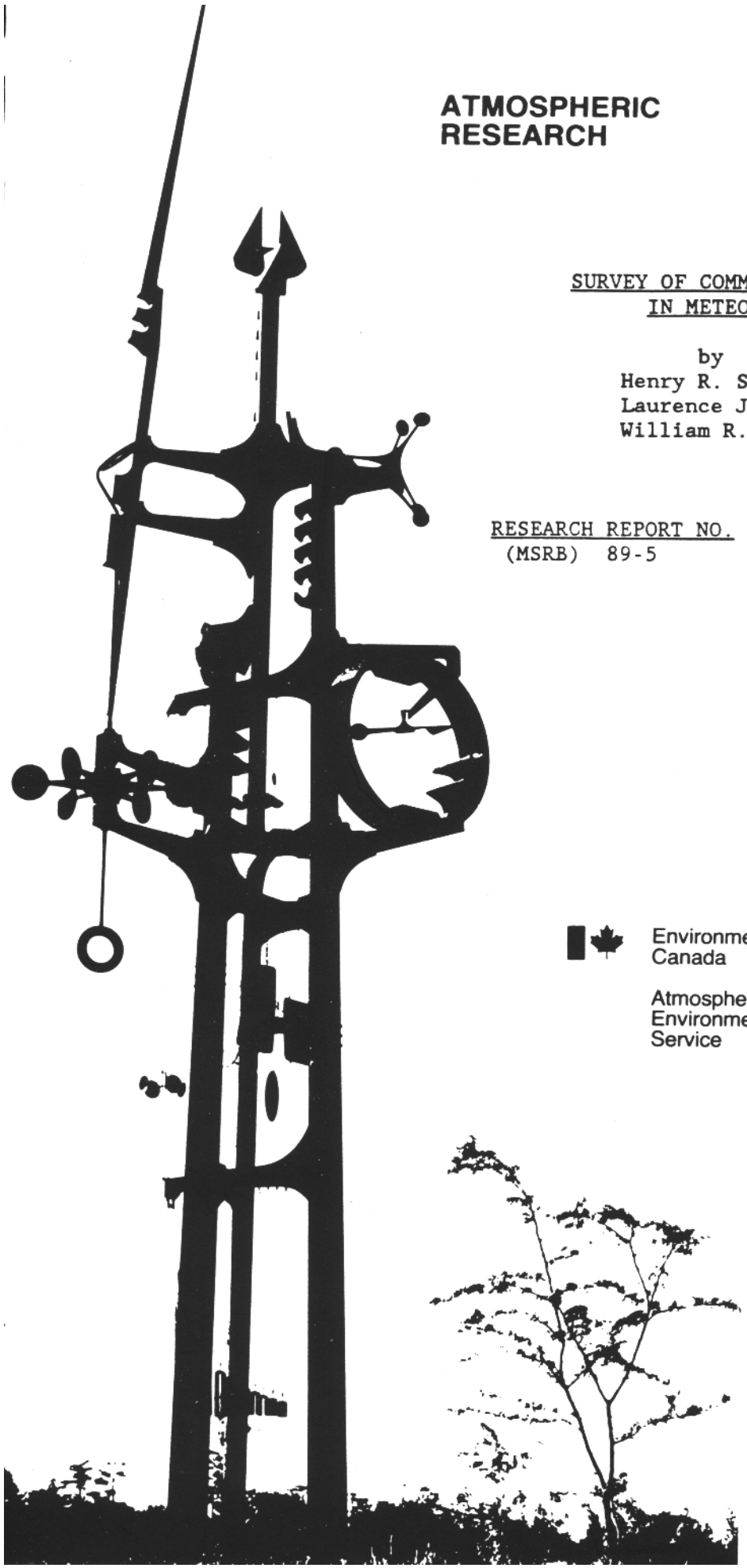


ATMOSPHERIC RESEARCH

RECHERCHE ATMOSPHÉRIQUE

SURVEY OF COMMON VERIFICATION METHODS IN METEOROLOGY

by
Henry R. Stanski
Laurence J. Wilson
William R. Burrows

RESEARCH REPORT NO.
(MSRB) 89-5

July, 1989

Environment Canada | Environnement Canada

Atmospheric Environment Service | Service de l'environnement atmosphérique

SURVEY OF COMMON VERIFICATION METHODS IN METEOROLOGY

by

Henry R. Stanski
Laurence J. Wilson
William R. Burrows

Second Edition
1989

Atmospheric Environment Service
Forecast Research Division
4905 Dufferin Street,
Downsview, Ontario,
Canada
M3H 5T4

ABSTRACT

Commonly used verification methods for weather elements and numerical weather prediction model forecasts are described in terms of a general verification model. The advantages and disadvantages of each verification method are identified and discussed. Furthermore, numerous examples using meteorological data are provided. New ideas in verification are included in a separate chapter.

