Ian B. Mason

Integrated verification procedures for forecasts and warnings

Consultancy Report

June 1999

Consultancy brief

Task 1:

Review existing and planned AIFS verification modules:

- maximum and minimum temperature verification (TEMPV);
- quantitative rainfall verification (RAINV);
- fire weather verification;
- TAF verification;
- 7-day forecast verification; and
- verification of model output forecasts (MOF)

and make recommendations for any changes in the theoretical and mathematical basis of those modules.

Task 2:

Review the existing and planned output formats from AIFS verification modules (as listed under task 1) and make recommendations for any new designs, or design changes, of those outputs which will make them easy to interpret by various users of such information,

The two types of output format envisaged are:
Output format type A: Diagnostic output for individual forecasters, forecasting teams, numerical modelers and the developers of Model Output Forecasts (MOF); and
Output Format Type B: Simpler output for weather services users, the media, Bureau management and Government.

Take into account:
information provided by the Bureau of Meteorology Verification Group on the verification output format requirements of various verification users

Task 3

Make recommendations on the design (*theoretical and mathematical basis* and *output format*) of future AIFS modules for the verification of the following types of forecasts:

- qualitative forecasts of precipitation for capital and provincial cities; and
- wind forecasts and warnings.

Task 4

Produce a final report detailing the findings and recommendations resulting from Tasks 1 to 3.

Preface

Integrated Verification Procedures for Forecasts and Warnings, a Consultancy Report by Ian Mason, consists of three parts:

- Summary Report;
- Report on Tasks 1 and 2; and
- Report on Task 3.

Most of the important information is covered in the Summary Report. However, the reader can obtain a better understanding of the full Report by reading selected sections of a general nature from the Report on Tasks 1 and 2 and the Report on Task 3. These selected sections have been collated below to facilitate such reading.

"…

In my opinion, the most important thing for us <methodologists> to do (in the context of events such as the BOM verification workshop), is to provide the operational community with sound and useful approaches to verification problems - and along the way point out the deficiencies in current methods and practices. As long as this community fails to grasp the true nature of verification problems and the potential benefits of a re-oriented approach (including methods, etc.), forecast verification will not be viewed as a high priority or especially useful enterprise. It is not far off the mark (in my opinion) to say that we need a revolution in the arena of forecast verification.

In addition to approaches, methods, ..., another important future consideration is the design and implementation of real-time operational verification systems.

…"

e-mail from Allan Murphy to Ian Mason 22 April 1997

SUMMARY REPORT: CONTENTS

Introduction

This is the third and final report in response to the consultancy brief on integrated verification procedures for forecasts and warnings. The brief is at the front of this report.

Tasks 1 and 2, relating to current AIFS modules and output formats, were covered in the first report. The second report covered task 3, on future AIFS modules for qualitative precipitation forecasts and wind forecasts and warnings.

This report summarises the findings and recommendations of the first two reports, with some brief explanatory text accompanying the listing of recommendations. It also includes as appendices an annotated bibliography of some basic papers on forecast verification, and an edited version of a paper on signal detection theory and ROC analysis presented at the Verification Workshop.

These reports have focussed on theoretical and mathematical issues in forecast verification, as required by the brief.

Summary

Overview

The theoretical basis for these reports is the distributions-oriented approach to forecast verification developed by Murphy and Winkler (1987), and the overarching recommendation is that this approach should be a guiding principle in development of AIFS verification modules.

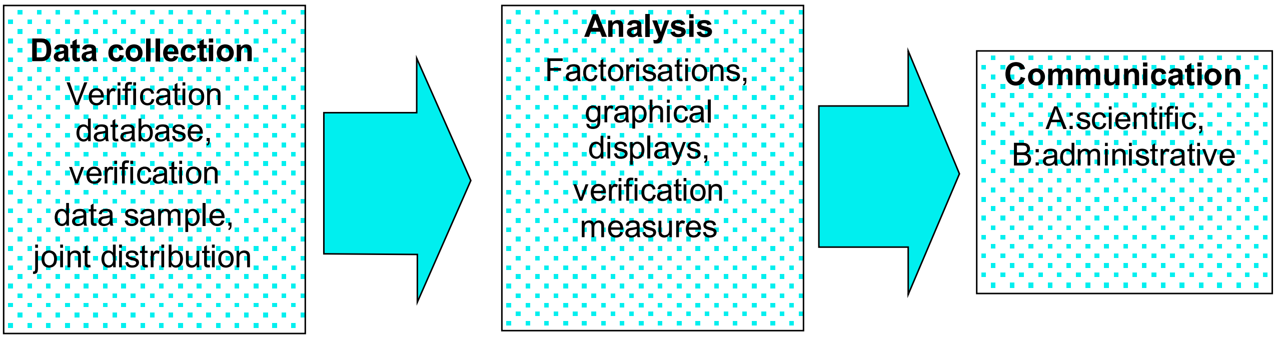
These reports place more emphasis than Murphy and Winkler on initial extraction of the verification data set, since this is often a challenging task in an operational setting.

A listing of the individual recommendations is at sections 2.3 and 2.4 of this report.

A significant addition made in these reports to the suite of distributions-oriented methods is the use of verification measures and methods from signal detection theory. These methods solve some old problems with traditional verification measures, and provide new insights into the nature of forecasting skill. A brief introduction to SDT-based methods is at Appendix 3.2 to this report.

A simple 3-stage structure is used to provide a framework within which the relationships between parts of the verification process can be considered. The stages are data collection, analysis and communication with stakeholders. This process model is discussed in more detail in section 8.1 of the first report. The next page of this report gives a diagrammatic outline.

A structure for verification


Data collection
Verification database, verification data sample, joint distribution
Analysis
Factorisations, graphical displays, verification measures
Communication
A:scientific,
B:administrative

Data collection

Verification database VDB: (forecast f, observed x) pairs plus associated information

Verification data sample: (f,x) pairs from VDB by selection/transformation

Joint distribution: contingency table p(f,x)
contains all the non time dependent information

Analysis

Factorisations: calibration-refinement (CR) p(f,x) = p(f).p(x|f)
likelihood-base rate (LBR) p(f,x) = p(x). p(f|x)

Graphical output: scatterplots, histograms, box plots, conditional histograms, etc

CR factorisation: p(f): predictive distribution histogram
p(x|f): reliability diagram

LBR factorisation p(x): climatological distribution
p(f|x): likelihood diagrams, ROC

Verification measures: means, sds, correlations, MSE and partitions, (d’,β), (Δm,s), Az, etc

Communication

Scientific: forecasters, researchers

Administrative: users/public, political, management

Current AIFS modules: overview

- Current AIFS FVS modules provide a sound basic system and use widely accepted methods of analysis and measures of forecast quality.

- The "diagnostic" or distributions-oriented (DO) framework for forecast verification can be implemented using the present system as a foundation.

- The theoretical framework on which this Report is based is a 3-stage model for the verification process. These stages are

  1. *Data collection*. The output of this stage is the two basic data structures of DO verification, the verification data set and the joint distribution of forecasts and observations.

  2. *Analysis*. In the DO approach this stage involves factoring the joint distribution into marginal and conditional distributions, and calculation of measures of performance and graphical displays of forecast quality.

  3. *Communication*. This stage involves selection of appropriate forms for communication of the results to stakeholders.

- The Report seeks to demonstrate application of the DO approach in each module.

- Measures of forecasts quality for the 2x2 case (POD, FAR, etc) are considered deficient and potentially misleading due to dependence on sample climate and decision threshold. Measures based on signal detection theory (SDT) are more reliable. For non-probabilistic forecasts the measures (d', β) and Az are recommended.

- With regard to specific modules:

  Mean absolute error (MAE) and bias are recommended as basic summary measures of accuracy in temperature forecasting.

  The SDT measures (d', β) or Az are recommended as basic summary measures for rain forecasts in the RAINV format.

  MAE and bias are recommended for numerical fire danger ratings, and SDT indices for ratings collapsed onto the warning threshold.

  A 4x2 joint distribution is proposed for TAF verification, in which INTER and TEMPO appear as separate forecasts. The TAF module should be thoroughly tested on validated data.

  For 7-day forecasts the Priestly skill score and MAE or % of forecasts with lower error than climatology are recommended.

  The variety of forecasts produced by MOF makes a simple recommendation difficult. Overall, MAE and bias are suitable for a "first look" at accuracy.

Current AIFS modules: recommendations

Temperature

Recommendation: that the tabulated values presented for TEMPV be maintained for continuity.

> Some of the performance measures in TEMPV, and other current AIFS modules, are theoretically questionable, but should be maintained at least until there is a general consensus for discarding them. This is for continuity and for comparison with forecasts from other organisations.

Recommendation: That error distributions be provided for signed and absolute errors, and in cumulative form for absolute errors.

> The full error distributions provide information that is not available from summary measures such as mean absolute or mean square error.

Recommendation: That a combination of persistence and climatology be investigated for use as a baseline for the skill score.

> A combination of persistence and climatology may be a more demanding zero for skill than either separately (Murphy 1992, Williams 1997). The best available baseline should be used.

Recommendation: That the scatterplot display include values for slope and intercept of the fitted regression line.

> Good statistical practice requires that fitted models be accompanied by the values of the parameters of the model (and appropriate "goodness of fit" statistics).

Recommendation: That the facility to extract subsets of the verification data sample (VDS) on user-defined thresholds be provided.

> This facilitates ancillary analyses of forecasting skill. For example, accuracy in forecasting large changes, or climatologically rare events, or values near significant thresholds.

Recommendation: That the joint distribution of forecast and observed temperature be provided. for the actual forecasts and for the persistence/climatology baseline forecasts.

> The joint distribution is the starting point for distributions-oriented verification.

Recommendation: That conditional and marginal distributions be available as an option for all temperature forecasts and for the persistence/climatology baseline forecasts.

> Factorisation of the joint distribution into conditional and marginal distributions is a fundamental component of distributions-oriented verification.

Recommendation: That the possible usefulness of box plots, conditional quantile plots, and other appropriate graphic displays for verification data, be investigated.

> These (and other) graphical displays are widely used in statistics to investigate the properties of data sets. Meteorologists concerned with forecast verification could make more use of such displays.

Recommendation: That the measures d',β and Az be available for temperature forecasts collapsed to yes/no forecasts by thresholding on a critical temperature. This temperature should be variable by the user.

> There is often interest in the skill of a forecasting system around specific thresholds, eg frost. The SDT indices are the best measures available of skill in forecasting 0/1 events. d' and Az are measures of discrimination capacity, β indicates the decision threshold.

Recommendation: That scatterplots of observed vs forecast temperatures should be available for all guidance forecasts, and for the official forecasts, with statistical parameters of the fitted line and values of MAE for each.

> Scatterplots are good basic data displays. Where a linear model is plausible it should be provided with relevant parameter values.

Recommendation: That verification information as detailed above for individual forecasters be available at logon at the start of each forecaster's shift.

Information on performance is important feedback for individuals, and needs to be frequent and quick to be most effective.

Recommendation: That MAE be adopted as a single basic measure of temperature forecast accuracy for public information throughout Australia, and be referred to as "average accuracy".

Mean absolute error (MAE) is simple and easily explained to a non-expert. Root mean square error may be more satisfying from some theoretical perspectives, but is not as easy to understand. "Average accuracy" seems more positive than "average error".

Rain

Recommendation: That the current verification measures used in the AIFS rain FVS be maintained for continuity.

