200, with a low score being better than a high score.* A perfect score of 0 occurs when the forecast and observed gradients are the same, even though the pressure values may be different. The worst possible score of 200 is attainable when the pressure gradients are exactly reversed. A "bad" score doesn't necessarily mean a bad prog. The S1 score is sensitive to the exact placing of pressure gradients, so that even if the prog had the pressure centers well-placed, if gradients are strong then a lower score may occur if the strengths of the forecast and observed gradients were substantially different. This discourages forecasting cyclone development, and is a draw-back of the score.

2) The S1 score has a seasonal trend because weaker pressure gradients occur in the summer, thus summer S1 scores are larger than winter scores because the denominator G_1 is smaller in summer.

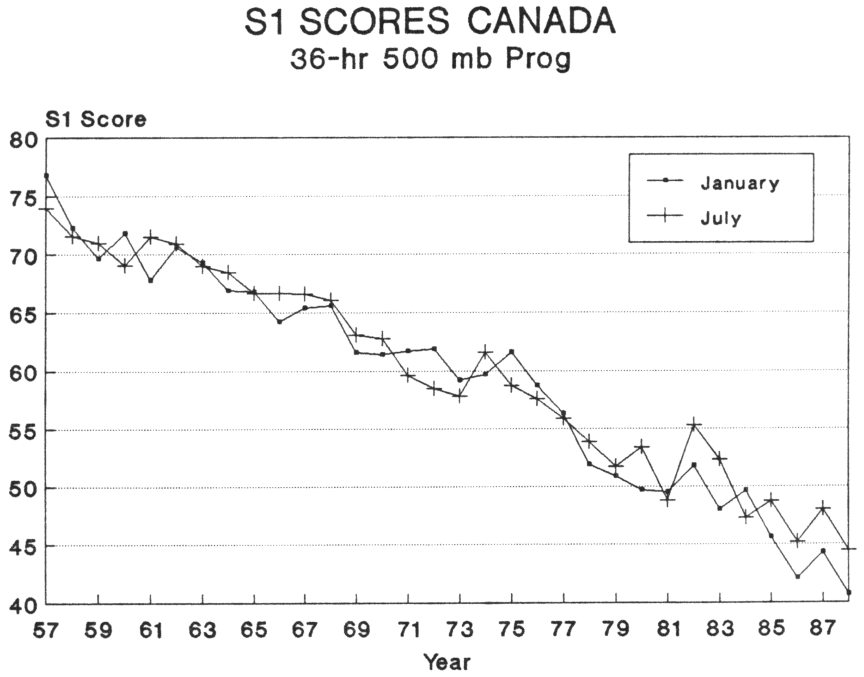


Figure 3.5. S1 scores for CMC forecasts.

3) Lower S1 scores are more easily obtainable for large, intense pressure systems than for weak, poorly defined pressure patterns.

4) The S1 score is more sensitive to forecasts for an individual system when a smaller area is being verified. Lower S1 scores can be achieved by avoiding the placement of high and low centers on grid points.

5) The S1 score suffers from the same problem that many quantitative scores do, namely what constitutes a significant improvement? For example, how much "better" is a score of 75 than a score of 65? A long term trend of lowering scores does indicate, however, steady improvement in the accuracy of forecasts of atmospheric flow.

Figure 3.5 shows S1 scores for CMC 500 mb hemispheric 36-hr forecasts for January and July for 1957 to 1988. Since 1957 there have been many model changes as modelling theory and computers improved, and steady improvement of the forecasts is evident from the lowering scores. Based on S1 scores, we can also say that a 96-hr 500 mb forecast in 1988 (scores not shown) is about as accurate as a 36-hr forecast in 1957, since the scores are of about equal value.

3.3 SUBJECTIVE MEASURES OF NWP MODEL SKILL

The objective measures of NWP model skill discussed above are calculated from data at regular points directly from model output from known formulae and can easily be automated. Another class of measures are *subjective* because they require interpretation of rules which usually depend on the meteorological situation in order to obtain the result. Much can be done by computer, but some interpretation by meteorologists is still required. To date the most useful skill measures to forecasters have been those involving verification of model skill at predicting surface pressure systems and fronts.

3.3.1 Surface Pressure Center Verification

A popular subjective means of measuring characteristics and skill of an NWP model is to document errors associated with its forecasts of *surface pressure systems*. These measures are of great use to forecasters, who must make use of NWP map-format forecasts of weather-producing pressure systems in addition to point values of model variables. Generally the errors are time-averaged to give *systematic* characteristics of model forecasts. Results are subjective to some degree because they require a scheme for segregating different kinds of pressure systems based on their dynamics and/or geographical location and a scheme for matching up observed and forecast pressure centers for verification. Although many aspects of the verification can be done by computer, much of the work in studies to date has been done by knowledgeable humans. Important points to consider and a suggested verification scheme follow below.

We must bear in mind when verification of NWP model output is done for individual cases that the results may reflect both errors in the model formulation and errors in the analysis used to initialize the model, and that it will often be difficult to separate the errors in a quantitative manner. Hopefully, time-averaging the errors will leave model error as the dominant error component. However this may not be true as a general rule for some geographical areas, such as the west coasts of most continents, where the large-scale flow is coming directly off an oceanic region where data is sparse.

To proceed we must decide what features are important when verifying pressure centers. This may vary between users, but generally a forecaster is interested in *speed and position errors for cyclones and anticyclones, and center-pressure errors for cyclones.* A gauge of how the NWP model is handling baroclinic (frontogenetic and heat transport) processes can be found by measuring 500mb-1000mb *thickness errors* over pressure centers. *Another rather crucial point to consider is whether or not the NWP model forecasts a particular pressure center.* For example, we have to associate a particular observed pressure with one on the NWP prog which we hope corresponds to the observed center. For strong centers this is usually not difficult, but it can be difficult for weak or multi-centered weather systems. *We must make decisions about the boundaries of geographical areas in which to verify pressure centers, because as forecasters are well aware, model performance varies substantially between diverse areas such as land versus ocean, or mountainous terrain versus flat terrain.* Including too large an area can hide important information. It is also necessary to stratify the data by monthly or seasonal periods, and by pressure center characteristics.

An early published study of systematic errors in operational USA progs was by Leary (1971). She took into consideration many of the important aspects of pressure center verification just mentioned. A subsequent study by Silberberg and Bosart (1981) updated and expanded on Leary's (1971) work and added some case studies to the discussion. They document some practical rules for establishing a methodology for verification of pressure centers which takes into account the points discussed above. Following in Section 3.3.1 is a proposed pressure center verification methodology which closely parallels Leary (1971) and Silberberg and Bosart (1982). A few added features are included from independent unpublished work by Alexander and Burrows (1981) done for NWP training workshops presented to forecasters at Canadian operational weather centers in 1980/1981. The user can calculate all or some of the measures, depending on his resources and interests. Since several stratifications are possible, it is advisable to work with several years of data.

3.3.2 A Suggested Method for Verification of Pressure-Center Forecasts

1. Inclusion Rules:

1) Verify all *low pressure systems* that have a closed circulation, that is, *at least 1 closed isobar* in either prog or verification analysis. The center location and pressure can be computed with bi-cubic-spline interpolation of the surrounding pressure field or by subjective estimation. Flag cases when the model forecasts a poorly-defined cyclonic center with no closed isobar and check that the computed center position and pressure are reasonable. If not, correct the data with a subjective estimate of the central pressure and location or ignore the center if this is not possible.

2) For *ambiguous low pressure centers* a subjective examination may be advisable. For example, when multi-center low pressure areas are observed but fewer centers are forecast on the prog, match first the prog center and observed low center which have the lowest central pressure, and which appear to be meteorologically consistent with the 500-1000 mb thickness field. Proceed with additional centers if possible.

3) Verify *anticyclones* that have a well defined center (closed or nearly closed isobar) on the verifying analysis.

4) Record data where it occurs for *centers that were forecast but not observed and centers that were observed but not forecast*. When no closed center on a prog corresponds to a closed observed center, but a trough or ridge on the prog does correspond, record the 8x track error (see below for definition) along the direction of the observed center.

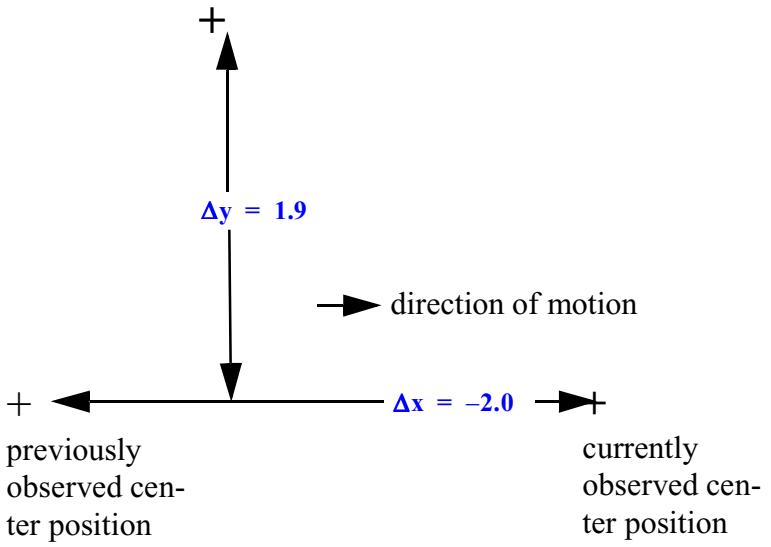
2. Rules for Collecting Data:

For *observed* and *forecast* pressure centers record the following data:

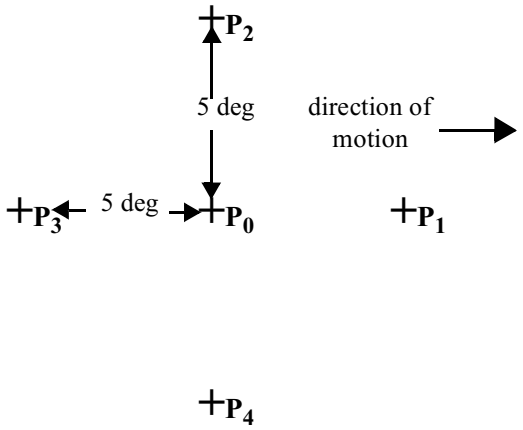
1) The *latitude and longitude* of each pressure center, the *central pressure*, and the *500-1000 mb thickness over the center*. For anticyclones, record a reasonable estimate of the *surface temperature* and/or 500-1000mb thickness near the center.

2) The current longitude and latitude *position errors*. An error of more than 1 degree is deemed to be significant. Resolve the position errors into distance in degrees of latitude and direction to 16 points of the compass. For example, a low pressure center forecast at 45N and 55W that verified at 47N 53W was in error by about 2.3 degrees of latitude to the south-west.