RÉSUMÉ

Les méthodes de vérification généralement utilisées par le modèle de prévision et prédiction d'éléments atmosphériques et numériques sont décrites en terme générale du modèle de vérification. Les avantages et désavantages pour chacune de méthodes de vérification sont identifiés et discutés. En plus, de nombreux exemples utilisant les données météorologiques sont inclus. De nouvelles idées dn vérification sont inclues dans une section indépendente.

FOREWORD

This document is an extensively revised version of Module VIII, Verification of Forecasts, by Henry Stanski (1982). The original document was written as part of an internal AES training course, but has enjoyed wide circulation and interest nationally and internationally as a handy guide to verification methods. Since 1982, there have been significant advances in verification methodology and changes in thinking about purposes and design philosophy of verification methods for weather forecasts. This document is designed to take into account those changes, while hopefully retaining the "handy catalogue" characteristics that contributed to the success of the original document.

Material for the present document is taken from two main sources: the original document, and notes from a laboratory on verification methodology that was presented at the Second AES-CMOS* Operational Meteorology Workshop, Halifax, Oct 14 to 16, 1987. To this has been added numerous new examples, and information from published literature on verification methods. In particular, a paper on a general framework for verification (Murphy and Winkler, 1987) has influenced the contents of the document.

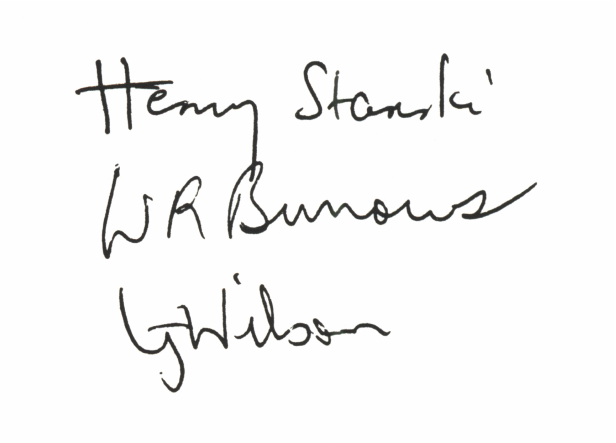
The objectives of this manual are:

1. To explain and encourage practical applications of verification methods through understanding and example.

2. To bring some consistency of thought in AES to the design, execution, and interpretation of weather forecast verification.

We have deliberately limited the contents to those verification methods and measures that have enjoyed widespread general use in the past several years, from our perspective. Newer methods that perhaps are not yet widely used are described in Chapter 4. If your favourite verification score has been omitted, please let us know, and we can consider its inclusion in the third edition.

* Atmospheric Environment Service, Canadian Meteorological and Oceanographic Society



Preface to the Electronic Edition

We decided to undertake the task of producing an electronic edition of this document for two main reasons:

1. Fourteen years after it was first issued as an internal report, we are still getting requests for the paper copy. While we remain happy to fulfull these requests, distribution by electronic means will be easier and will make the document more readily accessible.
2. Members of the international meteorological verification community are in the process of developing a web-site on forecast verification. This site would seem a logical repository for an electronic edition of this document.
3. The development of computer technology over the years means that the formatting could be improved and graphics could be enhanced with the use of colour.

We have not changed the contents of the document, nor the order of the material. It was tempting to do so, especially to update some parts which are now somewhat out of date, but we felt it better to leave the contents as in the original, in order to put reasonable limits on the task of producing this version. This version is shorter than the original, 81 pages instead of 115. This was achieved through the elimination of white space, and through the use of a smaller font for the text. The text was recovered from the original "WORD for DOS" files, but not all the characters translated correctly. We have proofread the text to try to eliminate these problems, but some may have slipped through. Please let the authors know of any editorial problems, and they will be corrected. Figures which came from other documents have been scanned from the original "cut-and-paste" versions at as high a resolution as the pdf conversion software would allow. Those figures which were recoverable from the original "Harvard graphics" files were imported and reformatted as necessary. The flowcharts in Chapter 1 were reentered and colour coded.

We have been pleased and rather amazed at the success enjoyed by this document over the past 14 years. We hope that this version will be useful too, and would be glad for any comments or suggestions for improvement.

Laurence J. Wilson, <lawrence.wilson@ec.gc.ca>

William R. Burrows, <William.Burrows@ec.gc.ca>

Henry Stanski, <Henry.Stanski2@ec.gc.ca>

ACKNOWLEDGEMENTS

We would like to thank the various individuals and agencies for their co-operation in supplying data for the tables and figures. Appropriate references of such sources are given throughout the document and in section seven.