As with TEMPV, some of the performance measures currently provided are theoretically questionable but should be maintained until they fall into disuse. This is for continuity and for comparison with forecasts from other organisations.

Recommendation: That the full CR and LBR factorisations for the 8x8 and 2x2 JDs be available as an option.

Output equivalent to the joint distribution is already available in the AIFS FVS. The calibration-refinement (CR) and likelihood-base rate (LBR) factorisations are fundamental to distributions-oriented verification.

Recommendation: That the SDT-based indices d', β and Az be calculated for rain forecasts as outlined.

SDT provides the best measures available at present for pure skill in forecasting a 0/1 event. d' and Az are measures of discrimination, β indicates decision threshold.

Recommendation: That values of d' and β for individual forecasters most recent adequately large sample of forecasts be available at logon, together with the outcome (forecast and observed categories) of their previous forecast.

Feedback on performance is important for learning, and needs to be frequent and rapid to be most effective.

Recommendation: That the form for RAINV forecasts be incorporated into operational AIFS forms for standard OWR issue times, to reduce the risk that they will be overlooked or done after the start of the validity period of the forecast.

Variability in the operational setting of the forecasts should be minimised.

Fire Weather

Recommendation: That use of transformations to normality and box plots be investigated to enhance the clarity of graphical displays of skewed data such as fire danger ratings.

Many basic statistical procedures assume normally distributed data. If the data is not normal, methods established in statistics to handle this situation should be used.

Recommendation: That where a linear model is fitted to data, the parameters of the model and other statistical information as above be provided.

Good statistical practice requires that fitted models be accompanied by the values of the parameters of the model (and appropriate “goodness of fit” statistics).

Recommendation: That the capability to select subsets of the VDS be provided.

This is to facilitate detailed analysis of skill, for example near thresholds for issue of advices and warnings.

Recommendation: That an option to download the VDS to standard spreadsheet programs be provided in the AIFS fire weather verification system.

This is to facilitate use of statistical software that may not be available in AIFS.

Recommendation: That the primary measure of skill for wind speed and direction and fire danger ratings be in the form of a table of values of d’ and β for successive category boundaries on the joint distribution.

The SDT measures d’ and β provide the best succinct description currently available of forecasting performance for 0/1 events.

TAF verification

Recommendation: That values for CSI be calculated for the 2x2 JD.

“Critical Success Index” (CSI) is widely used as a performance measure, and is likely to be required for comparison with forecasts produced by other organisations. Some of the performance measures currently in use, including CSI, are theoretically questionable, but should be maintained at least until there is a general consensus for discarding them.

Recommendation: That values of the scores presented in AIFS should be accompanied by confidence intervals using the methods discussed in Seaman et al (1996).

Significance testing is rarely undertaken in forecast verification, and the issue has not been pushed in this Report. Nevertheless, where feasible it should be done, and particularly where small differences between competing forecasting systems may be important.

Recommendation: That the AIFS system be exhaustively tested on validated data sets before use as the Bureau’s official system for TAF verification.

Statements on forecast accuracy are likely to be closely scrutinised by both customers and potential competitors. The system must be well-tested and robust.

Recommendation: That a system for verifying TTF and Code Grey forecasts be developed.

These forecasts have significant commercial implications for customers.

Recommendation: That the possibility of double counting in combining contingency tables be investigated and eliminated if found.

The present AIFS TAF verification system is somewhat opaque in some details. The possibility of double counting arose in examination of some samples of output.

Recommendation: That verification results be available for the validity period of the TAF in 3-hour blocks as a routine option (in addition to the currently proposed variable forecast age).

Shorter period and longer periods of the TAFs are used differently by the industry. The average deterioration in skill with time is likely to be of interest.

Recommendation: That the next revision of the TAF verification module includes verification of temperature and QNH, and consideration be given to real time verification of at least these quantities.

On the general principle that all forecasts should be verified. The skill of temperature and QNH forecasts will be of interest when objective guidance becomes available for the TAF.

Recommendation: That the reliability of PROB30 and PROB40 forecasts be assessed by extracting the relative frequencies of corresponding forecast events.

The actual probability of events given specific forecasts is important for operational decision-making.

Recommendation: That the raw verification data set be available as an option, for both the actual and persistence forecasts, and for subsets to be selectable on user-defined criteria.

This is mainly to facilitate examination of particular cases, and forecasting performance when forecast or observed weather is near significant thresholds (alternate minima).

Recommendation: That contingency tables include row and column totals, plus the joint relative frequencies.

To save manual calculations.

Recommendation: That 4 (forecasts) x 2 (observation) contingency tables as described in the text be available as an option.

To assess the overall operational implications of INTER and TEMPO the forecasts are categorised into four classes: No alt, INTER, TEMPO and alt, and the observations determined following Ross Keith's algorithm as in the present AIFS TAF module.

Recommendation: That the components of the CR and LBR factorisations be available as an option for both the actual forecasts and persistence, in both the 2x2 and 4x2 forms of the joint distribution.

The CR and LBR factorisations are fundamental to distributions-oriented verification.

Recommendation: That values of d', β and Az be calculated for all 2x2 forecast observed contingency tables and presented with estimated confidence intervals, and that Az be calculated for the 4x2 JD.

The SDT measures d' or Az are the best available for discrimination capacity. Az is appropriate when forecasts are made at a range of confidence levels (the 4x2 JD).

7-day forecasts

Recommendation: That the Priestly skill score be used as a summary measure of skill for 7-day outlooks. Smoothed daily means should be used as the climatology.

The Priestly skill score provides a figure for % improvement over climatology, and thus a better indication than MAE of the lead time at which the forecasts add no value to climatology.

Recommendation: That (d', β) be used as summary indicators of skill for 7-day rain forecasts prepared as yes/no forecasts. For probabilistic forecasts Az should be used.

The SDT measures d' and β are the best available verification measures for yes/no forecasts. Az is better for probabilistic forecasts.

MOF

Recommendation: That MOF be verified in a separated AIFS FVS module and summary measure of accuracy for the previous 30 issues be available for each operational issue.

It has been observed that biases can develop in MOF-type forecasts, particularly at the extremes of the ENSO cycle. It is important that operational forecasters know and allow for these biases.

Recommendation: That MAE and bias be used as summary measures of skill for MOF temperature forecasts.

MAE is a good summary measure performance measure for temperature forecasts. Knowledge of bias is operationally important.

Recommendation: That (d', β) for the rain/no rain threshold be used as a summary indicator of skill for MOF rain forecasts.

The SDT measures d' and β are the best available performance measures for yes/no forecasts.

Recommendation: That MAE and bias for wind speed be used as summary indicators of skill for MOF wind forecasts.

MAE is a good summary performance measure for continuous predictands. Knowledge of bias is operationally important

Recommendation: That MAE and bias be used as summary measures of performance for the MOF forecasts listed in the text.

See above.

Proposed AIFS modules

Recommendations in this section are more detailed than those of the first report, as the AIFS modules for these products have not as yet been produced.

In outline, the 3-stage verification process is recommended using distributions-oriented methods in both modules. Methods for evaluation of probability forecasts for qualitative predictands are well developed. For wind forecasts and warnings the main problems are in development of a verification data set with adequate time and space resolution to capture the main features, and hourly or shorter sampling of (forecast, observation) pairs is recommended.

Qualitative precipitation forecasts

A verification system for probabilistic precipitation forecasts should have the following components:

Data

A comprehensive basic verification data set should be available, containing in addition to the sequence of matched pairs of forecasts and observations enough additional information to fully characterise the forecasting situation.

The basic data set should be flexible, so that subsidiary data sets can be derived from it by selection of cases on user-defined criteria or by transforming variables

Analysis

The joint distribution of forecasts and observations should be available, together with the CR and LBR factorisations, since these are the basic data structures for DO verification. The means and variances corresponding to each of these factorisations should be available, and the correlation coefficient.

Analyses and summary measures should include

- MSE in the raw form and components of the partitions described in the text
- The skill score SS based on MSE using both long-run and sample climate as baselines for skill. Use of the optimal combination of climatology and persistence as a no-skill baseline should be investigated. Components of the partition of SS should be available.
- SDT-based measures (Δm,s) and Az, together with variances.
- The sharpness diagram (p(f) histogram), reliability diagram and likelihood diagram.
- The ROC on both linear probability and “binormal” axes.

Communication

Category A

Two levels of output should be provided for category A users. These are

1. Comprehensive output to provide a complete description of all aspects of the forecasts. This should comprise
    - listings of the basic verification data set and any derived data sets,
    - the joint distribution with CR and LBR factorisations,
    - graphical output of p(f), the reliability diagram, likelihood diagrams and the ROC in linear probability and bi-normal form
    - the Brier score and skill score together with components of the various partitions
    - (Δm,s) for the SDT model with estimated variances, and

- Az and estimated variance.

2. Output suitable for daily feedback on performance to forecasters and forecasting teams. The skill score based on squared error is recommended for this purpose, ie

$$SS = 1 - (f-x)^2/(c-x)^2$$

where f is the forecast probability for rain, $x \in \{0,1\}$ is the observation and c is the corresponding long-run climatological probability of rain. SS can be expressed as percent improvement over climatology.

Investigation of the use of an optimal combination of climatology and persistence is recommended.

At regular intervals, as sufficient data accumulates, reliability diagrams should be available. How much data is sufficient is difficult to define, but about 100 forecasts is probably the minimum.

Category B

For category B users there are two main features of probabilistic forecasts that require description. These are calibration and skill or discrimination.

The REL component of the CR partition of MSE is most often used to describe calibration, although it does not provide an indication of whether the forecasts are over- or under-confident. In fact there is currently no completely satisfactory summary measure of calibration, and at least for management it is desirable also to show the calibration diagram with a brief description in lay terms.

The summary measure of skill most often encountered is the skill score SS based on MSE, expressed as % improvement over climatology. It can be simply explained as percent improvement in the accuracy of the forecasts over unchanging forecasts of the average probability of rain. Az is more theoretically satisfying as a measure of discrimination, but is more difficult to explain to non-specialists.

Output recommended for Category B users (weather services users, the media, Bureau management and Government) is

- The REL component of the partition of the MSE skill score based on the CR factorisation.
- Reliability diagrams should also be provided to management, and possibly to other B users at the discretion of management. The reason for this caution is that the sample relative frequencies plotted in reliability diagrams are unstable for small samples, and can be misleading for non-experts.
- The MSE-based skill score SS using long-run climatology as the baseline.

Wind forecasts and warnings

Data

The main recommendation relating to verification of wind forecasts and warnings is that the basic verification data set should include hourly values of forecast and observed wind speed and direction, plus values at any intermediate times at which either forecasts were issued or wind speed passed through one of the warning thresholds.

The basic VDS should also include whatever additional information is necessary to adequately describe the situation, including for warnings the facility to link to text files containing comments or non-standard observation that can be entered by RFC staff in real time. These kinds of data are sometimes lost or forgotten in the period between the warning situation and verification.

It should be possible to extract subsidiary VDS by selection on the basis of user-supplied criteria, or by transformation of variables.

Separate joint distributions should be available for direction, speed and warnings together with the corresponding CR and LBR factorisations.

Analysis

As a summary performance measure for wind direction the correlation coefficient is recommended.

For wind speed it is recommended that the 7x7 JD be collapsed into a sequence of six 2x2 JDs by thresholding on successive category boundaries, and the SDT measures (d', β) be calculated for each threshold. The correlation coefficient should also be provided.

For the 2x2 warning JD d' and β are recommended as summary performance measures.