3) Check back 6 (or 12) hours for the previous position of the observed and forecast pressure centers. When a "historical" position can be established for both, calculate the errors in the rate of central pressure change, speed, direction of motion to 16 points of the compass, and *(Δx, Δy) track errors*. The latter are *"along track error"* and *"cross-track error"* to the nearest tenth of a degree of latitude, respectively, and can be calculated as follows:



4) Place a cross grid (+) D degrees of latitude on a side over the observed and forecast cyclone pressure centers with the x axis pointing along the respective directions of motion. A suggested value for D for extratropical systems is 5 degrees of latitude (555.5 km). Add the MSL pressures at each end of the cross and subtract 4 times the central pressure value from the result. This gives a measure of the strength of the *circulation (C)* of the flow around the cyclone. Calculate the *"circulation error"* by subtracting the observed value from the forecast value. This error could also be calculated separately as a gradient error for any of the 4 directions.



$$C^f = P_1^f + P_2^f + P_3^f + P_4^f - 4P_0^f$$

$$C^o = P_1^o + P_2^o + P_3^o + P_4^o - 4P_0^o$$

$$E = C^f - C^o$$

5) Some *comments about the nature of the pressure center being verified* should be recorded along with the information above to facilitate stratification and study of the results. Classifications (1)-(7) below apply to extra-tropical cyclones and are taken from Leary (1971). Her classification system is based on a progression from the youngest to the oldest cyclones and the increasing penetration with age of the sea-level cyclone into the strong 500-1000 mb thickness gradient toward the cold air. The reader is referred to her article for further information. Classifications (8)-(10) apply to "warm core" systems as opposed to the "cold core" extra-tropical cyclones. The suggested classifications for low pressure systems are:

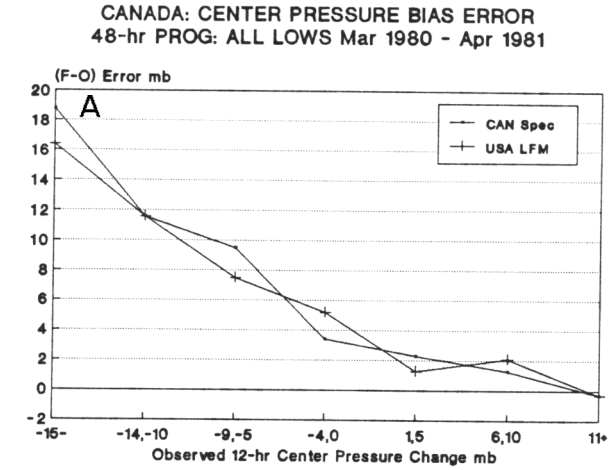
(1) Lee side cyclone	(2) Frontal wave
(3) Wave cyclone	(4) Occluding cyclone
(5) Occluded cyclone	(6) Cold low
(7) Old cold low	(8) Summer heat low
(9) Polar low	(10) Tropical storm

There are obviously many ways to classify and present the data. The user should decide what aspects are important to him and concentrate on those. Examples of what can be done follow. Figure 3.6a which shows the central pressure bias error (average P^f-P^o error) as a function of observed 12-hour pressure change for all lows over Canada and adjacent waters during March 1980 to April 1981, for 48-h progs by the Canadian operational spectral version 8 NWP model. We see that the error increases rapidly with the rate of deepening of a low, and that central pressure is

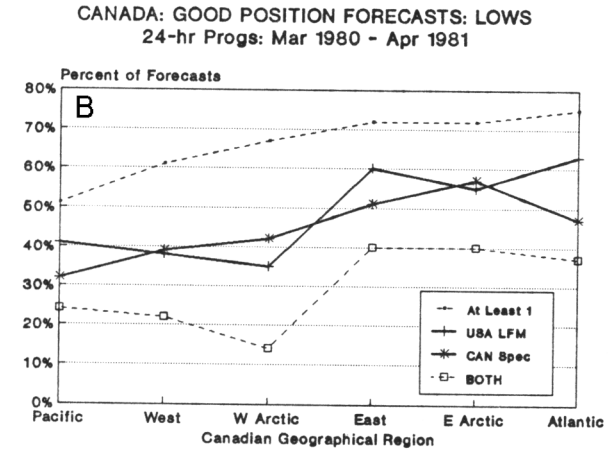
forecast to be too high for all lows except those filling at more than 11 mb in the last 12 hours.

An example of a means of comparing two models is Figure 3.6b. Here we see the percentage of all lows forecast with "good" position errors (2 degrees of latitude or less for both along-track and cross-track position error) for 24 hour progs, plotted by geographic regions for both the Canadian spectral and USA's LFM (Limited Fine Mesh) operational models for March 1980-April 1981. The figure shows that both models made a greater number of "good" position forecasts over eastern and Atlantic regions than over western and Pacific regions, and the LFM out-performed the Canadian spectral model in nearly all regions. Interestingly, there were fewer good position forecasts by both models at the same time than for each individually, whereas if one could always correctly choose one of the model's forecasts, there would be greater success than for either model separately.

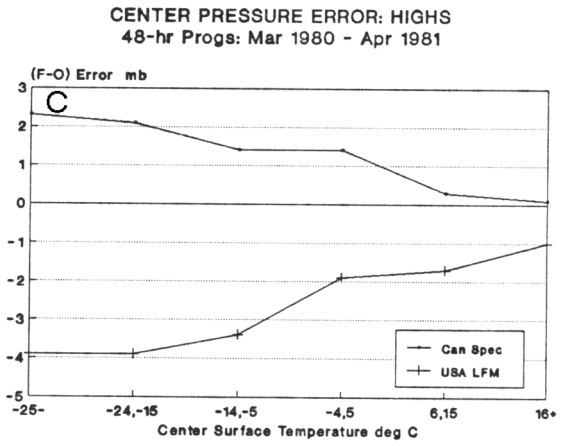
Figure 3.6c shows the mean central pressure error as a function of surface temperature near the center for 48-hr progs of high pressure centers valid at 12Z for USA's LFM and Canada's spectral models for March 1980 to April 1981. Both models are seen to have increasing error as temperature lowers, but the sign of the error is opposite. The LFM model tends to underforecast the central pressure of highs while the spectral model tends to overforecast the central pressure.



(a)



(b)



(c)

Figure 3.6. (a) Central pressure bias error as a function of 12-h pressure change. (b) compares "good" position errors of lows (2 degrees of latitude or less) for the Canadian Spectral and USA's LFM model. (c) Mean central pressure error of high pressure centres for Canadian Spectral and USA's LFM.

3.3.3 Model Dynamics Verification

The verification of model performance in simulation of atmospheric flow and structures is of major concern to modellers. Of interest here is how a model simulates the basic atmospheric parameters (wind, temperature, humidity, precipitation, cloud); how it simulates basic energy-conversion processes such as solar and terrestrial radiation, transports of heat, momentum, and moisture, and dissipation of kinetic energy; and its simulation of the structures of atmospheric disturbances. The means of doing these verifications are many and varied. While many of the verification methods presented here can be and are being used, a complete discussion is beyond the scope of this document.

4. NEW IDEAS IN VERIFICATION

4.1 SIGNAL DETECTION THEORY

Signal Detection Theory (SDT) is a verification procedure brought into meteorology by Mason (1982). It has been more widely applied in medicine and other sciences. Examples of problems for which it has been used include assessment of the ability to diagnose breast cancer from X-rays (Swets and Pickett, 1982), the ability to detect deception by means of polygraph tests (Szucko and Kleinmuntz, 1981), and the comparative evaluation of two methods for forecasting frost (Mason, 1980). In all of these cases, the aim was to assess the ability of the diagnostic system (i.e. the X-ray or the polygraph) or the forecast method to clearly discriminate between two alternative outcomes, for example, rain or no rain, or temperatures above or below an important threshold. SDT is therefore most applicable to two-state categorical weather elements, although multiple category elements could be verified as a sequence of two-category elements.

		Forecast		
		YES	NO	
OBSERVED	YES	X	Y	X+Y
	NO	Z	W	Z+W
		X+Z	Y+W	Total

Figure 4.1. Contingency table.

Consider a two category contingency table for rain occurrence as shown in figure 4.1. The four entries of the table can be referred to as "hits" (correct forecasts of rain), "correct rejections" (correct forecasts of no rain), "misses" (forecasts of no rain when rain occurred), and "false alarms" (forecasts of rain when no rain occurred). The SDT model makes use principally of two functions of these four entries: the hit rate and the false alarm rate. The hit rate is simply X/(X+Y) and is identically the probability of detection or the prefigurance. The hit rate can also be referred to as the percent of correct forecasts of rain given that rain was observed. The false alarm rate is Z/(Z+W), which is **NOT** the same as the false alarm ratio or the post-agreement described in section 2.6.2. Here the false alarm rate is the percent of forecasts of the event given that the event did not occur. These two measures imply a data stratification on the basis of the observation, and thus SDT can be included in the class of verification measures that require stratification by observation.

If, for a given contingency table, these two measures are plotted against each other on a graph, a single point results. It is most desirable that the hit rate be high and the false alarm rate be low. On the graph, the closer the point is to the upper left hand corner, the better the forecast.

SDT is in fact a generalization of these ideas to verification of probability forecasts. Suppose a verification dataset is stratified as for a reliability table into 10% wide categories, and occurrences and non-occurrences are tabulated for each category. Table 4.1 is an example. Suppose further that 30% is chosen as a threshold for forecasting

precipitation; precipitation is forecast if the probability is over 30%. Given that threshold, the entries of the table can be summed to produce the four entries of a two-by-two contingency table, the hit and false alarm rate calculated and a point plotted on a graph. If this process is repeated for a set of tables generated using 0, 10%, ...100% as thresholds, the result is a set of points on the graph that will usually form a smooth curve such as the one shown in figure 4.2. This curve is called the relative operating characteristic (ROC). Since a perfect forecast means all correct forecasts and no false alarms regardless of the threshold chosen, a perfect forecast is represented by a curve that lies along the left-hand side of the graph to the upper left corner, from there along the upper side of the graph. A convenient relative index associated with the ROC is the area under the curve, which decreases from 1 toward 0 as the curve moves away from the left and top sides of the box. A useless forecast is represented in this system by an area of 0.5, for a curve that lies along the diagonal. This is produced by a forecast system that incurs false alarms at the same rate as hits. Such a system cannot discriminate between occurrences and non-occurrences of the event. Perverse forecasts can be envisioned where the line lies below the 45 degree line, but fortunately we have never seen a forecast verify this way.