We also wish to express gratitude to Dr.A. Murphy and Dr.S. Venkatesh for reviewing the document. Dr.A. Murphy has been especially generous with his time whenever he was approached.

Table of Contents

1. A VERIFICATION FRAMEWORK

Verification is the assessment and quantification of the relationship between a *matched set of forecasts and observations*. Verification activities are useful only if they lead to some decision regarding the product being verified. That decision will either generate changes in the product or in the way forecasts are made, or it might be a "do nothing" decision which confirms that the product or service is satisfactory. The forecasts must be written with sufficient objectivity to be verifiable. While observations are assumed to be an accurate representation of reality, indeed some verification may require the assumption that a point observation adequately represents the events of an area.

It is imperative that the verification goal be established before the verification system is designed because the verification purpose has strong implications for the design.

Verification activities in meteorology are directed by two main types of goals :

1) ***Administrative***.... Initially, the Meteorological Service of Canada (since 1871) used verification to justify to Parliament the provision of a forecast service. In addition to justifying the cost of provision of a weather service, administrative uses of verification information include support for the purchase of major equipment such as larger computers, the determination of when or whether to replace a forecast product with a new one, and many other decisions on the optimum deployment of human and equipment resources in a weather service. Administrative verification is also done, on a continuing basis, to monitor the overall quality of forecasts and to track changes in their quality over periods of time.

2) ***Scientific***.... To identify the strengths and weaknesses of a forecast product in sufficient detail that actions can be specified which will lead to improvements in the forecast. That is, to provide information to direct research and development.

Ignored here is verification for utility or economic purposes. Utility verification requires the input of information regarding the economics of actions taken in response to the forecast, which are highly user-specific. Instead the attention is fixed on the attributes of the forecast from the point of view of the originator or generator of the forecast. Increasingly governments have taken the position that the public purse can afford to provide only a basic level of service, hence specialized forecasts and utility verification are often left to the users to do.

1.1 A Verification Model

Figure 1.1 presents a general model for verification of weather element forecasts, and it also serves to illustrate and summarize the types of decisions that must be made before specific verification measures can be chosen. We emphasize that this is only one model of verification strategy; other equally valid models may exist. However this model presents a consistent picture of the relative characteristics of verification measures and the types of decisions that lead to their selection.

All verification starts with a matched set of observations and forecasts (the parallelogram at the top). Once the matching has been done, the next steps in processing the datasets depend on the decisions that have to be made (the diamond shapes on the diagram). The most important decisions relate to the use of the verification output, classed as administrative or scientific, as described above. Examples of specific questions that are addressed in scientific verification are: "What does a temperature forecast of -10 °C really mean?" or "Can this forecast method discriminate reliably between VFR and IFR[1] conditions?" or "How accurately can I forecast extreme rainfalls?" Examples of specific administrative questions are: "Is the accuracy of the forecasts improving?" or "Are objective temperature forecasts better than subjective ones?"