Values for the more familiar “scores”, POD, FAR, bias, POFD, CSI, Hansen & Kuipers' score, the Heidke score, and proportion correct should also be provided.

Values for MSE with components of the basic, CR and LBR factorisations should also be available.

The skill score based on MSE should be available, using both long-run climatology and sample climate as the no-skill baseline, and use of the optimal combination of climatology and persistence should be investigated.

Graphical displays should include histograms for marginal distributions of both forecasts and observations with descriptive statistics (μ_f, μ_x, σ_f, σ_x, ρ_{fx}), bivariate histograms (eg Murphy 1997) and box plots for wind direction and speed. The box plots should be based on the raw observations, rather than categorised.

Communication

Category A

Output recommended for Category A users has two levels.

As a “first look” option, the joint distributions for direction, speed and warnings should be provided, together with descriptive statistics (μ_f, μ_x, σ_f, σ_x, ρ_{fx}) and values for the skill score standardised against long-run climate.

The second level includes the full DO verification and all statistics, graphs and verification measures.

Category B

As a single summary statistic for wind forecasts for B users, the MSE skill score using long-run climate should be provided, for both direction and speed, referred to as “improvement over climatology”, with climatology represented by an unchanging forecast of average conditions.

For warnings, presentation of the 2x2 JD alone may be satisfactory, with cells indicating numbers of occasions rather than relative frequencies. If a single number is required, d’ is recommended. It could be referred to simply as a measure of forecasting skill, with the information that zero represents forecasts with no skill and values above about 4.0 indicate almost perfect performance.

Appendices

Some basic documents: an annotated bibliography

The following papers and textbooks are one view of some interesting reasonably recent documents in this field, focussed largely on distributions-oriented methods and SDT since these are still new to most meteorologists. It is not intended to be comprehensive, but would be a basic collection and provide a lead into a fairly large literature.

A more exhaustive bibliography on forecast verification by Harold Brooks is on

http://www.nssl.noaa.gov/~brooks/verification/bibliography.html

A bibliography on SDT methods in non-meteorological fields is on

http://www.urmc.rochester.edu/smd/biostat/roc.html

Textbooks

Wilks, Daniel S. 1995. Statistical methods in the atmospheric sciences : an introduction. *Academic Press*, pp 467.

The currently favoured textbook on statistical methods in meteorology. Rather more space devoted to time series methods than they are really worth in meteorology, and little or nothing on SDT, but still an essential reference.

Katz, Richard W. and Allan H. Murphy, editors, 1997. *The Economic Value of Weather and Climate Forecasts*. Cambridge University Press, Cambridge.

The chapter on forecast verification by Murphy is alone worth the price of admission. The most comprehensive presentation of distributions-oriented methods available at present.

Murphy, A.H. and Richard W. Katz, editors, 1985. *Probability, Statistics and Decision Making in the Atmospheric Sciences*. Westview Press, Boulder and London.

Surveys of some developments in probability and statistics in the early 1980s that the editors thought might be of interest to atmospheric scientists, and reviews of some areas within the atmospheric sciences in which probability and statistics play a major role. Predates development of distributions-oriented methods, but still a valuable reference.

Particularly useful chapters on exploratory data analysis by Graedel and Kleiner, Bayesian inference by Winkler, and decision analysis by Winkler and Murphy. The section on forecast evaluation by Murphy and Daan is a good reference for pre-DO verification but a difficult notation makes it slow reading.

Stanski, H.R., L.J. Wilson and W.R. Burroughs, 1989. *Survey of common verification methods in meteorology*. Research Report No. 89-5, 114pp, Toronto: Canadian Atmospheric Environment Service.

Useful and frequently cited reference material. Appears to be a set of edited lecture notes.

Green, D.M. and J.A. Swets, 1966. *Signal Detection Theory and Psychophysics*. Reprinted 1974 Robert E. Kreiger New York. 479pp.

Classic text on SDT by two of the founding fathers. A valuable reference for basic mathematical and statistical results.

Swets, J.A., and R.M. Pickett 1982. Evaluation of diagnostic systems: methods from signal detection theory. *Academic Press, New York*.

More practically oriented than Green and Swets and includes examples from a variety of fields including weather forecasting. Has a listing of a FORTRAN program to fit ROC by maximum likelihood and provide variances of parameter estimates (a reasonably user-friendly working implementation of which can be downloaded from

http://www-radiology.uchicago.edu/krl/toppage11.htm .).

Largely devoted to designed testing of diagnostic accuracy in a medical setting.

Swets, John A. 1996. Signal detection theory and ROC analysis in psychology and diagnostics: collected papers. Lawrence Erlbaum Associates Inc, pp 308.

Highly recommended. If you can only get one book on SDT/ROC methods this is it. A collection of papers which include references to applications in meteorology and other fields, and analysis of some performance measures used in meteorology. Not quick or easy reading, but worth the effort.

Winterfeldt, Detlof von, and W. Edwards 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press, 604pp.

Contains as chapter 5 a classic paper on Bayesian statistics, originally by Edwards, Lindeman and Savage. The rest of this text is background relevant to studies of the value of weather forecasts in decision-making, and includes some interesting comments on SDT.

Distributions-oriented methods

Murphy, A. H. and Winkler, R.L., 1987. A general framework for forecast verification. *Monthly Weather Review, 115*, 1330-1338.

Essential reading for meteorologists concerned with forecast verification. The original presentation of the diagnostic or distributions-oriented approach. Murphy and Winkler state in the introduction…

“A need exists for a general framework for forecast verification. To be useful such a framework should (*inter alia*) (i) unify and impost some structure on the overall body of verification methodology, (ii) provide insights into the relationships among verification measures, and (iii) create a sound scientific basis for developing and/or choosing particular verification measures in specific contexts. Moreover, such a framework should minimize the number of distinct situations that must be considered. The primary purpose of this paper is to describe a framework that appears to meet many of these goals”.

Murphy, A. H. 1995. A coherent method of stratification within a general framework for forecast verification. *Monthly Weather Review, 123*, 1582-1588.

Extends the “general framework” to include stratification using covariates. This is the theoretical foundation for investigating the dependence of forecasting skill on factors such as synoptic weather type, identity of forecaster, ENSO regime etc.

Murphy, A. H. 1991. Forecast verification: its complexity and dimensionality. *Monthly Weather Review, 119*, 1590-1601.

> Important paper defining complexity and dimensionality in the context of forecast verification. The basic reason why distributions-oriented methods are essential. Murphy states…
>
> "Failure to take account of the complexity and dimensionality of verification problems may lead to an incomplete and inefficient body of verification methodology and, thereby, to erroneous conclusions regarding the absolute and relative quality and/or value of forecasting systems".

Murphy, A.H. 1996. Forecast verification: a diagnostic approach. in *Proceedings of Workshop on the Evaluation of Space Weather Forecasts*. Boulder, Colorado; 19-21 June 1996.

> One of the best reasonably succinct presentations of the DO approach.

Murphy, A. H. 1996. General decompositions of MSE-based skill scores: measures of some basic aspects of forecast quality. *Monthly Weather Review, 124*, 2353-2369.

> Presents three kinds of partition of MSE and the skill score based on MSE, using the basic factorisations of the distributions-oriented verification framework. Necessary background for any use of MSE as a performance measure (but has an error in comments on SDT-based methods).

Applications of distributions-oriented methods

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

> The first paper to apply distributions-oriented methods to a specific verification problem. The methods are applicable to forecasts of any quasi-continuous variable. Some interesting graphical displays (bivariate histograms, conditional quantile plots).

Brooks, H.E. and Charles A. Doswell III, 1996. A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Weather and Forecasting, 11*, 288-303.

> Very useful demonstration of an application of distributions-oriented methods.
>
> Verification of some maximum temperature forecasts for Oklahoma City used to compare conventional ("measures-oriented") with distributions-oriented methods. The authors state that they show "the vast wealth of additional information that can be obtained through a distributions-oriented verification over a "traditional" measures-oriented approach".
>
> Some interesting criticisms of the US NWS verification effort.

Murphy, A. H. and Winkler, R.L., 1992. Diagnostic verification of probability forecasts. *International Journal of Forecasting, 7*, 435-455.

> A model application of distributions-oriented methods to some US probability forecasts. Useful reference for partitions of the Brier score, and the attributes diagram (an elaboration of the reliability diagram). A blind spot in the direction of ROC/SDT methods.

Brown, Barbara G., Gregory Thompson, Roelof T. Bruintjes, Randy Bullock and Tressa Kane 1997. Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *Weather and Forecasting, 12*, 890-914.

Application of distributions-oriented verification to forecasts of aircraft icing. Good examples of box plots.

Murphy, A.H. 1996. The Finley affair: a signal event in the history of forecast verification. *Weather and Forecasting, 11*, 3-20.

Mostly a historical review of the development of scores for yes/no forecasts in response to publication of Finley's tornado forecasts in 1884. Verification measures proposed then by Finley himself, Peirce, Doolittle, and Gilbert are still in use. This paper is a fascinating exploration of theoretical issues raised by these scores, first discussed over 100 years ago and still relevant.

Methods from signal detection theory

Mason, I.B. 1982. A model for assessment of weather forecasts. *Australian Meteorological Magazine, 30*, 291-303.

The first paper in the refereed literature on application of ROC/SDT methods to weather forecasts. Shows ROCs for forecasts of a number of different weather events and demonstrates that the SDT model with Gaussian distributions provides a good model for the data.

Mason, I.B. 1982b. On scores for yes/no forecasts. *Preprints of papers delivered at the Ninth AMS Conference on Weather Forecasting and Analysis*, Seattle, Washington, 169-174.

Tried to show that all common scores for yes/no forecasts are problematic because they imply unrealistic ROCs. See Swets 1986 below.

Levi, Keith 1985. A signal detection framework for the evaluation of probabilistic forecasts. *Organizational behaviour and human decision processes, 36*, 143-166.

Appears to be the second refereed paper on application of SDT to weather forecasts. Interesting discussion of implications for decision-making.

The following three papers by John Swets are in the collection referred to under textbooks.

Swets, J.A. 1973. The relative operating characteristic in psychology. *Science, 182*, 990-1000.

A review of the development of ROC/SDT methods in psychology and engineering. Good background.

Swets, John A. 1988. Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293.

Shows that the ability to discriminate between two alternatives in many fields (all so far studied) can best be measured by ROC analysis and that a good performance measure is Az.

This paper concludes…

"…that the fundamental factors in accuracy testing are the same across diagnostic fields and that a successful science of accuracy testing exists. Instead of making isolated attempts to develop methods of testing for their own fields, evaluators could adapt the proven methods to specific purposes and contribute mutually to their general refinement".

Swets, John A. 1986. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin, 99*, 100-117.

Uses the ROC to evaluate some common performance measures, including proportion correct, Hansen and Kuipers' score, Frank Woodcock's Z, Yule's Q and some others. Shows that all imply a SDT model with some form of underlying distributions, and that those not consistent with a "regular" ROC, that is, most, are likely to be misleading.

The paper that Mason would have written if he could, rather than 1982b above.

Harvey, Lewis O. Jr, Kenneth R. Hammond, Cynthia M. Lusk, and Ernest F. Mross 1992The application of signal detection theory to weather forecasting behaviour. *Monthly Weather Review, 120*, 863-883.

Interesting paper by a group of psychologists at University of Colorado viewing weather forecasting as an example of human judgement under uncertainty.