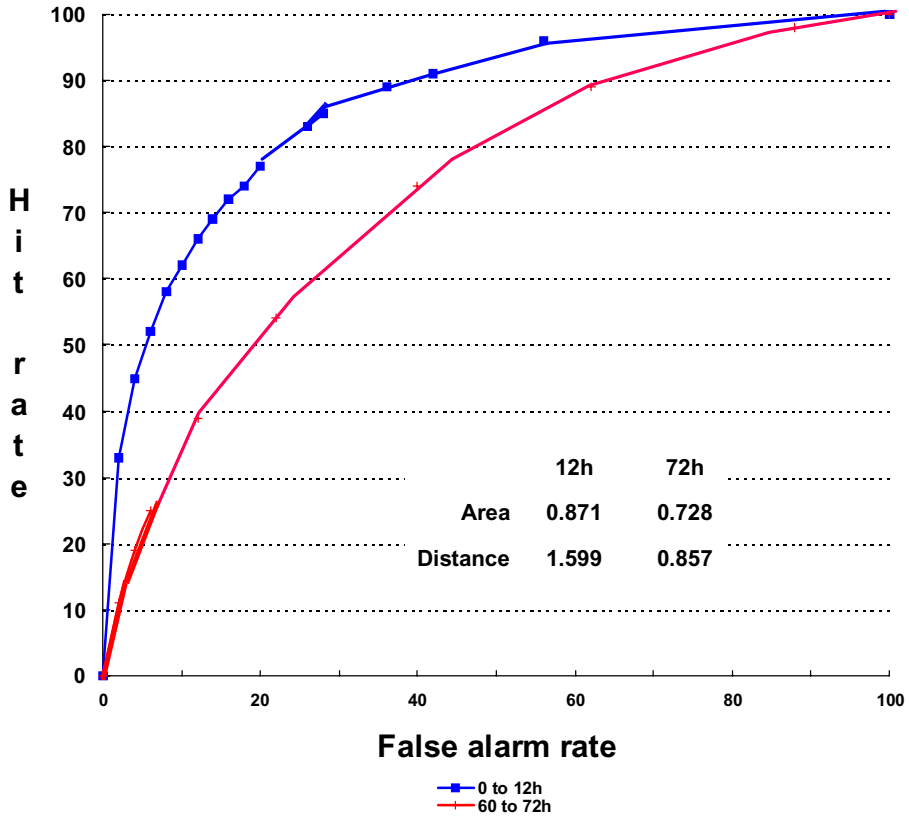


Figure 4.2. Graphing the relative operating characteristic curve (ROC). Data for CMC 12h POP, May 1987 to April 1988, 7250 cases.

Table 4.1: Precipitation forecast distributions stratified according to observation.

Probability range	# non-occurrences	# occurrences
0-9%	613	43
10-19%	1389	172
20-29%	1183	283
30-39%	936	350
40-49%	602	323
50-59%	327	287
60-69%	151	169
70-79%	88	163
80-89%	40	89
90-100%	22	41
Total	5331	1920

EXAMPLE: If 30% is chosen as the threshold for forecasting rain, the entries of Table 4.1 can be summed to produce the entries of a two-by-two contingency table. To help illustrate this, a horizontal line has been drawn at the 30% threshold. X is the sum of the occurrences column below the line; Y is the sum of the occurrences column above the line; Z is the sum of the non-occurrences column below the line; and W is the sum of the non-occurrences column above the line. X+Y is the sum of the whole right-hand column, given at the bottom of the table, and Z+W is the sum of the whole left-hand column. For the 30% threshold, the hit rate is X/(X+Y)= 0.741 and the false alarm rate is Z/(Z+W)= 0.406. These two values when plotted on the ROC graph give a point on the (lower) curve of figure 4.2. Other points on the curve are generated by "moving the line" on the table to different thresholds, and recalculating the hit rate and false alarm rate for these thresholds. SDT thus seeks to give information about a set of probability forecasts as they may be used in decision-making.

There is one other measure of importance in the SDT model. Consider figure 4.3, which represents the conditional distributions of forecast probabilities given the occurrence and non-occurrence of fog. The farther apart these two distributions, the greater the power of the forecast to discriminate occurrences from non-occurrences. One measure of the separation distance is the separation of the means of the two distributions. It is also evident that the discriminating power is weakened if the distributions have large dispersions, which increase the overlap for a given separation of means. Thus, the distance measure is normalized to the standard deviation of the distribution for non-occurrences (usually). However, the distance measure is deficient in that it implicitly assumes that the two subsamples are of equal size.

Probably the greatest advantage (or disadvantage) of SDT is that it can be used with non-numerical probabilistic forecasts as well as for categorical and numerical probability forecasts. The attributes that it measures are common to all these types of forecasts, and they can be directly compared. For example, it is possible, perhaps even desirable to verify numerical POP forecasts and worded precipitation forecasts together using this method. One can then make statements about the utility of the POP forecasts vs the worded forecasts for decision-making.

The other main advantage of SDT for forecast verification is that it is independent of the calibration of the forecast probability; reliability is NOT considered in any way. Recall that reliability tables, which imply stratification on the basis of forecast probability, address questions of reliability of probability forecasts. For SDT, if a threshold of 30% succeeds in separating the events into occurrences and non-occurrences perfectly, so be it. The number or word attached to the forecast is immaterial. This point is illustrated by figure 4.5, for two different precipitation forecast techniques. The curves say that the MOS forecast is slightly better than the perfect prog forecast. The calibration is indicated by the relative positions of the points along the two lines. A 55% MOS forecast threshold achieves about the same hit and false alarm rate as a 85% perfect prog threshold, but this doesn't affect the position of the lines relative to each other.

EXAMPLE: Interpretation of the ROC graph. Two ROC curves are shown on figure 4.2, one for a 0 to 12 hour forecast of POP, and the other for a 60 to 72 hour POP forecast. Both are from the operational POP forecast system at CMC(Canadian Meteorological Centre), and the verification sample is one year of forecasts ending in April, 1988. The curve for 60 to 72 hours is clearly lower than the other one, showing that the longer range forecasts form a poorer basis for decision. The approximate areas under the curves are given on the graph. 0.871 is quite a high value in our experience, and represents a very good forecast. Even the 72h area is well above the 0 skill value of 0.5. The "distance" values are measures of the separation of the means of the distributions of forecasts preceding occurrences and non-occurrences. To calculate these distances, normal distributions have been assumed, and the distances are expressed in terms of the standard deviation of the forecast probabilities preceding non-occurrences. A glance at figure 4.4 will help clarify the concept of the distance between the two distributions. Although the distributions are skewed, it is possible to see that the means will have greater separation for the 0 to 12h forecasts than for the 60 to 72h forecasts. The area under the ROC and the separation of distribution means are two related but different measures of the discriminating power of the forecast technique. The separation of means is expressed in terms of the standard deviation of one of the distributions because the dispersion of the two distributions affects the ability to discriminate as well: The greater the dispersion, the poorer the discrimination for a given separation of means. On the graphs, perfect discrimination is represented by a single solid bar on the left and a single hatched bar on the right. The 12h forecast begins to approach this ideal. No skill is represented by the two distributions lying on top of each other, with identical means.

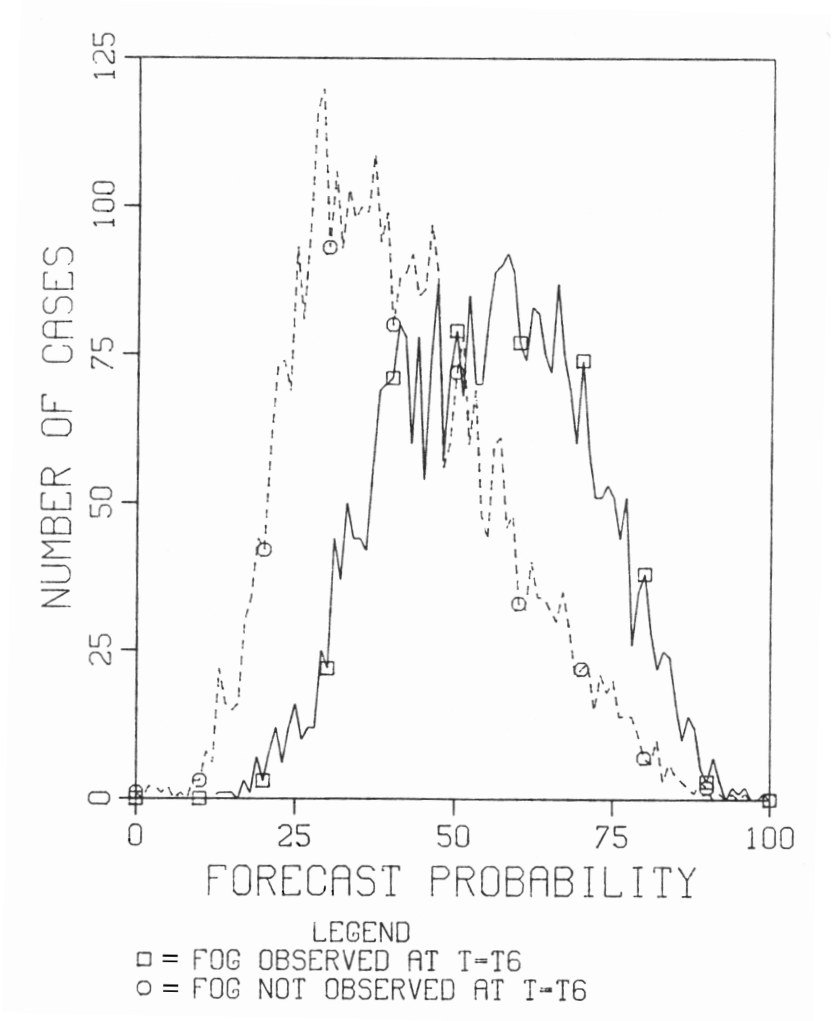


Figure 4.3. Conditional distribution of forecast probabilities given the occurrence and non-occurrence of fog.

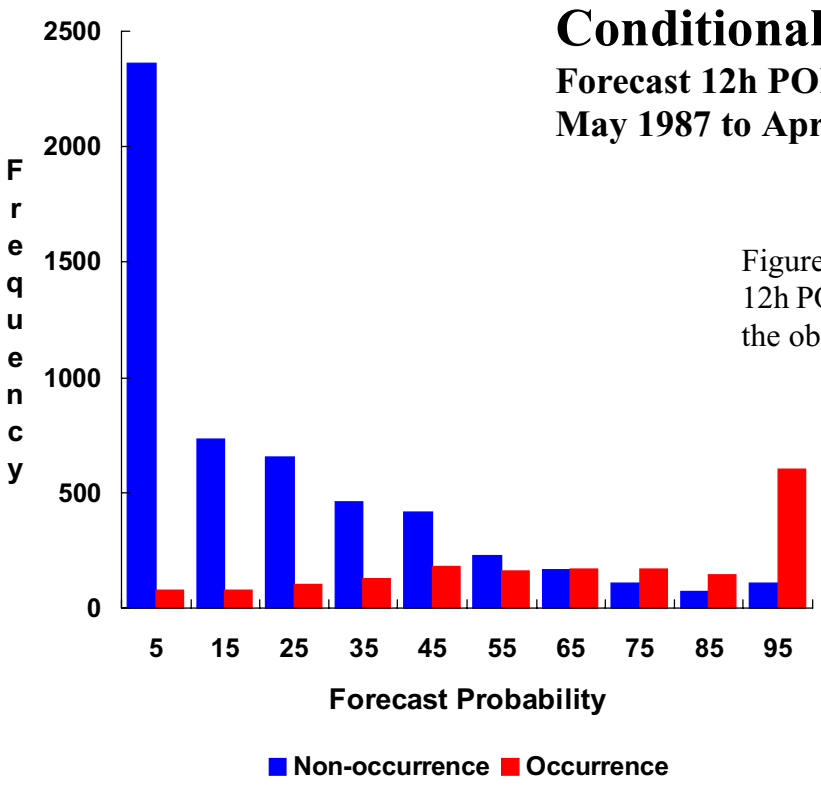


Figure 4.4 a) Histogram of the 0-12h POP forecasts conditioned on the observation.