1. VFR Visual Flight Rules
IFR Instrument Flight Rules

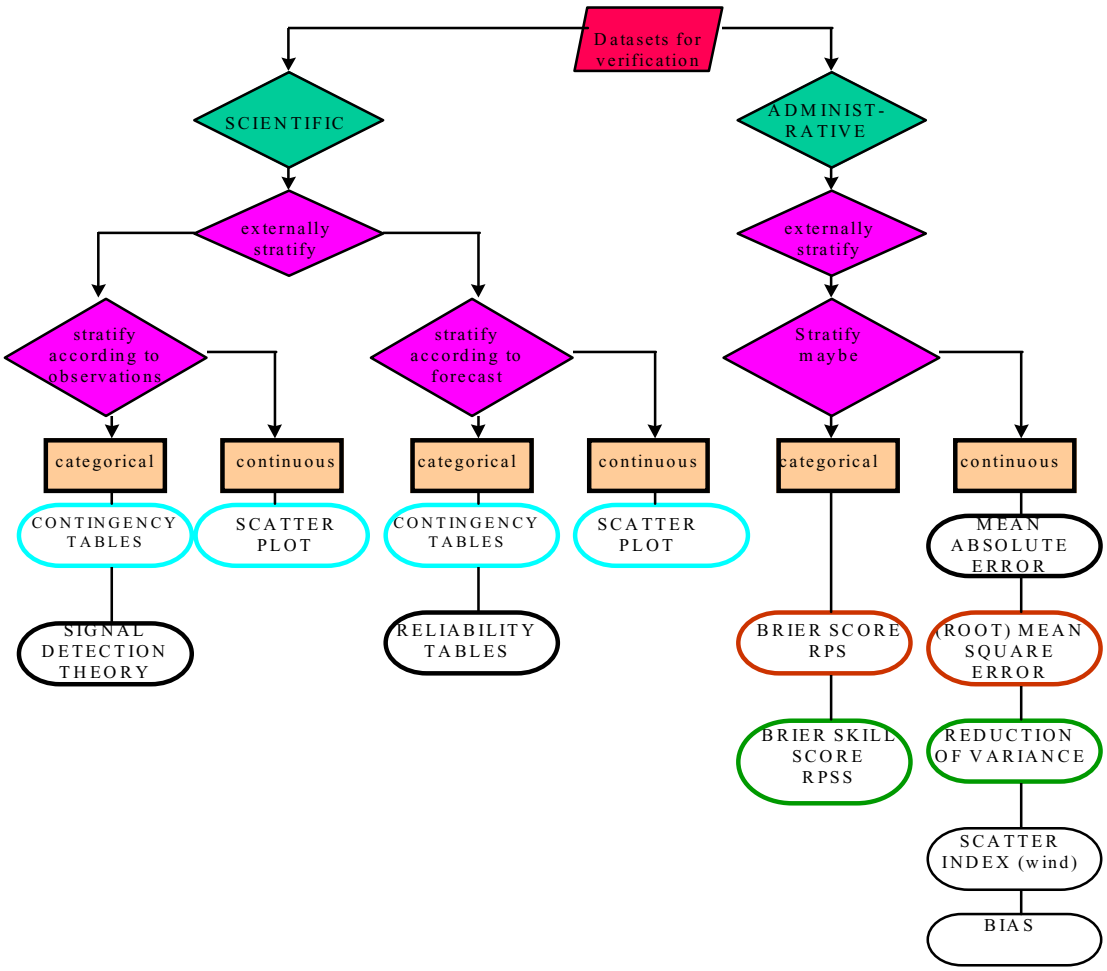


Figure 1.1. A general model for verification of weather element forecasts.

Once the purpose of the verification is established, the sample may be stratified to help meet the intended purpose. **Stratification** means separating the events of the sample into two or more groups according to a selection rule, then carrying out the verification for each group separately. "**External stratification**" refers to stratification by selection rules that are independent of the element being verified. A common type of external stratification that allows identification of variations in the verification is by time of day (diurnal) or by season. For example, it is usually true in Canada that precipitation forecasts are more accurate in winter than in summer because in the summer season the small scale variability of precipitation (convective) is hard to predict accurately (fig 1.2). External stratification can be done at any time in the process before computation of the actual verification statistics, and, may be done for either administrative or scientific purposes.

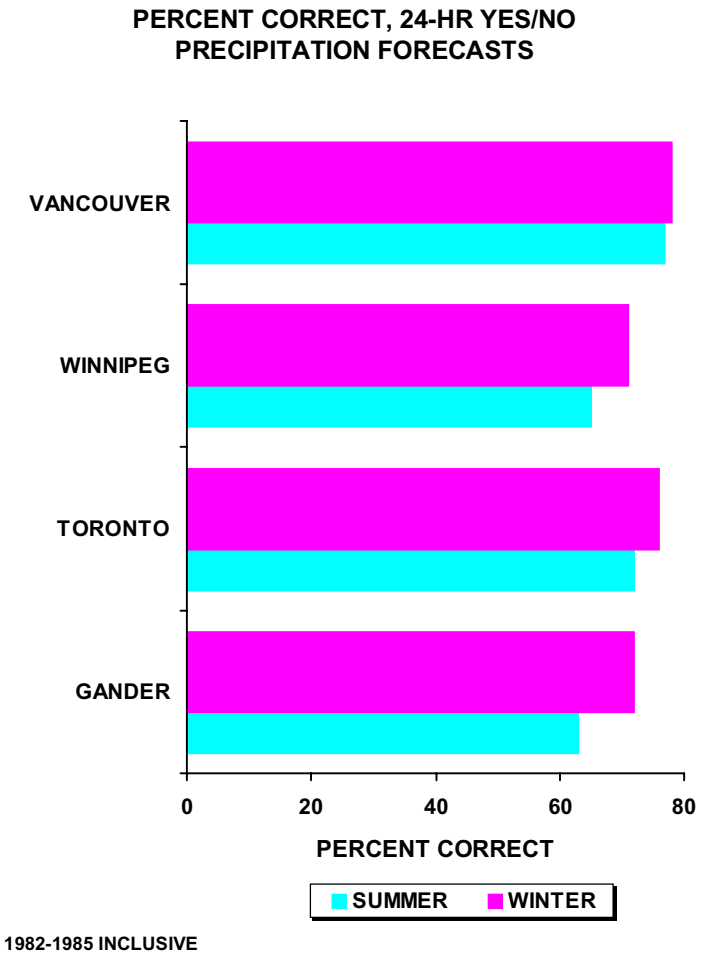


Figure 1.2. Percent correct for 24 h yes/no forecasts for four Canadian stations according to season. Winter means January, February and March; Summer means June, July and August.