Criticises the POD/FAR/CSI (etc) indices, and proposes SDT as "an appropriate model of forecasting behaviour (which) achieves the three goals stated by Murphy and Winkler, is compatible with joint probability decomposition and provides a simpler description of forecasting behaviour".

Includes an interesting analysis of some research data to learn whether stress affects forecast accuracy, decision criteria, or both. Accuracy was better under high stress conditions and stress caused a lowering of decision criteria.

Some applications of SDT/ROC methods in meteorology

McCoy, Mary Cairns 1986. Severe-storm-forecast results from the PROFS 1983 forecast experiment. *Bulletin American Meteorological Society*, *67*, 155-164.

Application of SDT to assessment of skill in severe storm forecasts. Good presentation of SDT.

Buizza, R., T. Petroliagis, T. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi 1998. Impact of model resolution and ensemble size on the performance of an Ensemble Prediction System. *Quarterly Journal of the Royal Meteorological Society, 124*, 1935-1960.

Uses the ROC to evaluate ensemble forecasts by ECMWF. (Presentation of SDT follows Stanski et al 1989, above)

Hamill, T.M., C. Snyder and R.E. Morss 1999. A comparison of probabilistic forecasts from bred, singular vector, and perturbed observation ensembles. Submitted to Monthly Weather Review February 1999, downloaded from

Uses ROC to compare wind forecasts produced by different ensemble methods at NCAR.

The weather forecast as a statistical decision: the signal detection model in verification

Methods from signal detection theory (SDT) have been recommended in these reports as providing the best means currently available to analyse the skill of forecasts for weather events.

This section is edited text of a paper presented at the Verification Workshop in November 1997. It was intended as a reasonably accessible introduction to the basic ideas and methods.

Introduction

This is an outline of an approach to forecast verification using methods from the mathematical theory of detection of signals in noise (SDT, for signal detection theory), and the use of the relative operating characteristic (ROC).

Methods based on SDT and the ROC are a useful addition to techniques currently used in forecast verification. They facilitate the assessment of pure forecasting skill, as distinct from an ability to communicate with users of forecasts (reliability or calibration). They also provide measures of skill that can be calculated for forecasts issued in any form, enabling valid comparisons between the skill of simple yes/no forecasts and those issued as risk ratings or probabilities, or in any other way.

ROC/SDT methods are compatible with and complement the distributions-oriented approach developed by Murphy and Winkler (1987).

These methods also provide criteria for good measures of forecasting skill and reveal deficiencies in current measures (Mason 1982b, 1989; Swets, 1986).

The methods described were originally developed by psychologists studying human sensory discrimination, including the ability of human observers to detect specific signals on radar screens in a military context, and also drew on statistical decision theory and detection of electromagnetic signals in noise. Swets (1973) gives a detailed account of the historical development of the field to that time, updated in 1988 (Swets, 1988). In a meteorological context, Mason (1980, 1982a,b, 1989) has discussed the applicability of the methods to forecast verification. Other presentations for meteorologists include Levi (1985), Harvey et al (1992) and Buizza et al (1999). Some good basic texts are Egan (1975), Swets and Pickett (1982), Macmillan and Creelman (1991) and Swets (1996).

This presentation is rather more discursive and heuristic than it would be in a formal setting, as the ideas are still unfamiliar to most meteorologists.

The signal detection model

There are many situations of practical importance in which it is necessary to decide among alternative courses of action on the basis of information which does not provide absolute certainty. A meteorologist deciding whether to forecast rain on the basis of the usual synoptic data is usually an example of such a situation. Another is the manager of a weather-sensitive business deciding on the basis of a weather forecast whether to take precautions against adverse weather. Others, among very many, include medical professionals making a diagnosis, engineers seeking metal fatigue in aircraft, police using polygraph lie detectors, and research scientists seeking to ascertain whether a particular experimental manipulation did or did not have a significant effect.

A common feature of these and many other formally similar situations is that the available information provides only a certain “weight of evidence” for occurrence of an event of interest. This weight of evidence varies from one occasion to the next and is usually not sufficient for certainty. A decision is made by comparing the current weight of evidence with a pre-determined decision threshold. The system (forecaster, doctor, engineer, etc) asserts a positive result when the evidence exceeds the threshold, and negative when it is less. Figure 1 illustrates this situation.

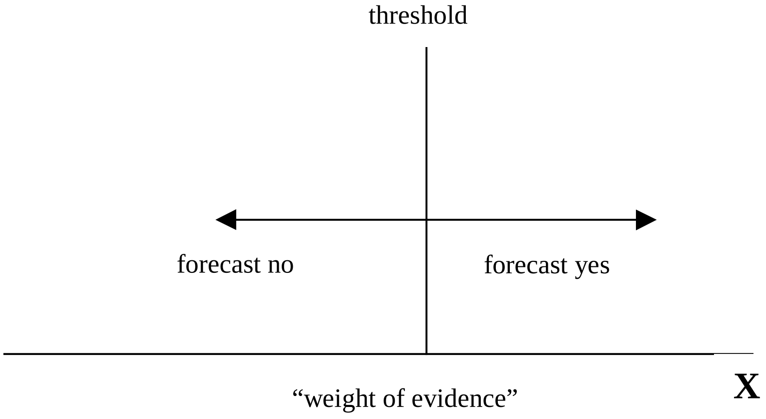


Figure 1: *Forecasting a two-state event under uncertainty*

There is a similarity to statistical hypothesis testing. The weight of evidence for a hypothesis is represented by a function of the data, usually something like Student's t or chi squared. Above a critical value a null hypothesis is rejected, and below it, accepted.

The model illustrated in Fig.1 is developed further by supposing that the weight of evidence, which is assumed to be represented by a scalar quantity X, has a certain fixed and known probability density when the event of interest does not occur, denoted $f_0(x)$, and a different distribution $f_1(x)$ when the event does occur, as shown in figure 2.

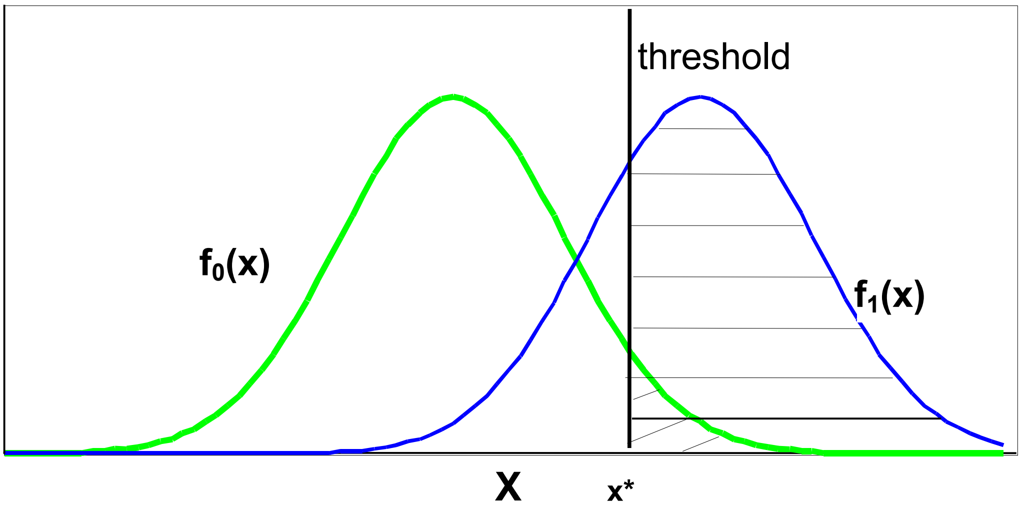


Figure 2: *Probability distributions for the "weight of evidence" X prior to non-occurrence, $f_0(x)$, and prior to occurrence, $f_1(x)$, of the event to be forecast. The diagonally hatched area represents the probability of a forecast of occurrence given that the event does not occur, and the horizontally hatched area the probability of a forecast of occurrence given that the event occurs.*

The decision threshold is indicated by x*. The form of the distributions does not need to be specified at this stage.

We are in the familiar territory of statistical hypothesis testing, where $f_0(x)$ would be the distribution of a test statistic under the null hypothesis and $f_1(x)$ the distribution of the statistic under a (simple) alternative hypothesis. When the model is applied to signal detection, $f_0(x)$ represents the distribution of incoming data when noise alone is present, and $f_1(x)$ the distribution when a signal is present in addition to noise. In the case of weather forecasting $f_0(x)$ is the distribution of the weight of evidence for a weather event prior to non-occurrence, and $f_1(x)$ prior to occurrence. The event is forecast when the weight of evidence is greater than x*.

X could be something like the Total Totals index in the case of thunderstorms, where x* would be a value of (say) 45. In general is simply a scalar variable monotonically related to the likelihood of the event. In the case of “subjective” forecasts X is a representation of the forecaster’s judgement of the degree to which the evidence favours occurrence rather than non-occurrence of an event.

Given the distributions $f_0(x)$ and $f_1(x)$, the location of the decision threshold x* determines some interesting probabilities. The area under $f_1(x)$ to the right of x* represents the probability of a forecast of occurrence given that the event does occur, or a hit or true positive (TP). This area is equivalent to Probability of Detection, POD. The area under $f_0(x)$ to the right of x* represents the conditional probability of a forecast of occurrence given that the event does not occur, ie the probability of a false alarm, sometimes referred to as a false positive (FP). This area is equivalent to Probability of False Detection (POFD). The other two areas, under $f_0(x)$ and $f_1(x)$ to the left of x*, are the probabilities of true negatives (TN) and of false negatives (FN, or misses), respectively. POFD is analogous to the probability of a type 1 error, and POD to the power of a statistical test, or 1 – the probability of a type 2 error.

The probability of each of these four combinations of forecast and event (TN, FN, TP, FP) changes as x* moves along the weight of evidence axis. When the decision threshold is low, ie x* is toward the left end of the X continuum, then the event will be forecast relatively often, practically all of the occurrences will be correctly forecast (POD near 1.0) but there will be many false positives (POFD also approaches 1.0). Just how many false positives is determined by x* and the nature and relative separation of the distributions. Conversely, when x* is towards the right end of the X axis, POFD will be low but so will POD; there will be many misses.

(The forecasts have been assumed to be unequivocal assertions that the event will or will not occur. In the case of forecasts issued as probabilities there are K-1 thresholds on the “weight of evidence” axis, corresponding to the K allowed values for the probabilities (eg 0, 2%, 5%, 10%, ... 100%).)

X is related to the probability of the event. Given the distributional form of $f_0(x)$ and $f_1(x)$, and the climatological (prior) probability of the event, p_C, a numerical value for the threshold x* can be converted into a threshold probability p* by Bayes’ rule:

$$p^* = \Pr\{\text{event}|X=x^*\} = R/(1+R) \qquad (1)$$

where $R = [p_C/(1-p_C)][f_1(x^*)/f_0(x^*)]$

There is an optimal location for x* or p* for decision-making which maximises the expected value of the forecasts and which is determined by the relative benefits and costs of the four possible outcomes referred to above. In the well known cost-loss decision model (Thompson & Brier, 1955; Murphy, 1977), the optimal p* is equal to C/L, where C is the cost of precautions against the event, and L is the loss if the event occurs and no precautions have been taken.

Calculation of the parameters of the SDT model for yes/no forecasts

If a single set of verified yes/no forecasts is available then, subject to assumptions about the form of f_0 and f_1, the separation of the means of these distributions can be calculated, and also the implied location of the decision threshold.