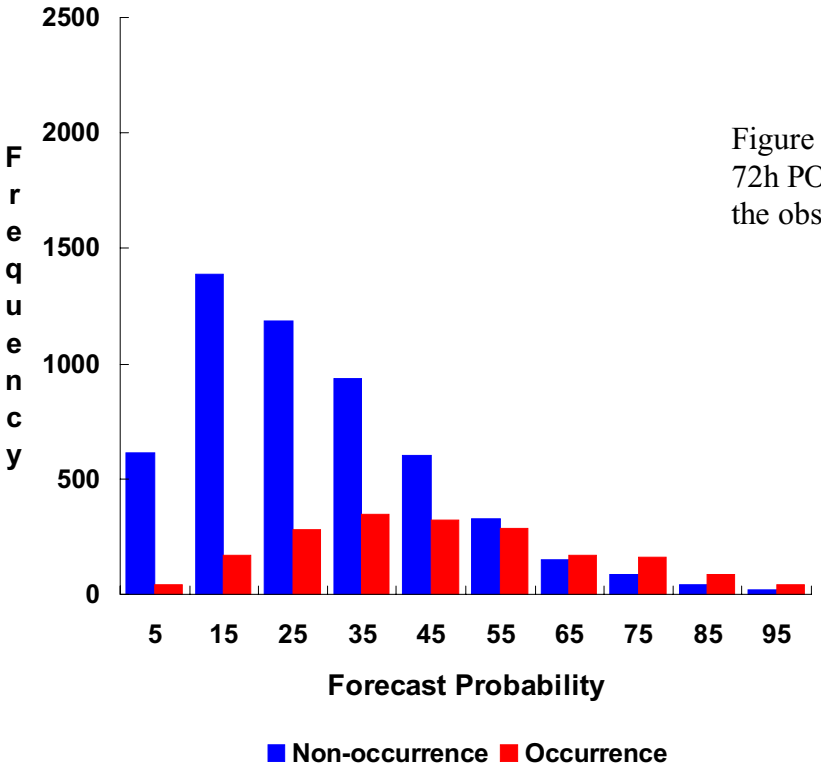


Figure 4.4 b) Histogram of the 60-72h POP forecasts conditioned on the observation.

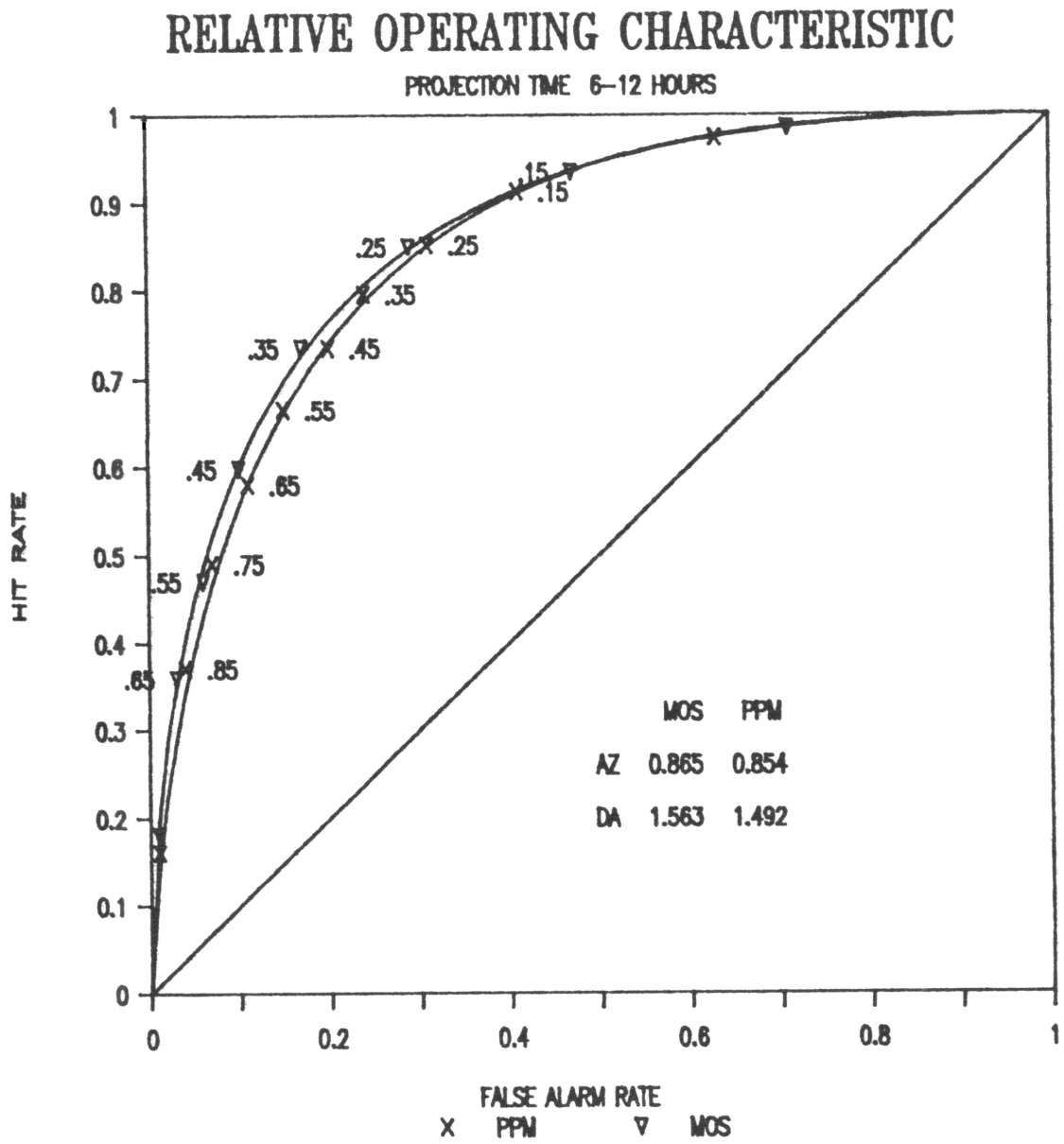


Figure 4.5. ROC curves for two different precipitation (POP) forecast techniques. "AZ" is the area under the ROC curve and "DA" is the distance between the means of the distributions of forecast probability preceeding the event and the non-event. Sample size 7400 events.

SDT is in the "bandwagon" stage right now. There is a tendency to overstate its utility in verification, and to expect it to satisfy more verification needs than it can. It is used a great deal to verify the various PROFS(Program for Regional Observing and Forecasting Services) forecasting experiments for severe weather. While it is an important new tool in verification, it is necessary to keep its limitations in perspective: SDT is based on stratification by observation and therefore can say nothing about reliability, and does not deal with missed events except indirectly. SDT strength lies in its ability to describe the cost of increased false alarms when thresholds are relaxed for severe weather forecasting, and also its ability to permit verification of numerical and worded probability and categorical forecasts in one system.

4.2 STATISTICALLY FORECASTING THE ERROR IN NWP MODELS

NWP model skill will vary with time for three main reasons: the quality of the initial analysis, baroclinic and/or barotropic instability of the large scale flow, and model systematic errors. Recently there has been research on the extent to which NWP model skill can be predicted. The motivation for this is simple: along with NWP forecasts can we provide a measure of their expected skill? Such an estimate would be of great use to a user in terms of attaching credibility to a particular forecast.

Palmer and Tibaldi (1988) studied the problem of predicting the skill of medium range forecasts out to 10 days, using ECMWF model output and statistical methods. Four sets of potential predictors were tried. The first set was a measure of consistency between adjacent forecasts: the "spread", as measured by computing the RMS difference and anomaly correlation coefficients (zAC) between matched sets of yesterday's day n+1 and today's day n forecasts. The second predictor set used descriptors of the large scale flow as predictors: for the forecast 500 mb heights, a regression analysis is done between their RMS error and matching EOF (Empirical Orthogonal Functions) coefficients. The third predictor set was a "proxy" measure of initial analysis errors, where the RMS error of yesterday's day 1 forecast with today's observations was considered to be a measure of today's initial analysis error. This was then correlated with all of today's day n+1 forecasts to give an estimate of the growth of initial analysis errors. The fourth predictor set was RMS difference between the initial 500 mb height and the 500 mb height forecasts, and is a measure of persistence. It can be regarded as a proxy measure for the degree of instability of the basic atmospheric flow.

While there were some areas of success, overall the results of Palmer and Tibaldi's studies were disappointing when the prediction methods were tested on one winter of independent medium range forecast data. They concluded that some aspects of the low frequency component of forecast skill variability can be satisfactorily predicted though, high frequency variability remains unpredicted. They seemed to achieve the best results with the second and fourth predictor sets. In another study, Chen (1989) also presents evidence that the persistence of the latest model integration is significantly correlated with the skill of medium range forecasts.

Palmer and Tibaldi found that model systematic error and flow barotropic instability are important factors for variability in medium range forecasts, while baroclinic instability (growth of cyclones) is likely the dominant mechanism for variability of *short range* forecasts. Work is currently underway by W. R. Burrows to study the problem of predicting by statistical methods the variability of short range forecasts over Canada using a variety of predictors and skill measures, some involving the "baroclinic" measures of the skill of model prediction of cyclones described here in Section 3.4.

4.3 INTERRELATIONSHIPS BETWEEN OBJECTIVE AND SUBJECTIVE GUIDANCE

There is considerable debate and controversy in the meteorological community concerning the respective contributions to weather forecasts by "man" (i.e. forecasters) and "machine" (i.e. numerical and/or statistical models, and "expert" systems). Especially sensitive are situations where objective (numerical/statistical) forecasts are provided as guidance for the preparation of the corresponding subjective forecast. The traditional verification method is to evaluate the individual contributions of forecasters and models by comparing the value of overall measures of performance(i.e. such as mean absolute error, skill scores, etc.). Typical results show that there is relatively little difference between the scores for objective and subjective forecasts, especially for the longer lead times. As a consequence, the deduction was made that subjective forecasts contain very little information that was not already contained in the objective forecasts.