It is clear from the examples of scientific verification questions that there will usually be a need for further stratification of the sample. If extreme forecasts are of interest, for example, the sample must be stratified to separate those events from non-extreme events. This type of stratification may be called "**internal stratification**" because categorization rules are determined according to the purpose of the verification, using the element that is being verified. There are two ways of doing internal stratification, and Figure 1.1 suggests they have quite different implications for the types of verification results that are obtained.

Stratification according to the observation means defining categories according to observed values of the weather element. Then, verification measures can be calculated for each category of the observations, and such statistics are said to be **conditioned on the observation**. An example is a conditional distribution of forecasts given a specific value or range of values of the observation.

Stratification according to the forecast means defining categories according to forecast values of the weather element. Analogously, statistics calculated for these categories are said to be **conditioned on the forecast**. Examples of conditional distributions of both types are described in Chapter 2.

Which type of stratification to choose depends on the question that is to be answered by the verification. Indeed, many questions will require both types of stratification to provide a complete answer. Furthermore, as suggested by the diagram, different verification measures (the oval boxes at the bottom) imply one or the other type of

stratification and give different information about the product.

Administrative verification, on the other hand, is less concerned with performance variations according to different values of the predictand. Instead, the questions asked are of a more general nature and suggest summarizing information in some way. Thus internal stratification may be done, but rarely is in our experience. The overall thrust of administrative verification is to represent the quality of a product in as few numbers as possible, or to facilitate comparisons or identification of trends. The summary nature af administrative verification is easily associated with "**summary scoring rules**", as Figure 1.1 shows.

Occasionally, the drive to summarize is carried to extremes, namely there have been attempts to provide management with a single number that represents the quality of all forecasts issued by the service. This need to summarize verification information in a single number puts tremendous pressure on the design of the verification system to ensure that: a) the chosen score reliably measures what is desired; and b) the component events are treated fairly in the composition of the score. A common complaint about summary verification is that all events are treated equally in the averaging process. This is done for convenience (a simple average is easy to compute), but also because it is difficult to find a weighting scheme that accurately and objectively reflects the importance of the component events for the intended purpose, without interfering with other desirable attributes of the score. How to weight the component events in a summary verification remains an unsolved problem.

Summary verification scores have been oversold in the past. Forecasters still express frustration when trying to use them to answer specific scientific questions. Their summary nature limits their use for scientific purposes because of the lack of stratification of the verification in terms of antecedent conditions. A summary score cannot say for example, how well precipitation is forecast in cutoff low situations compared to transient wave situations; it only says how well precipitation is forecast overall. It cannot tell under what conditions the Regional Finite Element model (RFE) is most preferred to the Spectral model; it can only say that the RFE is slightly better or slightly worse than the Spectral model on average.

1.2 Predictand Types

After all the decisions about stratification have been made, and the exact question(s) to be answered by the verification system have been posed, one is in a position to choose appropriate measures to answer the questions. The set of rectangular boxes on Figure 1.1 suggests that the choice of appropriate measures depends also on the nature of the forecasts to be verified. For verification purposes, there are two distinct types: continuous and categorical/probabilistic. **Continuous predictands** are those elements where a specific value or range of values is forecast. Among weather elements, only temperature and wind are nearly always forecast this way. For example, "Low near -10 °C tonight", or "Winds west 15 kmh gusting to 25". **Categorical predictands** are those elements for which the forecast is of the occurrence of the event in one of two or more mutually exclusive and exhaustive categories of the element. Examples are the occurrence of measurable precipitation (two categories - either it rains or it doesn't), or precipitation type (usually three categories, frozen, freezing or liquid).

Some elements may be forecast either categorically or continuously, the choice depending mainly on user requirements for detail in the forecast. If there were a demand for more detail, forecasters would presumably be willing to forecast precipitation amount in mm, but, unless heavy precipitation is expected, a categorical Yes or No, or a probability forecast of measurable precipitation is sufficient. Verification systems respond to user requirements for detail in a similar way; a forecast expressed as a continuous variable may be verified categorically because that is all the information that is needed by the user. Ceiling is a good example of this; it may be forecast to the nearest 100 feet over at least part of its range, but it is usually verified according to categories that are significant to aviation.

Probability forecasts are viewed as more general categorical forecasts. They apply to categorized predictands as well, but each category is assigned a probability of occurrence. The probabilities must add up to 1 over all the categories of a predictand (something must happen). Conversely, a categorical forecast is a restrictive probability forecast where only the probabilities 0% and 100% are allowed for all categories. Categorical forecasts imply certainty that the chosen category will occur.