We assume that f_0 and f_1 are Gaussian in form with equal variances and means separated by d’. (anticipating a little, the assumption that the distributions are Gaussian, or at least Gaussian to within a monotonic transformation, is well supported by data. The implied variances are not usually exactly equal, but this is disregarded for the present). These assumptions make it possible to make the x scale quantitative, and derive a specific value for the threshold x*, simply by looking up POD and POFD in tables of the standard normal deviate. For example, suppose the verification array is

		event		
		no	yes	
forecast	no	247	8	
	yes	49	66	
				370

Then POD = 66/(8+66) = 0.892 and POFD = 49/(247+49) = 0.166

The standard normal deviates corresponding to these taken as areas under f_1 & f_0 can be found from any table of areas under the standard normal curve or using the NORMAL function in common spreadsheets. They are -1.237 and +0.970 respectively, in units of the common standard deviation of f_1 and f_0.

Thus the decision threshold is 1.237 units to the left of the mean of f_1 and 0.970 units to the right of the mean of f_0 Hence the separation of the means is 2.207 Figure 3 illustrates what is going on.

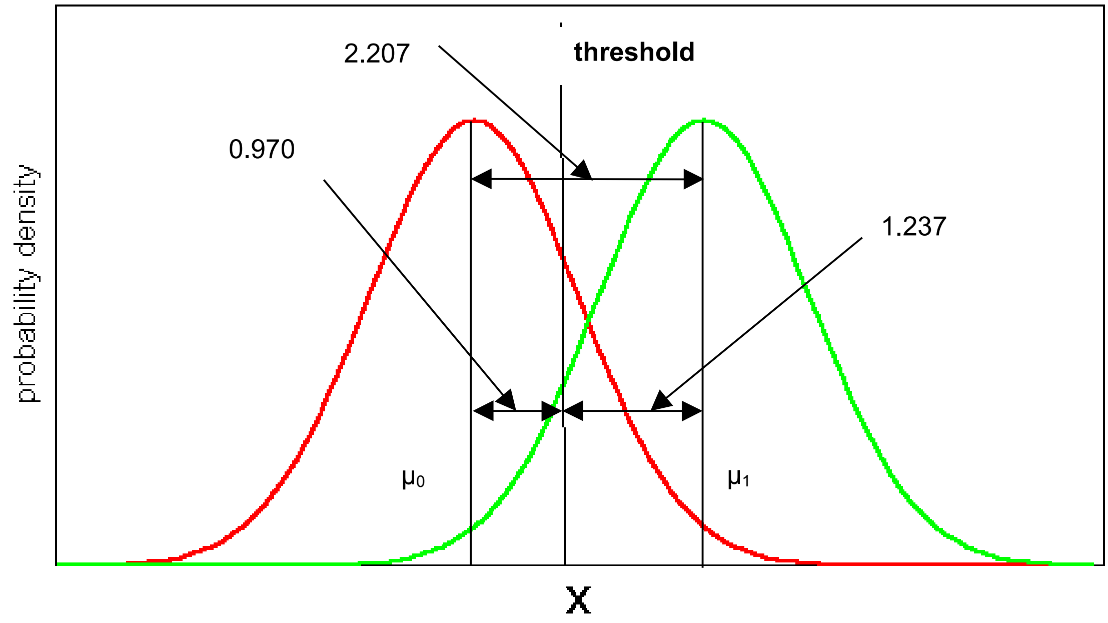


Figure 3: *Calculation of d'. $f_0(x)$ is the distribution of X prior to non-occurrences with mean μ_0 and $f_1(x)$ prior to occurrences with mean μ_1. The common standard deviation is 1.0. The implied decision threshold is 1.237 standard deviations to the left of μ_1 and 0.970 standard deviations to the right of μ_0, so the separation of the means, d', is 2.207.*

The separation of the means, conventionally denoted d' when the variances are equal, is used as an index of the intrinsic discrimination capacity of a forecasting system. If d' is small then POD and POFD are only slightly different, ie a forecast of occurrence is only slightly more likely before an occurrence than it is before a nonoccurrence. If d' is large then POD»POFD and the system is showing a high level of discrimination capacity.

The location of the threshold, x*, is usually indexed by a likelihood ratio β=f1(x*)/f0(x*), the ratio of the ordinates of the distributions at x*. β can be expressed as a threshold probability p* via the odds form of Bayes' formula, which gives

$$p^* = \Pr\{event|X=x^*\} = R/(1+R)$$

where

$$R = \omega_O \cdot \beta$$

and

$$\omega_O = \Pr[E=1]/(1-\Pr[E=1])$$

(the prior odds on an occurrence).

It is interesting to note that all scores expressible in terms of the elements of the 2x2 verification array can be expressed in terms of POD, POFD and pr[E=1] = p_c, the (sample) climatological probability of the event.

The relative operating characteristic

A single set of yes/no forecasts is not sufficient to determine the performance of a forecasting system for all thresholds, due to the need to assume equality of the standard deviations of the underlying distributions. An adequate description of performance requires specification of the ratio of these standard deviations, which in turn requires knowledge of the performance of the system at different thresholds.

The overall performance of a forecasting system for any threshold is determined by the nature and parameters of f_0 and f_1, and can be described empirically for real forecasts by graphing the variation of POD with POFD as x*, or p*, varies, using forecasts for discrete events (eg rain/no rain) issued as ratings of risk or probabilities. The threshold probability is stepped through the range of forecast probabilities used, and values for POD and POFD calculated at each step. Table 1 shows the process, for some rain forecasts for Canberra.

Table 1: *Calculation of POD and POFD as functions of threshold probability, p*, for some forecasts of rain in Canberra. N1 is the number of forecasts of the corresponding probability followed by occurrences of rain, N0 is the number followed by no rain. N(N1≥p*) is the number of forecasts of rain probability greater than or equal to the corresponding forecast followed by rain. N(N0≥p*) is the corresponding quantity for forecasts followed by no rain. POD is the proportion of occurrences of rain preceded by a forecast greater than or equal to the probability in the left column, and POFD the corresponding quantity for no rain*.

FCST PROB	N1	N0	N(N1≥p*)	N(N0≥p*)	POD	POFD
0.00	0	3	82	282	1.000	1.000
0.02	1	47	82	279	1.000	0.989
0.05	5	89	81	232	0.988	0.823
0.10	13	64	76	143	0.927	0.507
0.20	12	38	63	79	0.768	0.280
0.30	16	17	51	41	0.622	0.145
0.40	12	12	35	24	0.427	0.085
0.50	10	8	23	12	0.280	0.043
0.60	4	2	13	4	0.159	0.014
0.70	6	1	9	2	0.110	0.007
0.80	2	1	3	1	0.037	0.004
0.90	1	0	1	0	0.012	0.000
0.95	0	0	0	0	0.000	0.000
1.00	0	0	0	0	0.000	0.000
TOTALS	82	282				

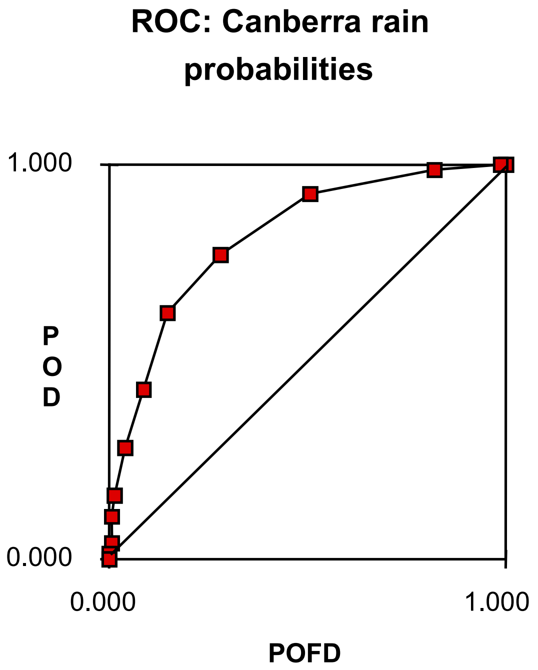
The ROC is a graph of POD (Y axis) against POFD (X axis) for all values of p*. Figure 3 shows the data of table 1 plotted in this way. The form of the resulting curve is purely empirical, determined by the data. While the possible usefulness of these axes is suggested by the SDT model, there is no modelling involved in calculation of the quantities plotted, and no assumptions about underlying distributions.

To orient ourselves on the ROC, it is useful to note several of its properties.

Firstly, the major diagonal represents forecasts which have no skill. In this case learning the forecast does not change one's opinion about the event. This can be shown using Bayes' rule, which shows how new information changes the probability of an event. Recalling that POD = Pr{forecast>=p*|event} and POFD = Pr{forecast>=p*|no event}, and putting p=Pr{event|forecast>=p*} and p_0 = Pr{event} ie the probability of the event before getting the forecast, Bayes' rule in the odds form provides

$$p/(1-p) = [p_0/(1-p_0)]*(POD/POFD) \qquad (2)$$

Figure 3: *Empirical ROC for probability forecasts for rain issued in Canberra at 06.00 for 12.00*



Hence if POD = POFD, the probability of the event is the same after getting the forecast as it was before; the forecasts make no difference to the user's opinion about the probability of rain. It therefore seems reasonable to say that a forecasting system that produces forecasts which plot on the major diagonal of the ROC has shown no skill.

Second, the further the ROC-point is from the major diagonal, the more skilful the forecasts. This is evident from the fact that moving up or to the left either increases POD or reduces POFD, or both.

Perfect skill is indicated by an ROC from 0,0 to 0,1 to 1,1.

The overall performance of the forecasting system, for all threshold probabilities, is indicated by the location of the whole curve in the unit square. A measure of this is the area under the empirical ROC, when the points are connected by straight lines, sometimes denoted PA. PA ranges from 0.5 for forecasts with no skill, to 1.0 for forecasts with perfect skill.

The bi-normal ROC

Referring to the model in fig 2, it is possible to compare ROCs generated by specific distributions with empirical data like figure 3.

Figure 4 shows a family of ROC curves generated by a moving threshold x* when the distributions are Gaussian with equal variances. The four curves correspond to distributions whose means are separated by 0.5, 1.0, 1.5 and 2.0 standard deviations. The similarity of form with the data of figure 3 is evident.

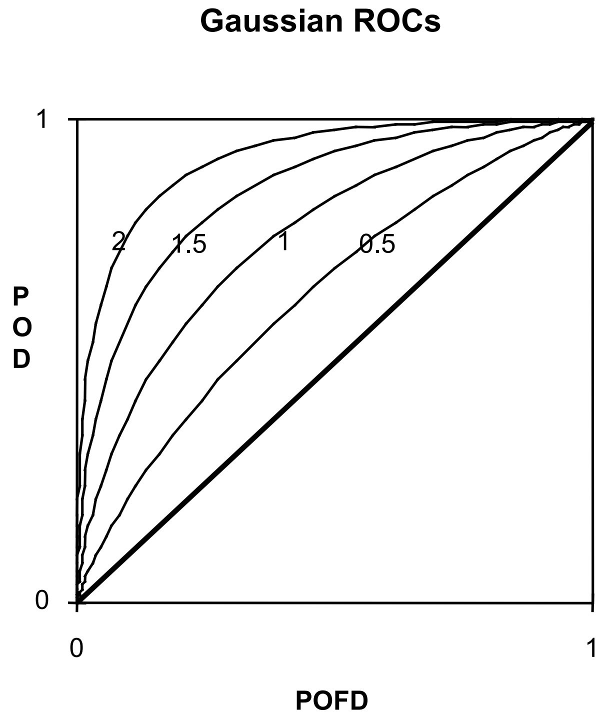


Figure 4: *ROCs generated by moving a threshold through equal variance Gaussian distributions with means separated by 0.5, 1.0, 1.5 and 2.0 standard deviations*.