Murphy and Winkler (1987) in proposing a general framework for forecast verification have changed the focus from which forecast performs best, to investigating the interaction of the two forecasts, as two complimentary sources of information. Murphy, Chen and Clemen (1988) state that "since the purpose of providing the forecaster with guidance is presumably to enhance the quality of the official subjective forecasts (as opposed to developing a rationale for replacing forecasters with objective models), an approach based on relative performance appears to be inappropriate."

A number of papers by Clemen and Murphy (1986), Murphy, Chen and Brown (1987) and Murphy, Chen and Clemen (1988) have applied the concept of forecast verification based on joint distributions of forecasts and observations. The conclusions following these studies were: (1) subjective forecasts contain information not included in the objective forecasts and (2) subjective forecasts do not make full use of the information contained in the objective forecasts.

The above papers are exciting to read. The simplicity of the verification approach is reflected by commenting to oneself that the method seems so obvious, but, "Why didn't I do it myself?" The papers also shift our focus from verification measures, to more basic measures of performance (Murphy and Winkler, 1987). As stated in previous chapters, summary verification measures are quite useful when the primary objective is to compare forecast procedures in some overall sense. Summary measures are not helpful when the object is to understand the strengths and weaknesses of the forecast, or to improve the forecast performance and accuracy.

5. VERIFICATION, CANADIAN EXPERIENCE

The (Canadian) National Aviation Terminal Forecast (FT) Verification program is an automated minute-by-minute verification system of the terminal forecast with climatology and persistence as the standards for comparison. This program is an excellent example of strengths and weaknesses generally found in verification schemes for administrative purposes. The essential features of this verification scheme are discussed here as an example.

A terminal forecast is categorized into a number of mutually exclusive, operationally significant ceiling/visibility ranges. Table 5.1 lists the categories that were selected as the National standard.

Table 5.1: Operationally significant ceiling and visibility ranges used at Winnipeg International Airport and used as the National FT verification standard.

Ceiling and Visibility Combinations					
Ranges	CIGS (ft) with	VSBYS (mi)		VSBYS (mi) with	CIGS (ft)
1	0-100	0 or more	or	0-3/8	0 or more
2	200	1/2 or more	or	1/2-5/8	200 or more
3	300-400	3/4 or more	or	3/4	300 or more
4	500-900	1 or more	or	1-1 1/2	500 or more
5	1000-2400	3 or more	or	3 or more	1000-2400
6	2500 or more	3 or more	or	3 or more	2500 or more

The ceiling/visibility standard can be traced to the verification effort of the 1950's. Ceiling and visibility categories were used in a 6x6 contingency table to verify the major terminals in each administration region. Monthly statistics for the regional office as a whole were sent to headquarters for analysis. During the 1970's the reports were no longer sent to headquarters, and eventually the regional verification effort had taken other priorities. In the early 70's I can remember receiving a personalized verification summary of my forecasts for Goose Bay and Frobisher Bay (now called Iqaluit), plus a little eyeball to eyeball talk with the Chief Meteorologist. Some credit these scores with my move from operations to research, but my interpretation of what they mean is not clear (Note from Henry Stanski).

5.1 Probability Interpretation Conversion

Each part-period of the terminal forecast is interpreted separately in terms of a probability for each of the six ceiling/visibility ranges. The probability P_i for each category is expressed in decimal form, such that:

$$\sum_{i=1}^{6} P_i = 1 \quad \text{and each } P_i \geq 0$$

where a probability of zero implies the event is not expected to occur, and one if the event is certain to occur. For a probability greater than zero but less than one, the probability of the event is indicated, e.g. if P_i is 25%, there is one chance in four that the event will occur.

A forecast involving a single category is assigned a probability of 1.0 in that category, and 0.0 for the other 5 categories. The probability assigned in multi-category forecasts depends upon the interpretation of the descriptive terms, variable (VRBL), occasional (OCNL) and risk (RSK). (A drawback in the National FT verification system is that interpretation of the terms VRBL and OCNL is not strictly correct. The assumption is to interpret the descriptive terms as a probability of occurrence and not as a percentage time of occurrence). The interpretations in Table 5.2

were reached by "consensus".

Table 5.2: Probability interpretation of descriptive forecast terms in the National FT Verification system. (From Reid, 1978)

Forecast	Probability Interpretation	Ratio of occurrence (A:B)
A vrbl B	A occurs as often as B	1:1
A ocnl B	A occurs 75% of the time and B 25%	3:1
A rsk B	A occurs 90% of the time and B 10%	9:1

Table 5.3: Examples of converting descriptive forecast terms into probability forecasts.

Terminal Forecast	Range	Probability Ranges					
		1	2	3	4	5	6
20 SCT C100 OVC	6	0	0	0	0	0	1.0
20 SCT C100 OVC VRBL C20 BKN 100 OVC	6V5	0	0	0	0	0.5	0.5
20 SCT C100 OVC OCNL C20 BKN 100 OVC	6O5	0	0	0	0	0.25	0.75
C2 X 1/2F RSK C0 X 0F	2R1	0.1	0.9	0	0	0	0
CLR VRBL C0 X 0F	6V1	0.5	0	0	0	0	0.5
C6 OVC 1R-F OCNL	4O3R2						
C4 OVC 3/4R-F RSK	(Reid)	0	0.1	0.3	0.6	0	0
C2 OVC 1/2RF	(Math)	0	0.08	0.23	0.69	0	0
where O = OCNL, V = VRBL, and R = RSK							

Table 5.4 shows the RPS obtained from a single forecast as a function of the category of occurrence. It illustrates how the score takes into account the proximity of the forecast and observed categories of the event Table 5.4 demonstrates several characteristics of the RPS:

1. The only way to get a perfect score is to forecast categorically for the category that occurs; any distribution of probabilities over more than one category reduces the maximum possible score.
2. Concentrating the distribution of forecast probabilities around the category that occurs gives a better score than spreading them out over all categories, even if the observed category corresponds to the mode of the probability distribution. (cf. examples 2,3,4 in the table)
3. Forecasting equal probability between two non-adjacent categories results in equal scores for the occurrence of either of the categories or the category in between, even if it was assigned zero probability of occurrence (cf. example 5 in the table).

Like the Brier Score, the RPS is often used for comparison of forecasts from different sources. The next set of figures illustrate the use of the score for comparison with unskilled forecasts of persistence and climatology, and also shows variations of the score with different climatologies of the verification sample. All the following examples are taken from output of the National FT verification program. Scores were computed on a monthly and sometimes annual basis.

.

Table 5.4: Probability interpretation of various descriptive forecasts and the RPS according to observed category.

		RPS according to Observed Weather Category					
Forecast Category	Probability (P_1,P_2,..,P_6)	1	2	3	4	5	6
6	(0,0,0,0,0,1)	.000	.200	.400	.600	.800	1.000
3	(0,0,1,0,0,0)	.600	.800	1.000	.800	.600	.400
3R2	(0,.1,.9,0,0,0)	.638	.838	.998	.798	.598	.398
2R1	(.1,0,.9,0,0,0)	.676	.836	.996	.792	.596	.396
3O2	(0,.25,.75,0,0,0)	.688	.888	.988	.788	.588	.388
3O1	(.25,0,.75,0,0,0)	.775	.875	.975	.775	.575	.375
3O2R1	(.08,.23,.69,0,0,0)	.736	.904	.980	.780	.580	.380
5O3R1	(.08,0,.23,0,.69,0)	.471	.639	.807	.883	.959	.759
4V3	(0,0,.5,.5,0,0)	.550	.750	.950	.950	.750	.500
3V1	(.5,0,.5,0,0,0)	.900	.900	.900	.700	.500	.300
3V2R1	(.06,.47,.47,0,0,0)	.779	.955	.943	.745	.543	.343
5V3R1	(.06,0,.47,0,.47,0)	.558	.734	.910	.898	.886	.686

Table 5.5: Monthly averaged RPS for the first 12 hours at several Canadian terminals, with the number of forecasts verified indicated. (F)orecast, (C)limatology, (P)ersistence.

Station	Month/Year	RPS(F)	RPS(C)	RPS(P)	Events
Toronto	6/88	.998	.993	.998	112
Winnipeg	6/88	.992	.991	.987	109
Toronto	2/83	.912	.921	.876	35

Table 5.5 shows the RPS for three stations averaged over a month. Scores for persistence and climatology forecasts are also given. The scores for June,1988 are very high. This is attributable to the fact that Winnipeg and Toronto were in the midst of a widespread drought, with little change in the weather over the month. At Toronto, persistence tied the forecast, while climatology suffered a small reduction as a penalty for not making categorical forecasts. Scores are similar at Winnipeg, but the lower score for persistence shows "more" variability in the weather. Forecasters in Toronto and Winnipeg had little opportunity to demonstrate their abilities during a period when weather conditions were static. The scores for the February data, on the other hand, are lower reflecting more change in the weather. Comparing the sets of scores indicates that differences due to changes in climatology or persistence may be as great or greater than differences between forecast and "standard" scores. This means it is difficult to draw conclusions on forecast accuracy from scores produced on different datasets, especially if the samples are small and therefore subject to greater between sample variance.

Figures 5.1 to 5.4 illustrate the influence the verification period has on the summary interpretation. Figure 5.1 is a plot of the RPS values for Toronto (YYZ) averaged daily for the month. This trace suggests a seasonal trend, but the large oscillations which modulate the seasonal trend indicate a lack of data and not real changes in forecast skill.

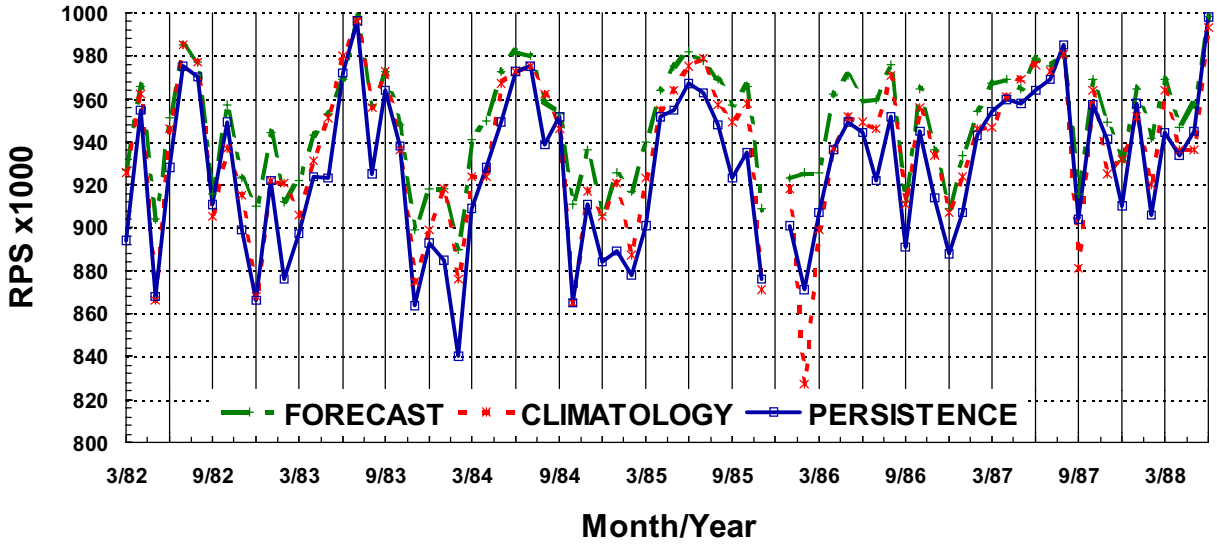


Figure 5.1. Average RPS values for Toronto, daily averages for the month.