1.3 Survey of verification measures

The verification measures listed at the bottom of Figure 1.1 are paired to show correspondence. For example,

contingency tables and scatter plots are completely analogous, each providing the same types of information, the former for categorical predictands and the latter for continuous predictands. The summary scores are also paired this way: The Brier Score (category=2) and the Ranked Probability score (category >2) measure exactly the same characteristics of a probability or categorical forecast as the mean squared error measures for a continuous forecast, for example. Note that there is no categorical forecast analogue to the mean absolute error. The expected analogue, mean probability error is not used because it is not **strictly proper** (meaning that it is possible to improve the score by systematically forecasting probability values other than one's best estimate). The two measures "reliability tables" and "signal detection theory" imply stratification according to forecast and observation respectively, while contingency tables and scatter plots are more general, permitting stratification either or both ways. Signal detection theory is discussed in Chapter 4 as a new idea in verification; it is not widely used at present.

Scores listed in Figure 1.1 and described in Chapter 2 are of only three types: linear scoring rules, quadratic scoring rules, and skill scores. Quadratic scoring rules give weights to errors according to their square while linear scores weight errors linearly. Thus, quadratic scoring rules give relatively higher weights to large errors in the sample than do linear rules, and are most useful when large errors are relatively more serious than small errors.

Skill scores are designed to evaluate forecasts relative to a standard. The standard is chosen to represent an unskilled forecast. The three standards that are usually used, in increasing order of sophistication are: "chance", "persistence", and "climatology". Chance represents a pure guess and requires no prior knowledge; persistence is a no change forecast and requires knowledge of only the initial weather conditions; and climatology is a forecast of the long term average weather, and requires a knowledge of the history of the weather. The form of the skill score is:

$$SS = \frac{SC - ST}{PS - ST}$$

where SC is the score achieved by the forecast, ST is the score obtained using the standard forecast and PS is the score for a perfect forecast. Skill scores may be formed using any of several of the summary scores. The most common skill scores are those based on the Brier Score (Brier Skill Score), on the rank probability score (Rank Probability Skill Score), on the marginal sums of the contingency table (Heidke Score), and on the mean absolute error. The most frequently used standard is climatology, but the Heidke score is almost always associated with chance. No matter what score the skill score is based on, they are always in the same format and measure the attribute of skill.

1.4 A NWP Verification Model

Figure 1.3 shows a general outline for NWP model verification which is consistent with the outline for weather element verification in Figure 1.1. Verification starts with a dataset which consists of forecast and observation data which must be matched both spatially and temporally. For spatial matching, observation data must either be analyzed to the grid of the forecast data, or the forecast data must be interpolated in space to observation points. The former is most often done, which sometimes leads to the criticism that the analyzed observation dataset is not as representative of truth as it was before the interpolation. This criticism becomes especially important to the verification if the analysis is an output trial field of the model which is being verified. In that case, there is potential bias in the results in favour of the model. Thus it is important in model verification to keep in mind the processing procedure for the verifying observations when evaluating results.

Effective NWP model verification also requires that decisions be made about the purpose of the verification before a system is designed (all of the diamond shapes on the figure). "Administrative verification" answers questions about trends in model accuracy and skill and can be used to compare the accuracy of two different models. Similar to weather element verification for administrative purposes, there is a tendency to reduce the results to a few numerical values through the use of summary scores.

"Scientific verification" involves answering questions about spatial and/or temporal variations in the performance of the model, to provide information that can be fed back to model developers to improve the model or to forecasters on how to modify the guidance. One common example is the verification of such features as low pressure areas or fronts. The verification dataset must be carefully stratified according to the characteristics of the features that are to be examined, and this can be done either on the basis of the observed characteristics or on the basis of the forecasts. Case studies are an extreme example of stratification of this type: a single case or event (or storm) is chosen for detailed analysis in an attempt to reveal the strengths and limitations of the model's simulation of the structure

and evolution of the storm.