ROCs generated by Gaussian distributions can be linearised by plotting on double probability axes or, equivalently, axes linear in the standard normal deviate corresponding to the probabilities POD & POFD.

Table 2 shows POD and POFD from table 1, with two additional columns showing the transformation to these deviates.

Figure 4 shows the data of fig 3, plotted on axes transformed in this way.

POD	POFD	Z(POD)	Z(POFD)
1.000	1.000		
1.000	0.989		-2.303
0.988	0.823	-2.251	-0.926
0.927	0.507	-1.453	-0.018
0.768	0.280	-0.733	0.582
0.622	0.145	-0.311	1.056
0.427	0.085	0.184	1.372
0.280	0.043	0.581	1.722
0.159	0.014	1.000	2.192
0.110	0.007	1.228	2.453
0.037	0.004	1.792	2.692
0.012	0.000	2.251	
0.000	0.000		
0.000	0.000		

Table 2: *Calculation of standard normal deviates of POD and POFD*

Figure 4: *Data points represent the same data as figure 3, plotted on axes transformed to the standard normal deviates of those in figure 3. The straight line is the ROC generated by the SDT model in which the means are separated by 1.232 (units sd of POFD distribution), and the ratio of the sds of the distributions is 1.097.*

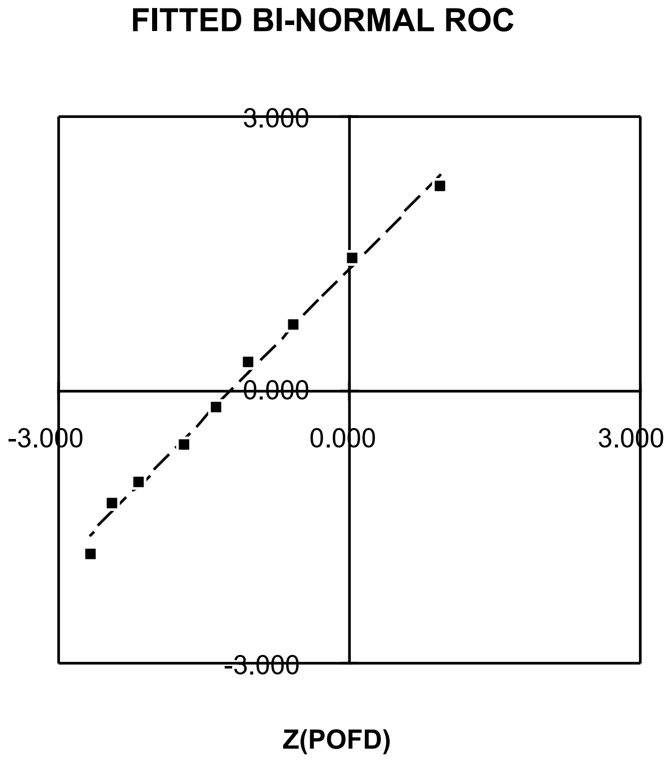


Figure 4 illustrates a robust empirical finding in the field of forecast verification, that the relationship between POD and POFD as decision threshold changes is very close to linear when plotted on double probability axes. The coefficient of linear correlation for the data in fig 3 is 0.9968. Mason (1982) showed that "bi-normal" ROCs for a wide variety of meteorological predictands follow this linear model. Some ROCs have a slight degree of curvature, but even for these, linearity is a very good first approximation (Swets, 1986).

The linearity of empirical ROCs plotted in this way supports the use of Gaussian distributions in the SDT model. Strictly speaking, the linearity of bi-normal ROCs implies only that the underlying distributions can be transformed to Gaussian form by a monotonic transformation.

Computer programs are available to fit the SDT model to empirical data. There is a FORTRAN listing of one such program, RSCORE, in a text by Swets and Pickett (1982), and a version of the same program can be downloaded from http://www-radiology.uchicago.edu/krl/toppage11.htm .

A significant benefit of RSCORE is that it provides variances for the parameter estimates. Seaman et al (1996) have recently commented on the importance of this for assessment of the significance of apparent differences in skill. The absence of significance tests has been a weakness of most published comparative forecast verification.

Summary measures of skill based on the ROC

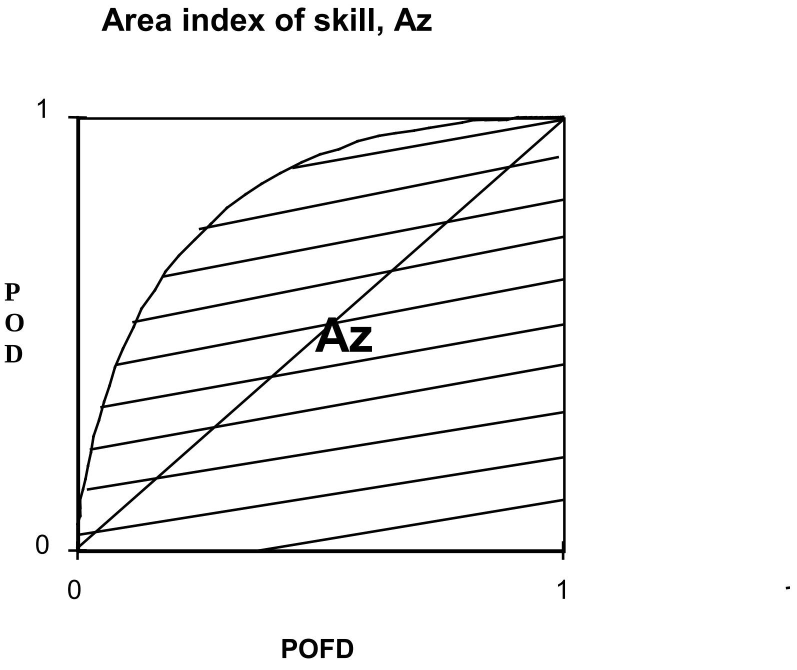
A satisfactory description of forecasting skill in the SDT model requires specification of both the slope and intercept of the straight line representing the system's performance on the bi-normal ROC or in terms of the SDT model the separation of the means of the f_0 and f_1 distributions and the ratio of their standard deviations.

When equality of variances is assumed (slope=1.0) the separation of the means is denoted d'. When this is not assumed, the separation of the means is usually denoted Δm, and is estimated by the X-intercept in units of the sd of the f_0 distribution. The slope of the line, s, provides the ratio of the standard deviation of the f_0 distribution to that of the f_1 distribution. Thus a complete description of the system's intrinsic skill is given by the pair of numbers (Δm,s). Given these parameters the ROC can be reconstructed and the system's performance specified for any and all decision thresholds.

There are occasions when a single-number summary index of skill is desirable. If only a set of yes/no forecasts is available, this only indicates the system's performance at one decision threshold, and hence provides only one point on the ROC. In this case it is necessary to assume a slope s=1.0 for the ROC, and the separation of the means d' provides the index of skill.

Another single number measure of skill, which is recommended by Swets (1986) is the area under the fitted bi-normal ROC transformed back to axes linear in probability. This area is denoted A_Z. Figure 5 shows the bi-normal ROC of fig 4 transformed in this way.

Figure 5: *The smooth curve from 0,0 to 1,1 is the straight line of fig 4, generated by moving a threshold through an SDT model with Δm = 1.232 and s = 1.097, transformed to axes linear in probability. The hatched area is Az, a recommended index of forecasting skill*



Using the fitted curve minimises random sampling variability, and variability due to differing spacing of data points on different ROCs. Az can be found for any set of data that can be plotted on ROC axes, and thus facilitates the comparison of different kinds of forecasts. An estimate of the variance of estimates of Az is provided by the program RSCORE referred to above.

Az has an interpretation as expected proportion correct in a particular kind of discrimination task known in psychology as a two alternative, forced choice task. In weather forecasting this task would involve presenting a forecaster or forecasting system with a series of paired data sets, one of which was followed by the weather event of interest and the other not, the task being to decide which was which. The experimental advantage of this design is that the decision threshold must correspond to a constant probability of 50%, and the climatological probability is also 50%, which eliminates those sources of variability. For non-experts, Az might be described as a standardised form of proportion correct.

It is still the case, however, that a complete description of the intrinsic skill of a set of forecasts for a two-state predictand requires both parameters of the ROC (slope and intercept), so a single number index of skill must generally lose some information and Az is no exception to this rule.

Comments and conclusions

The process of formulation of a forecast for a binary weather event can be modelled as a statistical decision. The variation of quantities analogous to the probability of a type 1 error and to the power of a statistical test, derived from verified forecasts, follow closely a model based on the classical theory of statistical hypothesis testing with underlying Gaussian distributions. Application of this model to assessment of the skill of diagnostic systems was developed by psychologists and engineers seeking to measure the capacity of human and electronic observation systems to detect signals in noise.

The ROC, in the weather forecasting context a graph of POD against POFD as decision threshold varies, is a useful way of assessing pure meteorological skill, and provides either a two-parameter description of skill for all thresholds (Δm,s), or a single number index of skill, d' or Az.

Use of methods from SDT facilitates the assessment of pure skill, in the sense of discrimination capacity, for predictions made in a variety of formats and relatively uncontaminated by variations in the reliability or calibration of forecast probabilities.

Computer programs are available to fit the model, and provide estimates of the variance of fitted parameters, making possible statistical assessment of the significance of differences in skill.

There is a firm basis in classical statistical theory and in numerous empirical studies for the validity of the SDT model in forecast verification, and the power and generality of the results obtainable suggest the time required to become familiar with the use of these methods is likely to be well repaid.

References

Egan, J.P. 1975. Signal detection theory and ROC analysis. *Academic Press*, 277 pp.

Harvey, L.O Jr, K.R. Hammond, C.M. Lusk and E.F. Moss, 1992. The application of signal detection theory to weather forecasting behaviour. *Monthly Weather Review, 120*, 863-883.

Levi, K. 1985. A signal detection framework for the evaluation of probabilistic forecasts. *Organizational behaviour and human decision processes, 36*, 143-166.

Mason, I. 1980. Decision-theoretic evaluation of probabilistic predictions. In The *collection of papers presented at the WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting*, Nice, 8-12 September 1980, 219-228.

Mason, I. 1982a. A model for assessment of weather forecasts. Australian *Meteorological Magazine, 30,* 291-303.

Mason, I. 1982b. On scores for yes/no forecasts. *Preprints of papers delivered at the American Meteorological Society Ninth Conference on Weather Forecasting and Analysis*, Seattle, Washington 169-174.

Mason, I. 1989. Dependence of the critical success index on sample climate and threshold probability. *Australian Meteorological Magazine, 37*, 75-81.

Murphy, A.H. 1977. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review, 105,* 803-816.

Murphy, A.H. and R.W. Winkler, 1987. A general framework for forecast verification. *Monthly Weather Review, 115*, 1330-1338.

Seaman, R., I. Mason and F. Woodcock, 1996. Confidence intervals for some performance measures of yes/no forecasts. *Australian Meteorological Magazine, 45*, 49-53.

Swets, J.A. 1973. The relative operating characteristic in psychology. *Science, 182,* 990-1000.

Swets, J.A. 1986. Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin, 99,* 181-198.

Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293.

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

Swets, J.A., 1996. Signal detection theory and ROC analysis in psychology and diagnostics: collected papers. *Lawrence Erlbaum and Associates Inc*, 308 pp.

Thompson, J.C. and G.W. Brier 1955. The economic utility of weather forecasts. *Monthly Weather Review, 83*, 249-254.

Ian B. Mason

Integrated verification procedures for forecasts and warnings

Report on Consultancy Tasks 1 and 2

Task 1

Review existing and planned AIFS verification modules:

- maximum and minimum temperature verification (TEMPV);
- quantitative rainfall verification (RAINV);
- fire weather verification;
- TAF verification;
- 7-day forecast verification; and
- verification of model output forecasts (MOF)

and make recommendations for any changes in the theoretical and mathematical basis of those modules.

Task 2

Review the existing and planned output formats from AIFS verification modules (as listed under Task 1) and make recommendations for any new designs, or design changes, of those outputs which will make them easy to interpret by various users of such information.

The two types of output format envisaged are:
Output format type A: Diagnostic output for individual forecasters, forecasting teams, numerical modellers and the developers of Model Output Forecasts (MOF); and
Output Format Type B: Simpler output for weather services users, the media, Bureau management and Government.

Take into account:
information provided by the Bureau of Meteorology Verification Group on the verification output format requirements of various verification users.

"…. *all* our knowledge grows *only* through the correcting of our mistakes…"
Karl Popper (1968)

CONTENTS

Overview

- Current and proposed AIFS FVS modules provide a sound basic system and use widely accepted methods of analysis and measures of forecast quality.

- The “diagnostic” or distributions-oriented (DO) framework for forecast verification can be implemented using the present system as a foundation.

- The theoretical framework on which this Report is based is a 3-stage model for the verification process. These stages are

    4. *Data collection*. The output of this stage is the two basic data structures of DO verification, the verification data set and the joint distribution of forecasts and observations.

    5. *Analysis*. In the DO approach this stage involves factoring to joint distribution into marginal and conditional distributions, and calculation of measures of performance and graphical displays of forecast quality.

    6. *Communication*. This stage involves selection of appropriate forms for communication of the results to stakeholders.

- The Report seeks to demonstrate application of the DO approach in each module.

- Measures of forecasts quality for the 2x2 case (POD, FAR, etc) are considered deficient and potentially misleading due to dependence on sample climate and decision threshold. Measures based on signal detection theory (SDT) are more reliable. For non-probabilistic forecasts the measures (d’, β) and Az are recommended.

- With regard to specific modules:

    Mean absolute error (MAE) and bias are recommended as basic summary measures of accuracy in temperature forecasting.

    The SDT measures (d’,β) or Az are recommended as basic summary measures for rain forecasts in the RAINV format.

    MAE and bias are recommended for numerical fire danger ratings, and SDT indices for ratings collapsed onto the warning threshold.

    A 4x2 joint distribution is proposed for TAF verification, in which INTER and TEMPO appear as separate forecasts. The TAF module should be thoroughly tested on validated data.

    For 7-day forecasts the Priestly skill score and MAE or % of forecasts with lower error than climatology are recommended.

    The variety of forecasts produced by MOF makes a simple recommendation difficult. Overall, MAE and bias are suitable for a “first look” at accuracy.

Introduction

Background

In recent years the Bureau of Meteorology has been interested in developing an improved approach to measurement of the quality of its forecast and warning outputs, whether generated by human forecasters or NWP. The Australian Integrated Forecast System (AIFS), currently being implemented in Bureau forecasting offices, has created new possibilities for forecast verification with real time access to databases of forecasts and observations, and little or no requirement for manual processing.

In September 1997 a Forecast and Warning Verification Workshop was held to further develop and coordinate the Bureau's approach to verification. This workshop identified a number of significant issues and recommended strategies to address them. The purpose of the present consultancy is to outline ways to implement several of these strategies.

This report is in response to the brief on page 1, and combines tasks 1 and 2 of the full brief.

Issues addressed under task 1 mainly relate to the theoretical and mathematical basis for the AIFS verification modules, as required. It is clear however that verification must be seen as an integral part of the whole system, including the science of meteorology, operational forecasting, management and public accountability, and some wider comments are offered.

Task 2 is addressed under the heading Communication with users under each module.

In summary, the current AIFS Forecast Verification System is sound, and uses currently accepted methods of analysis and measures of forecast quality. It can be extended to implement some recent developments in the theory of forecast verification, notably the distributions-oriented (DO) framework and methods based on signal detection theory (SDT). These extensions can be undertaken using the present system as a foundation.

This Report criticises some of the measures of forecast quality at present in use, notably POD, FAR and related summary indices for yes/no forecasts, from the standpoint of methods for analysis of skill based on SDT. Nevertheless, it recognises the fact that the "traditional" methods are entrenched, and does not propose that they be abandoned. Measures considered more satisfactory are described and recommended.

Traditional approaches to verification

Methods currently used in weather forecast verification in Australia are typical of those described by Murphy (1997) as "measures oriented" (MO). The MO approach consists largely of (i) calculating quantities considered to be measures various aspects of forecasting performance such as bias, accuracy or skill, and (ii) drawing conclusions regarding absolute or relative performance on the basis of numerical values of these measures. In the absence of a coherent theoretical foundation for selection of measures is difficult to be confident that all aspects of forecast quality have been assessed, or that the measures are reliable for their intended purpose. The DO approach and SDT provide such a theoretical foundation.

New developments

Distributions-oriented methods

Murphy and Winkler (1987) introduced the diagnostic or distributions-oriented (DO) approach to verification. This approach provides a general structure for verification problems, based on two related data sets. These are firstly the verification data sample (or data set), which is simply the sequence of matched (forecast, observed) pairs, sometimes including relevant covariates, and secondly the forecast/observed contingency table, referred to as the joint distribution, which is assumed to contain all the non-time-dependent information relevant to forecast verification. The joint distribution is factored into marginal and conditional distributions using relationships familiar in statistics. The factors provide insights into basic characteristics of the forecasts, the observations and the relationships between them.

Perhaps the main advantage of DO methods is that they impose a formal structure on specific verification problems, as well as on the body of verification methodology. They focus on assessment of basic aspects of forecast quality revealed by the joint distribution and its factorisations, and thereby provide forecasters, managers and users of forecasts with information that is needed to improve all stages of the forecasting process. All summary measures of forecast quality can be derived from the joint distribution and its factors.

A brief introduction to DO methods is provided in an appendix to this report. Useful papers in addition to Murphy & Winkler (1987) are Murphy (1995), Brooks and Doswell (1996), Murphy et al (1989), Murphy & Winkler (1992), and Brown et al (1997). Murphy's ideas were most influential in this field. This Report relies heavily on his work.

Methods from signal detection theory

A second development, still regarded as innovative in weather forecast verification although with a long history in other fields, is the use of methods derived from signal detection theory (SDT) for assessment of forecasting skill. The relative operating characteristic (ROC) is the main analysis tool. The theoretical foundation is the Neyman-Pearson theory of statistical hypothesis testing. SDT-based methods were largely developed to analyse sensory sensitivity in humans and animals, and to separate intrinsic sensitivity from other factors that might bias reports, for example the perceived relative costs of misses and false alarms. (Swets, 1973). They have been highly developed in studies of diagnostic systems in medicine (eg Centor, 1991; Swets, 1996).

The main benefit of ROC/SDT methods in forecast verification is that they provide a means of assessing the capacity of a forecasting system to discriminate between occurrence and non-occurrence of an event, on the basis of its issued forecasts. This approach provides the only measures of forecasting skill available at present which are unaffected by either sample climate or decision threshold. This makes it possible to separate assessment of meteorological skill from other factors. These other factors include the reliability or calibration of forecasts (more an issue of communication with users than of meteorological skill), decision thresholds (which are determined by non-meteorological factors such as users' economic sensitivity to forecasts and weather, and attitudes to risk) and the frequency with which weather events present themselves to be forecast (sample climate).

A feature of ROC/SDT methods is an empirically validated model for the co-variation of basic parameters of forecasting skill with decision threshold. This can provide valuable insights into the properties of many current verification measures, and provides a criterion for selection of reliable measures (Mason 1982b; Swets 1986).

An outline of the theory and some implications for practice are provided in Appendix 9.4. Papers directly relevant to meteorology include, in addition to those mentioned above, Mason (1982b), Levi (1985), McCoy (1986), and Harvey et al (1992). References to the wider literature on ROC/SDT methods can be found in Swets (1996).

The structure of the verification process

In order to provide some structure for consideration of individual modules in this Report, forecast verification is modelled as a three-stage process.

These stages are

1. Extraction of the verification data sample and joint distribution of forecasts and observations. These are the basic data structures for any verification problem. The joint distribution contains all the non-time-dependent information relevant to assessment of forecasting skill.

2. Calculation of
   a) performance characteristics based on factorisation of the joint distribution into marginal and conditional probabilities (distributions-oriented verification) and
   b) verification measures, which summarise significant aspects of forecast quality.
3. Communication with stakeholders, involving selection of appropriate measures and forms of display, etc.

Further detail is provided in Appendix 8.1.4

This report

In this Report the current or pre-AIFS verification practices are reviewed for each of the verification areas listed in the Brief. Then the proposed or current AIFS module is considered, with particular attention to the theoretical and mathematical basis for measures used, and comments and recommendations are made where appropriate. Then an outline is given of a DO approach to verification in each case, with discussion of issues related to extraction of the verification data sample (VDS) and joint distribution (JD) of forecasts and observations and performance characteristics and verification measures derived from the JD and/or based on ROC/SDT methods. Communication with stakeholders is then considered, and some recommendations made regarding appropriate methods.

Appendices

Appendices to the Report contain background material which is common to all sections.

Recommendations

Recommendations appear in the main text in bold type, and are collected in a Summary at Appendix 8.3

Maximum and minimum temperature verification (TEMPV);

Temperature verification pre-AIFS

The official system for verification of routine maximum and minimum temperature forecasts In the Bureau has been TEMPV. This is a desktop computer system with manual input. Output is familiar to most Bureau meteorologists. A typical example is:

TEMPV statistics, RFC PM	Max	Min
Mean obs temp	**20.1**	**11.1**
Forecast bias	**-0.8**	**-1.9**
STDEV obs temp	**4.5**	**3.3**
STDEV forecast temp	**4.5**	**3.0**
Mean mod forecast error	**3.6**	**3.2**
RMS error pers forecast	**5.5**	**4.2**
RMS error forecast	**5.1**	**4.1**
RMS error (obs>mean+SD)	**6.7**	**5.3**
RMS error (obs<mean+SD)	**9.4**	**5.0**
Skill score (forecast)	**-0.3**	**-0.6**
Skill score (persistence)	**-0.5**	**-0.7**
Extreme positive error	**13.0**	**5.0**
Extreme negative error	**-12.0**	**-8.0**
No. of forecast errors >= 3 deg	**5**	**7**
No. of forecast errors >= 5 deg	**2**	**3**
80% of errors <= (deg)	**4.0**	**5.0**
90% of errors <= (deg)	**12.0**	**8.0**
Data pairs (forecast)	**16**	**16**

A similar table is provided for forecasts issued in the morning and in the afternoon.

The AIFS system

The AIFS system is based on TEMPV, with substantial extensions taking advantage of the capabilities of the AIFS platform. The outline below is based on the document **FVS (Forecast Verification System) HELP Max/Min Temperature Module** on http://servb.ho.bom.gov.au/fvs/docs/fvstehlp.html.