Figures 5.2 to 5.4 are plots of seasonal RPS values for Toronto (YYZ), Winnipeg (YWG) and Gander (YQX); the y-axis dimensions were held constant to aid comparison. The seasonal RPS values are the weighted averages for 3 month periods starting with January (labelled season 1). A seasonal trend is stronger at Toronto and Winnipeg, while at Gander the seasonal trend is weak. In general the separation among the curves seems to be influenced by:

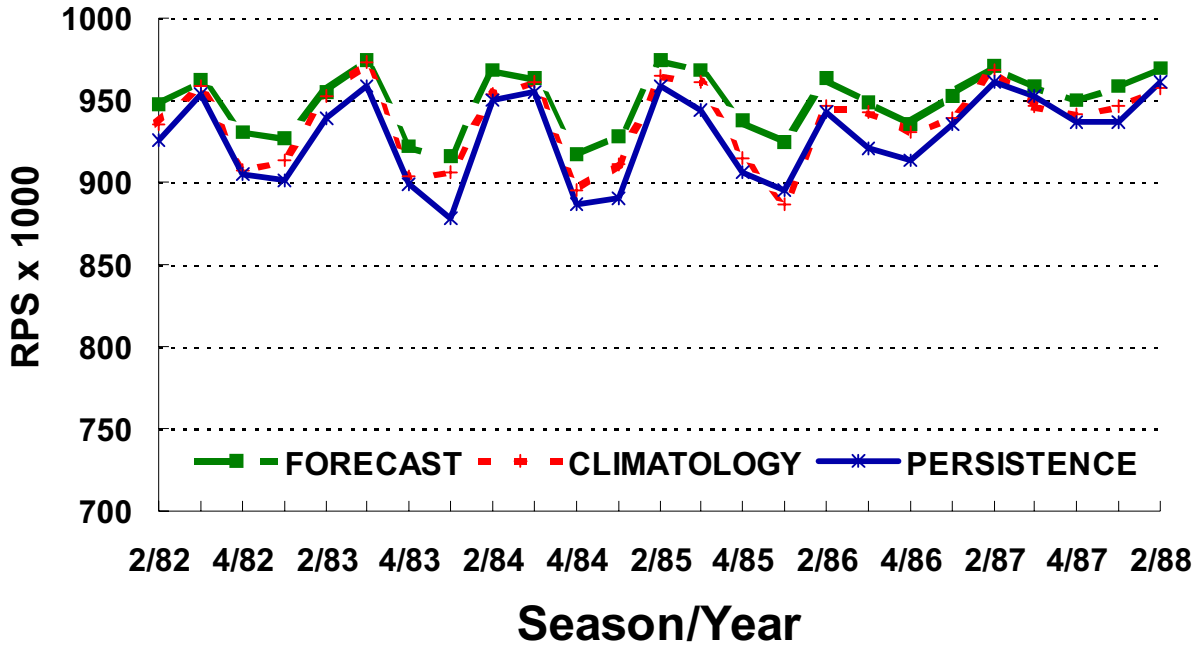
a) the season: The persistence in the weather results in higher RPS values and narrowing of the spread among the forecast, climatology and persistence curves. Seasons characterized with highly variable weather produce lower RPS values and the curves display greater separation.

b) the location: Also related to the variability of the weather. Winnipeg for example shows a tight saw-tooth pattern, and the amplitude of the oscillations changes little over the 24 hour period. At Gander and Toronto on the other hand, the amplitude of the oscillations increases in the second 12 hour period.

c) duration of the forecast: In the second 12 hour period, the amplitudes of the oscillations are greater and the forecast and climatology scores are closer rivals. The gap with persistence widens. However, compare Winnipeg with Toronto and Gander for changes in amplitude and the separation of forecast climatology and persistence.

Figure 5.5 displays the annual average(weighted) RPS values. The period selected was June to May to best accommodate the data available. With fewer data points a bar graph was used to better emphasize the similarities and differences. Seasonal trends have now been replaced with annual trends. The steady increase in the forecast RPS value for Toronto during the first 12 hour period cannot necessarily be interpreted as increased accuracy on the part of the forecaster, but may be due to other factors such as changes in the persistence of the weather. Note that the slope for persistence increases more rapidly than the slope for forecast and climatology. Recall that the RPS, like the Brier Score, is sensitive to the climatology of the verification sample. Thus, any comparison of scores between Winnipeg and Gander would be biased by the differences in climatology. Furthermore, analysis of year to year trends for specific stations must include an accompanying analysis of differences in year to year climatology.

a) Average Weighted Seasonal RPS for Toronto (1-12 h)



b) Average Weighted Seasonal RPS for Toronto (13-24 h)

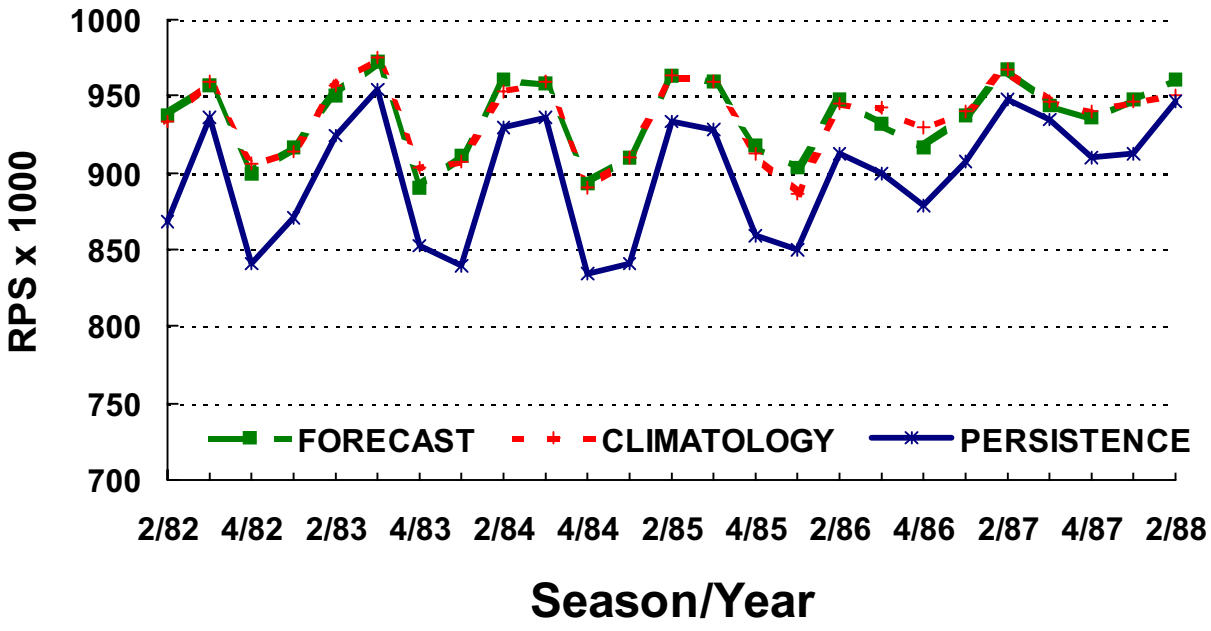
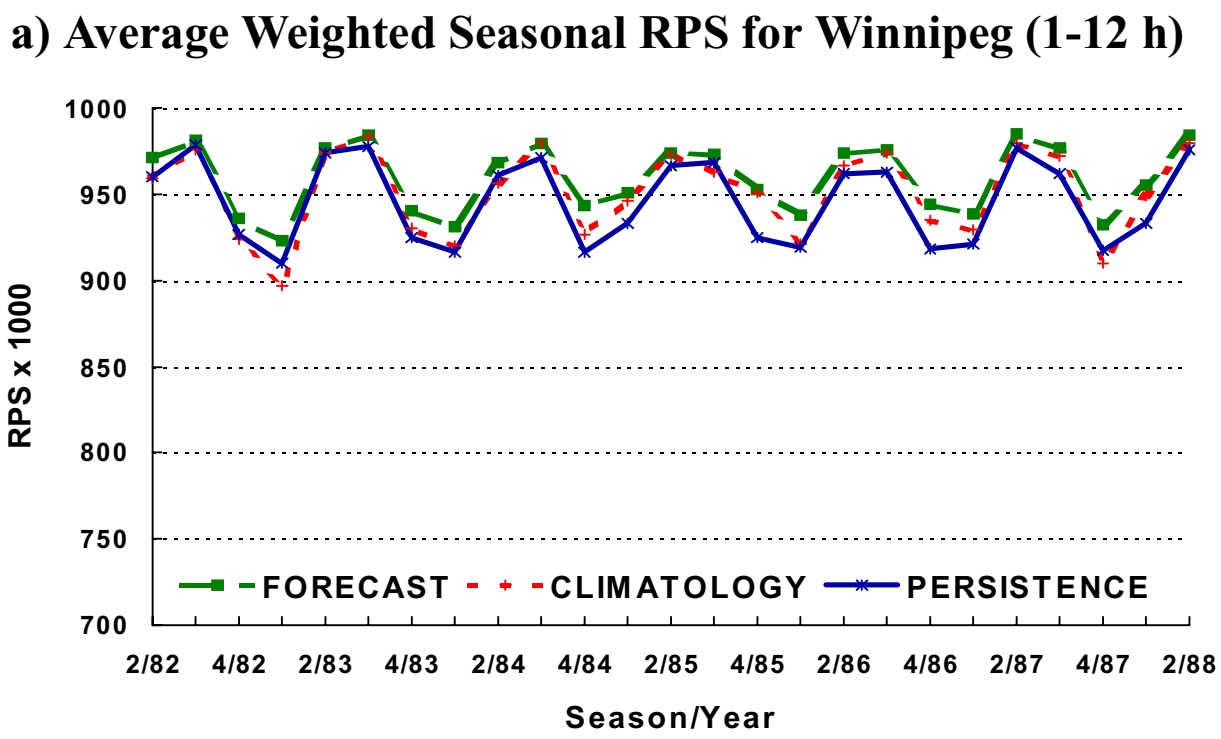


Figure 5.2. Seasonal (weighted) averaged RPS for aviation forecasts made at Toronto. a) 1-12h and b) 13-24h.