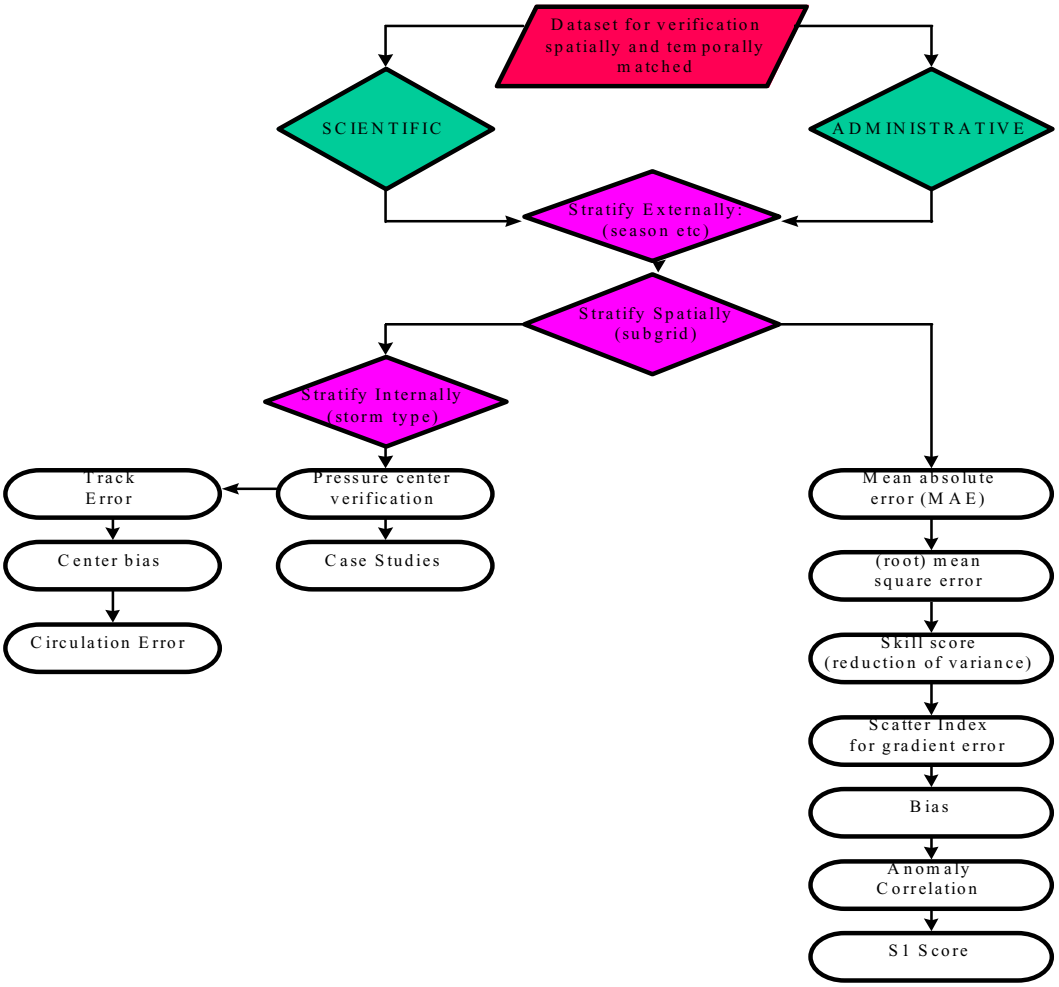


Figure 1.3. A general outline for NWP model verification.

Spatial stratification can be done either for administrative or scientific purposes. When done for administrative purposes, it is usually intended to reveal differences in accuracy over different portions of the model domain corresponding to administrative regions of the country (Atlantic region, Quebec region, etc.). When spatial stratification is done for scientific purposes, the subgrid areas are more likely to be chosen to reflect different climatological regimes such as mountainous areas, lee of mountains, east coast (for east coast storms) etc.

"External stratification" means dividing the verification dataset according to season or model run time or another selection rule that is independent of the parameter being verified. Stratification by season is most common and reveals differences in performance characteristics between seasons.

1.5 Attributes of a Forecast

One final comment about verification measures: NO SINGLE VERIFICATION MEASURE PROVIDES COMPLETE INFORMATION ABOUT THE QUALITY OF A PRODUCT. All give information about one or more aspects of the quality (often referred to as **attributes**) of a forecast product. Thus a verification system will often include computation of several measures chosen to describe the attributes that are most pertinent to fulfilling the goal

of the verification. The attributes that are associated with the different verification measures are defined in this section.

a) *Accuracy* is a general term indicating the level of agreement between forecast weather and true weather as represented by observations. The difference between an observed value and the forecast value is called the error. The smaller this difference is, the smaller the error and the greater the accuracy. The difficult aspect of accuracy is in communicating the idea of the scale of accuracy, i.e the boundaries or limits of acceptability. A statistical forecast for freezing rain may be considered accurate for statistical purposes, but not accurate enough to be of benefit to an operational forecaster. Discussions of accuracy often occur on different levels: one may speak of the average accuracy of the forecasts while another is only interested in the accuracy of the forecast for a particular event or on a particular day; while yet another may emphasize the fact that accuracy is inflated due to the abundance of `good' (easy to forecast) weather events.