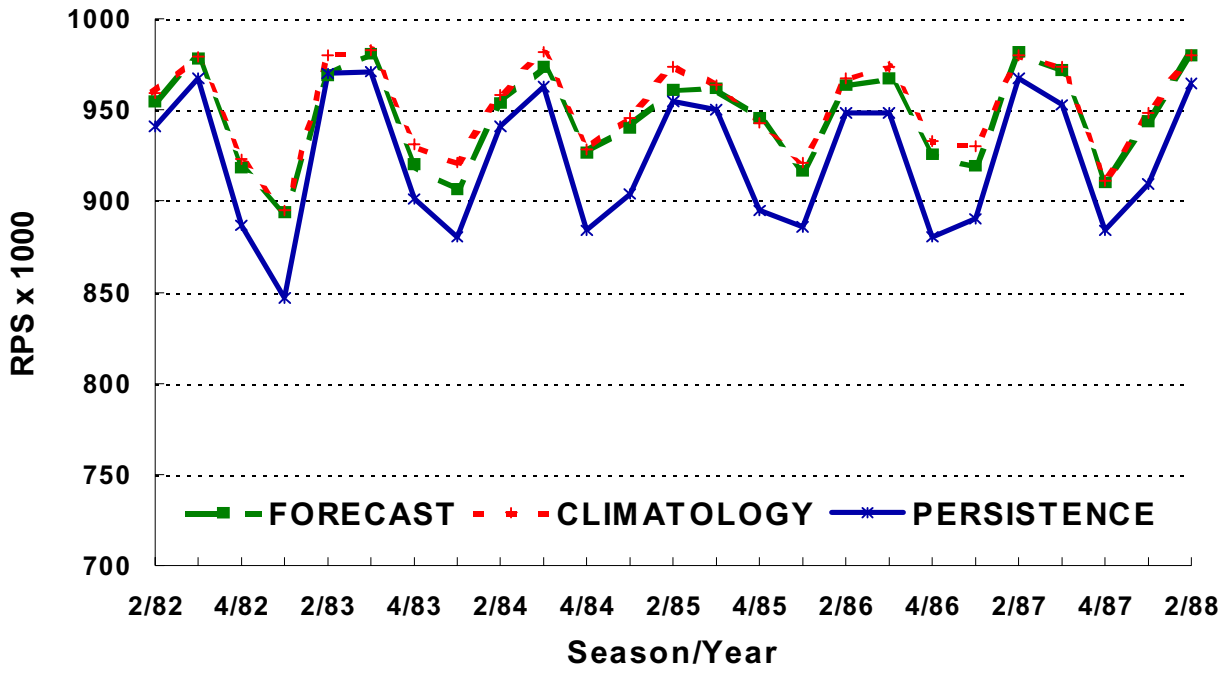


Figure 5.3 Seasonal (weighted) averaged RPS for aviation forecasts made at Winnipeg. a) 1-12h and b) 13-24h.

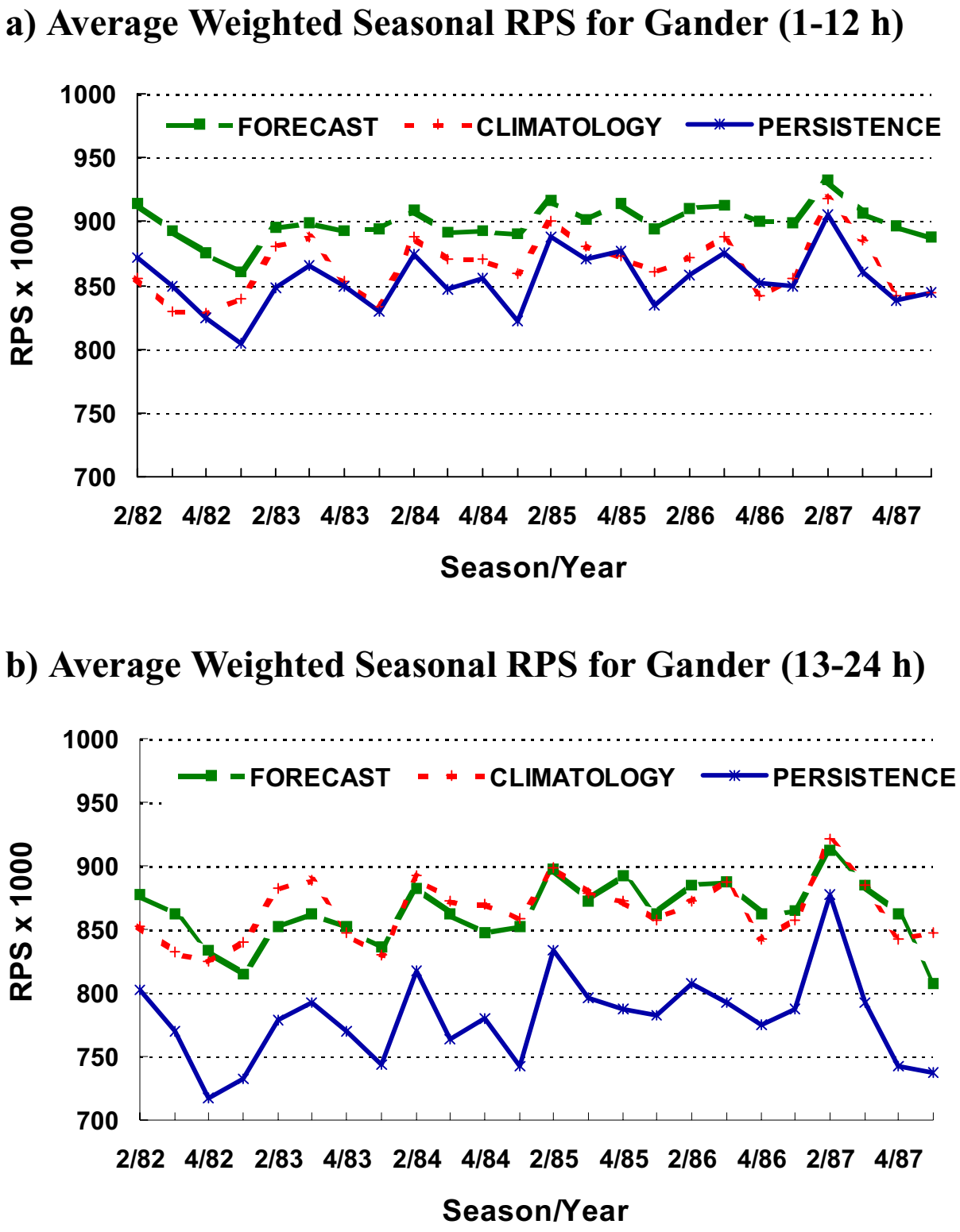


Figure 5.4. Seasonal (weighted) averaged RPS for aviation forecasts made at Gander. a) 1-12h and b) 13-24h.

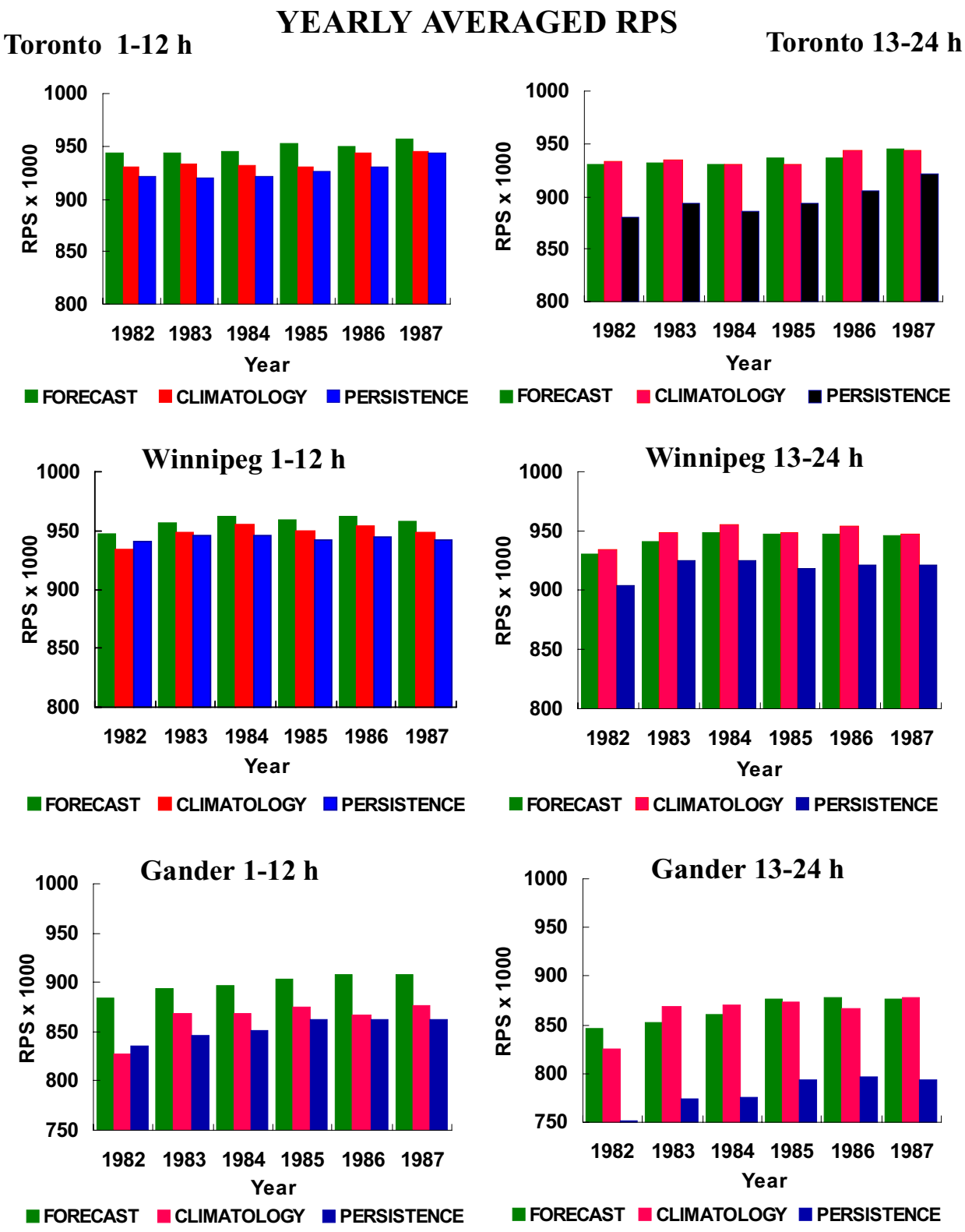


Figure 5.5. RPS for aviation forecasts displayed as weighted yearly values, with comparisons to climatology and persistence

5.2 Verification Suggestions

Several problems have been hinted at in the description of the verification scheme. The lessons to be learned from the experience are summarized in this subsection.:

1) Verification data should be archived on a computer and follow strict database formats. General purpose statistical/graphical programs should be able to access the verification database with minimal effort.