Accuracy measures are sometimes considered to be divisible into measures of other component attributes of forecasts (reliability and resolution) and/or observations (uncertainty). The Brier score is an example discussed in section 2.11.

b) *Skill*, or relative accuracy is defined as the accuracy of a forecast relative to the accuracy of forecasts produced by some standard procedure. Common standards, which are considered to have no skill (i.e. the standard forecast can be generated from the observations alone) are climatology, persistence and chance. These standards are also listed in decreasing order of sophistication. Skill scores provide a means of accounting for variations in accuracy which have nothing to do with the forecaster's ability to forecast

c) *Reliability* is equivalent to *bias*; and is simply the average agreement between the stated forecast value(label) of an element and the observed value(label). A distinction can be made between unconditional bias (overall reliability or systematic error) and conditional bias (reliability). Often, reliability can be improved by feeding back bias information to the forecaster, thereby affording the opportunity to remove it from future forecasts.

For continuous variables, the reliability may be expressed as the ability to forecast maximum temperatures to within plus or minus 3 degrees Celsius. The reliability for categorical/probability forecasts can be quantified in a reliability table where for example 70% POP forecasts (Probability of Precipitation) may correspond to an observed frequency of 90%. This POP forecasts underforecasts the true frequency by 20%. Reliability with respect to numerical progs has a slightly different meaning in as much as spatial characteristics are emphasized. A certain model for example may have the tendency to move lows too far north by 300nm and overforecast the deepening by 8mb. Reliability has the same interpretation for all forecasts.

d) *Resolution* is the ability of the forecast to sort or resolve the set of sample events into subsets with different frequency distributions. Resolution is related to the standard deviation or variance of the observations stratified by the forecast. Resolution is tied to the overall experience, and understanding of the forecaster, i.e resolution measures the state of the art. For example, if the distribution of observed temperatures when -10°C is forecast is greatly different from the distribution of observed temperatures when -5°C is forecast, the temperature forecasting system is said to have resolution. Note that it doesn't matter that the label be correct; the mean observed temperature for forecasts of -5°C could be 0 and the mean observed temperature for the -10°C forecasts could be -15°C; all that is important for resolution is that the forecast labels divide the sample into characteristically different components. A probability forecast is said to have resolution if the observed frequency of the event when 20% is forecast is noticeably different from the observed frequency when 70% is forecast, for example. Again, the labels don't have to be correct (that's reliability), only different in terms of the observations.

Model resolution is interpreted spatially usually, but the concept is the same. Using geopotential height as an example, if the mean observed height for an area where low heights were forecast is clearly different than the mean observed height for an area where high heights were forecast, then the model forecast is said to have resolution.

e) *Sharpness*, an attribute of forecasts alone, refers to the tendency to forecast extreme values. For probability forecasts, this is the tendency toward forecasting 0% and 100% probability, that is, the tendency toward categorical forecasts. Sharpness is not defined for non-probabilistic forecasts unless interval or multiple-category forecasts are permitted. Thus for continuous forecasts, it is the tendency to forecast extreme values. In both cases, it represents the tendency to "go out on a limb" and is directly related to the variance of the distribution of forecasts. Sharpness with respect to models may be found for example in the ability to forecast a hurricane, to position the low centres within a

multi-low system (i.e. to describe the fine scale structure), or to establish cyclogenesis. Note that for both models and weather elements, it doesn't matter for sharpness that the forecast be right or wrong, it is only the attempt to forecast fine scale detail or extreme values that matters. Sharpness can be increased by postprocessing the forecast using mathematical procedures such as "inflation", but only at the expense of reliability.

The essential difference between resolution and sharpness is that the former depends on both observations and forecasts while the latter depends only on the forecasts. It is possible to have a forecast which is sharp but has no resolution. For example, a probability forecast of rain/no rain is sharp by tending to forecast only 100 and 0%, but has resolution only if the forecast frequency of rain given that rain occurred, is significantly different from the forecast frequency of rain when rain didn't occur. Resolution implies sharpness, though.

f) *Uncertainty* is the variance of the observations in the verification sample and does not depend on the forecasts in any way. The term is applied to observation sets of categorical variables which consist of "1's" if the event occurred and "0's" if it didn't. Its analogue in observations of a continuous variable is the variance. It is usually considered to be related to the "difficulty" of the forecast set. Greater variance implies larger or more frequent changes in the weather element being verified, and when viewed as a time series, these elements are harder to forecast than more persistent weather situations. It is variations in the uncertainty between datasets that make it hazardous to compare verification statistics that are sensitive to uncertainty. For example, variations in the uncertainty of temperature make it inappropriate to compare temperature forecasts from one region to another without compensation for this factor.

Additional examples showing the interaction of reliability and resolution:

a) *low reliability and low resolution*...such forecasts (chance for example) have no skill and are least accurate of all.
b) *high reliability and low resolution*...climatology is an example of such characteristics.
c) *low reliability and high resolution*... an example possessing these attributes would be a POP forecast where for all 30% forecasts, 70% frequency was observed, and for all 90% forecasts, 20% frequency was observed. The forecast has high resolution but applies the wrong label.
d) *high reliability and high resolution*..This is the ideal. Short term persistence is an example of such a forecast.