2) The verification data should be saved in its raw or original form, archiving of processed scores and performance measures is optional. By archiving the raw data, the opportunity to go back and re-evaluate the data is possible.

3) The database should be organized such that various stratifications, such as issue time, valid time, season, location, duration etc are possible.

4) The verification archive should be a centralized process, the option/desire to regionalize the archiving procedure should be strongly avoided.

5) A mechanism for short term feedback of summary verification scores for individual forecasters or forecast offices is neither statistically beneficial nor resource efficient. However, with the rise of artificial intelligence one should keep an open mind with this suggestion.

6) Automatic decoding of elements from subjective forecasts (with/without formats) requires extensive testing under varied conditions and will probably result in a success rate of 85-95%. The decode program should have included a data dictionary (used to interpret the forecast) with the capability of learning new idea structures interactively. Caution should be taken when converting an area forecast into a point forecast, e.g. orographic considerations affecting only a portion of the city maybe included in the overall city forecast.

7) Human resources and the requirements for maintaining computer programing will be under-estimated.

8) Any attempt to design a verification system to meet several purposes at once, both administrative and scientific for example, should be avoided. Conflicts in the design features will make the system unsuitable for all purposes. Small is better than multi-purpose.

9) Ensure that the verification measures employed meet the intended purpose, and then exercise PATIENCE as sufficient data accumulates, which may take many years.

6. REFERENCES

Alexander, J.H.,and W.R. Burrows, 1981: Verification of NWP MSL progs-pressure center verification. A.E.S.Training Branch Internal Publication.

Appleman, H.S., 1960: A fallacy in the use of skill scores *Bull. Amer. Meteor.* **41**, 64-67

Bengtsson, L., 1985: Medium-range forecasting-the experience of ECMWF. Bull. Amer. Soc., **66**, 1133-1146.

Brantstator, G., 1986: The variability in skill of 72-hour global-scale NMC forecasts. *Mon. Wea. Rev.*, **114**, 2628-2639.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone, 1984: Classification and Regression Trees. Wadsworth & Brooks, California, 358pp.

Brier, G.W. and R.A. Allen, 1951: Verification of weather forecasts, Chapter in *Compendium of Meteorology*, American Meteorological Society, Boston, 841-848

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

Burrows, W.R., 1976: A diagnostic study of atmospheric spectral kinetic energetics. *J.Atmos.Sci.*, **33**, 2308-2320.

Burrows, W.R., 1990: Tuned perfect prog forecasts of mesoscale snowfall for southern Ontario. *J. Geophys. Res.* **95**, D3, 2127-2141.

Chen, W.Y., 1989: Another approach to forecasting forecast skill. *Mon.Wea.Rev.,* **117**, 427-435.

Clemen, R.T. and A.H. Murphy, 1986: Objective and subjective precipitation forecasts: statistical analysis of some interrelationships *Weather and Forecasting*, **1**, 56-64

Daley, R.,and R.M.Chervin, 1985: Statistical Significance Testing in Numerical Weather Prediction. *Mon.Wea.Rev.*, **113**, 814-826.

Epstein, E.S.,1969: A scoring system for probability forecasts of ranked categories *J. of Appl. Meteor.*, **8**, 985-987.

Finley, J.P., 1884: Tornado prediction, *American Meteorological J.*, **1**, 85-88.

Flueck, J.A., 1987: A study of some measures of forecast verification *Tenth Conference on Probability and Statistics in Atmospheric Sciences* Oct 6-8, 1987 Edmonton, Alta., Canada; Amer. Meteor. Soc. Boston, Mass.

Glahn, H. R., 1976: Forecast evaluation at techniques development laboratory. Chapter in *Weather Forecasting and Weather Forecasts: Models, Systems and Users*, National Center for Atmospheric Research, Boulder, Colo. 2, 831-838.

Hanssen, A.J. and W.J Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Koninklijk Nederlands Meteorologist Institua* Meded. Verhand, **81**, 2-15.

Heidke, P., 1926: Berechnung des Erfolges und der Guete der Windstarkevorhersagen in Sturmwarnungsdienst *Geografika Annaler* **8**, 310-349.

Hsu, W.-R. and A.H. Murphy, 1986: The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2**, 285-293.

Leary, C., 1971: Systematic errors in operational National Meteorological Center primitive equation surface prognoses. *Mon Wea.Rev.*, **99**, 409-413.

Mason, I., 1980: Decision-theoretic evaluation of probabilistic predictions. *WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting*, Nice, September 8-12, 1980, 219-228.

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

Miller, R.G. and D.L. Best, 1979: A model for converting probability forecasts to categorical forecasts, *Sixth Conference on Probability and Statistics in Atmospheric Sciences* Oct 9-12, 1979 Banff, Alta., Canada, Amer. Meteor. Soc.Boston, Mass., 98-102.

Miyakoda, K., G.D. Hembree, R.F. Strickler, and I. Shulman, 1972: Cumulative results of extended forecast experiments. Part1: Model performance for winter cases. *Mon.Wea.Rev.*, **100**, 336-355.

Muller, R.H. 1944: Verification of short-range forecasts (a survey of the literature) *Bull. Amer. Meteor. Soc.* **25**, 18-27, 47-53, 88-95.

Murphy, A.H., 1969: On the ranked probability score *J. of Appl. Meteor.*, **8**, 988-989.

Murphy, A.H., 1970: The ranked probability score and the probability score: a comparison *Mon. Wea. Rev.* **99**, 917-924.

Murphy, A.H., 1971: Scalar and vector partitions of the ranked probability score, *Mon. Wea. Rev.*, **100**, 701-708.

Murphy, A.H., 1973: A new vector partition of the probability score. J. *Appl. Meteor.* **12**, 595-600.

Murphy, A.H., B.G. Brown and Y. Chen 1989: Diagnostic verification of temperature forecasts *Wea. Forecasting* **4**, 485-501.

Murphy, A.H., Y. Chen and B.G Brown, 1987: Diagnostic verification of temperature forecasts: some preliminary results *Tenth Conference on Probability and Statistics in Atmospheric Sciences* Oct 6-8, 1987 Edmonton, Alta, *Amer. Meteor. Soc.*, 83-90.

Murphy, A.H., Y. Chen and R.T. Clemen, 1988: Statistical analysis of interrelationships between objective and subjective temperature forecasts *Mon. Wea. Rev.*, **116**, 2121-2131.

Murphy, A.H., and H. Daan, 1985: Forecast evaluation. Probability, Statistics, and Decision Making in the Atmospheric Sciences. A.H. Murphy and R.W Katz, Editors. Westview Press, Boulder, Colorado, 349-437.

Murphy, A. H. and E.S. Epstein, 1967: Verification of probabilistic predictions: a brief review, *J. of Applied Meteorology*, **6**, 748-755.

Murphy, A.H.,and E.S. Epstein, 1989: Skill Scores and Correlation Coefficients in Model Verification. *Mon. Wea. Rev.*, **117**, 572-582.

Murphy, A.H., W. Hsu, R.L. Winkler, and D.S. Wilks, 1985: The use of probabilities in subjective quantitative precipitation forecasts: some experimental results, *Mon. Wea. Rev.*, **113**, 2075-2089.

Murphy, A.H. and R.L. Winkler, 1987: A general framework for forecast verification *Mon. Wea. Rev.* **115**, 1330-1338.

Palmer, T.N.,and S.Tibaldi, 1988: On the prediction of forecast skill. *Mon. Wea. Rev.,***116**, 2453-2480.

Reid, J.D., 1978: Verification of ceiling and visibility forecasting. *Atmosphere-Ocean*, **16 (2)**, 177-186.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191-201.

Sanders, F., 1967: The verification of probability forecasts. *J. Appl. Meteor.*, **6**, 756-761.

Silberberg, S.R.,and L.F.Bosart, 1982: An analysis of systematic cyclone errors in the NMC LFM-II model during the 1978-79 cool season. *Mon. Wea Rev.*, **110**, 254-271.

Swets, J.A., and R.M. Pickett, 1982: Evaluation of diagnostic systems. Academic Press, New York, 253pp.

Szucko, J.J., and B. Kleinmuntz, 1981: Statistical versus clinical lie detection. *American Psychologist*, **36**, 488-496.

Teweles, S.and H.B.Wobus, 1954: Verification of prognostic charts. *Bull. Amer. Met. Soc.*, **35**, 455-463.

7. Data Sources

Fig. 1.2 Data from National Verification System monthly reports, Weather Services Directorate, Atmospheric Environment Service.

Fig 2.1 Evaluation of East Coast Marine Wind Forecast Techniques - Final Report SSC file #02SE.KM191-6-6384, Maclaren Plansearch Limited, 1987.

Fig 2.2 to 2.5 Data from Canadian Atlantic Storms Project (CASP), figures supplied by R. Sarrazin and L. Wilson.

Fig 2.9, 2.10 Verification of the 1981 Operational Probability of Precipitation Amount Forecasts, H. Stanski and L. Wilson, internal report.

Fig 2.11a,b Data from Canadian Meteorological Centre verification system.

Table 2.17 From Table 10, Wilson, L.J. and R. Sarrazin, 1989: A classical-REEP Short Range Forecast procedure. *Wea. Forecasting*, **4**, 502-516.

Fig 2.12 to 2.14 Data from National Verification System monthly reports, Weather Services Directorate, Atmospheric Environment Service.

Fig 3.1 *The Monthly Review*, Vol VI No. 4 May 1988, The Canadian Meteorological Centre.

Fig 3.2 after Brantstator, 1986.

Fig 3.3, 3.4 after Bengtsson, 1985.

Fig 3.5 data from Canadian Meteorological Centre verification system.

Fig 3.6 data from Alexander and Burrows, 1981.

Fig 4.2, Table 4.1 Data from Canadian Meteorological Centre verification system.

Fig 4.3 The Very Short Range Forecasting Project, Information for Meteorologists, L. Wilson and R. Sarrazin, Research Report MSRB-86-4, 1986.

Fig 4.4, 4.5 data from Canadian Meteorological Centre verification system.

Fig 4.6 Comparison of MOS and Perfect Prog Probability of Precipitation Forecasts using the Signal Detection Theory Model. L. Wilson and R. Sarrazin, 1987, WMO, *Programme on Short- and Medium-Range Weather Prediction Research (PSMP)*, Toulouse, France, 22-26 June 1987.

Table 5.2 after Reid, 1978.

Fig 5.1 to 5.8 Data from National Verification System monthly reports, Weather Services Directorate, Atmospheric Environment Service.