The AIFS system is interactive, from a menu. Data is automatically collected without need for any manual input. Forecasts of maximum and minimum temperature for all regional cities for which forecasts and observations are available can be verified. The forecasts verified are, at present, the official issued forecasts, GASM/NWP, and MOS and MOF guidance systems. The capacity also exists to verify forecasts from external sources.

Verification can be performed as soon as the relevant observation is available. Forecasts issued at lead times up to seven days can be handled.

A variety of tabular and graphical outputs are provided, together with various measures of forecasting performance, discussed in more detail below. The output variables provided by TEMPV are still available, for continuity.

The name of the originating forecaster is recorded, although the ability to collate data for individuals is not yet implemented.

Measures of performance provided in AIFS for temperature forecasts

The table of verification statistics provides the following:

(a) mean observed maximum and minimum temperatures,

(b) bias of the forecasts, equal to the difference between the average forecast and average observation,

(c) variances for the forecasts and observations,

(d) mean absolute error (MAE),

(e) root mean square error (RMSE),

(f) RMSE for persistence and for ‘climatology’, a forecast of the climatological mean for that day. Daily climatological means are found by fitting a smooth curve to monthly means.

(g) RMSE when observations are greater than the observed mean plus one standard deviation,

(h) RMSE when the observations are less than the observed mean minus one standard deviation,

(i) the skill score SS = 1 - MSE/Variance of observations,

(j) SS for persistence,

(k) largest positive and negative errors,

(l) number of errors greater than or equal to 3C,

(m) number of errors greater than or equal to 5C,

(n) absolute error amount (degrees C) that contains 80% of the forecasts,

(o) absolute error amount (degrees C) that contains 90% of the forecasts, and

(p) total number of (forecast, observed) pairs verified.

All these quantities can be calculated for official or guidance forecasts.

Comments

The performance statistics above are identical to the original TEMPV output. There is a lengthy time series of most of these statistics available, and it is desirable to keep them for continuity.

Recommendation: that the tabulated values presented for TEMPV be maintained for continuity.

Theoretical and mathematical properties of temperature verification measures

Root Mean Square Error (RMSE)

RMSE is a widely used measure of skill in temperature forecasting. If forecasters seek to minimise RMSE then they should forecast the mean of their judgmental distribution.

An important practical point is that RMSE is affected by different aspects of accuracy, and by factors that are not strictly related to accuracy at all. This can be seen from a basic decomposition of the mean square error (MSE),

$$\mathrm{MSE}(f,x) = (\mu_f - \mu_{x)}{}^2 + \sigma_f^2 + \sigma_x^2 - 2\sigma_f\sigma_x\rho_{fx} \qquad (1)$$

(Murphy 1988), where μ_f and μ_x are the mean forecast and observed temperatures respectively, σ_f^2 and σ_x^2 are the corresponding variances and $\sigma_f\sigma_x\rho_{fx}$ is the covariance between the forecasts and observations.

Each of these components of MSE refers to a different aspect of forecasting performance, with different implications for assessment of forecasting skill. Bias, as described by the first term in the above equation, may be quite easy to eliminate if it is consistent. Williams' (1997) work on temperature forecasting in NSW showed that a significant improvement in MSE can be obtained by recognising the contribution of bias and removing it. The second term, the variance of the forecasts σ_f^2, is not directly affected by the observations, but contributes to MSE, and MSE can be reduced to some degree by artificially reducing the variance of the forecasts without reference to the observations (Barnston 1992). The third term, variance of the observations, is the MSE of a constant forecast of the (sample) climatological mean temperature. This is not under the control of the forecaster, so it is questionable whether it should affect measures of forecasting skill. Also, it can vary between locations and between forecast sets at the same location, and thus introduces an element of variability into MSE that is not related to variations in forecasting skill. The last term is (twice) the covariance of forecast and observed temperatures, arguably the part that is most directly related to meteorological skill in forecasting.

Hence unpartitioned values of MSE should be taken with caution, particularly if used to compare forecast sets with differing climatologies.

Components of Murphy's partitions of the MSE are available as part of the scatterplot option in the AIFS temperature verification module.

Mean Absolute Error (or MMFE, mean modulus of forecast error)

MAE is a useful summary statistic and is the main single measure of temperature forecast quality in the US NWS. If forecasters seek to minimise MAE then they will forecast the median of their judgmental temperature distribution. Using MAE is equivalent to saying that the "seriousness" of an error is equal to the size of the error, ie a linear penalty function.

MAE can be partitioned in a similar way to MSE, but this has had little or no attention in the verification literature.

Skill score

The skill score for temperature forecasts, sometimes referred to in Australia as the Priestly skill score, is

SS=1-MSE (forecasts)/MSE(climatology)

which is equal to 1-MSE(forecasts)/σ_x^2 *if it is assumed that the variance of the observations in the verification sample is the same as long-term climatology*. Use of sample climatology as a reference standard facilitates a neat expression for SS (below) but does not give the forecasting system credit for recognising differences between sample and long-term climate. The values of SS calculated in the AIFS module takes the preferable approach, using long-term daily climatology rather than sample values.

A decomposition of SS similar to that of MSE above can be obtained (Murphy 1988). This is

$$SS = \rho_{fx}^2 - [\rho_{fx} - \sigma_f/\sigma_x]^2 - [(\mu_f - \mu_x)/\sigma_x]^2 \qquad (2)$$

The first term is the proportion of variance in the observed temperatures explained by linear regression on the forecasts. The second term is zero when the regression line has zero intercept and unit slope, ie perfect forecasts., otherwise it acts to reduce SS. The third term is a measure of bias. The second and third terms are often numerically quite small, so SS is approximately equal to the proportion of variance explained by the forecasts.

Further information on interpretation of the components of MSE and SS is in Murphy (1988) and (1997). Values of the components of these partitions are provided in the AIFS module in the scatterplot option.

There are other ways of partitioning MSE, for which see Murphy (1988).

In practice, the implication of the above is that raw, unpartitioned values of MSE and SS are subject to variations associated with different aspects of forecasting skill, and in the case of MSE not associated with forecasting skill at all. It is desirable that comparisons between forecast sets are based on the components, rather than on the unpartitioned values.

New measures of forecast quality

Bayesian Correlation Score

A new score for non-probabilistic forecasts of continuous predictands has been proposed by Krzysztofowicz (1992), which may have theoretical advantages over MSE or SS as a summary measure of quality for temperature forecasts. The Bayesian Correlation Score (BCS) is defined in terms of the parameters of the fitted regression model. Its main advantage is that it orders alternative forecast systems in terms of their expected economic value to users. BCS is defined by

$$BCS = [\sigma^2/(a^2\ S^2) + 1]^{-1/2} \qquad (3)$$

where σ^2 is the variance of the forecast temperatures about the regression line, a is the slope of the regression line and S^2 is the climatological variance of the observed temperatures. There is little experience with this measure as yet, although it is understood that the Canadian national weather service has used it.

In view of the plethora of measures of accuracy already available in the AIFS temperature FVS it is difficult to recommend yet another. Nevertheless, the claimed monotonicity with expected economic value would be a significant advantage, so BCS may be worth further investigation.

LEPS score

Another score suitable for evaluation of temperature forecasts which has been developed recently is the Linear Error in Probability Space (LEPS) score (Ward and Folland 1991; Potts et al 1996). The magnitude of the penalty for forecast errors is scaled according to the location of the forecast and observation in the cumulative probability curve, so that errors near the climatological mean are penalised more severely than those in the tails of the distribution. The underlying rationale is that forecasts of a quantity in regions of its probability distribution where it is relatively more likely should be easier than for values that are climatologically less likely.

It is not immediately self-evident that this is always the case. Climatologically rare events may be preceded by clearly anomalous precursors, so that the forecast is not especially difficult. Conversely, there is a range of meteorological situations than can lead to observations near the climatological mean, not all of which are necessarily easy from a forecasting viewpoint.

Wilks (1995) comments that “it is too soon to judge the extent to which LEPS-based skill scores will be useful or well-accepted…”. The paper by Potts et al (1996) gives extensive discussion of the LEPS score in relation to MSE. As with the BCS above, LEPS warrants further investigation, but could not be seen as a high priority.

Indices based on the distribution of errors

The output also provides some parts of the error distribution. These are the largest positive and negative errors, the number of errors greater than or equal to 3C, the number of errors greater than or equal to 5C, the absolute error that contains 80% of the forecasts and the absolute error that contains 90% of the forecasts.

It is preferable to give the full error distribution, both signed and in terms of absolute errors, and also in cumulative form for absolute errors. Then all the information on the error distribution is available.

Recommendation: That error distributions be provided for signed and absolute errors, and in cumulative form for absolute errors.

Persistence and climatology as baselines for skill

SS is a skill score that uses climatology as a baseline for skill, so SS indicates proportional improvement in skill over a forecast of the climatological mean temperature for that day. The AIFS module also calculates some scores using persistence as a forecast.

It has been found that a linear combination of climatology and persistence can outperform either separately (Murphy, 1992, 1996). There are methods of estimating the optimal linear combination, but a simple mean may capture most of the available improvement. Since it is desirable to assess forecasts against the most accurate no-skill baseline, it is recommended that measures of skill be considered for a combination of persistence and climatology.

Recommendation: That a combination of persistence and climatology be investigated for use as a baseline for the skill score.

Other output from the AIFS module

A variety of tabular displays is available. Of particular note is a capability to compare forecasts from different sources (eg official and MOS).

Graphical displays include time series of forecast and observed temperatures with an interactive capability to display forecast and observed temperatures for individual dates and the forecaster ID.

A very useful display is the scatterplot of forecast vs observed temperatures, with an interactive capability similar to that described for the time series display.

Regression analysis is useful in analysis of temperature forecasts, as it reveals consistent biases and facilitates their correction in real time. Williams (1997) used regression to identify significant biases in guidance forecasts, with a demonstrable and significant impact on the quality of temperature forecasts.

The current AIFS temperature module provides a scatterplot with fitted regression line and only needs the addition of slope and intercept for the line to be useable by forecasters to reduce systematic biases in real time. Ideally t-values and the correlation coefficient should also be provided.

Recommendation: that the scatterplot display include values for slope and intercept of the fitted regression line.

Comments

The graphics output of the AIFS temperature FVS gives a useful picture of forecast performance, and should be sufficient for most requirements except exhaustive evaluation of forecasts in a research environment.

Implementation of the distributions-oriented framework in temperature verification

This section outlines briefly an approach to temperature forecast verification using distributions-oriented (DO) methods, based on Murphy and Winkler's (1987) general framework. Implementation of this approach in the AIFS verification module would require no additional data collection.

Examples of DO verification of temperature forecasts are in Brooks & Doswell (1996) and Murphy et al (1989), on both of which the following discussion relies.

Implementation of the full DO framework generates a substantial amount of information. The forecasts and observations are analysed separately and in terms of their relationships. Thus a full DO analysis is essentially a tool for research into the basic properties and capabilities of a forecasting system (machine or human). A subset of the information generated can be selected for specific user groups.

Brooks and Doswell comment on the DO approach to temperature verification as follows.

> "If the approach to verification is limited to simple summary measures, the richness of the relationship between forecasts and observations is lost. What appear as issues of fundamental importance when considering a distributions-oriented approach to verification cannot even be asked with a measures-oriented approach, since the presentation of the data does not allow the issues to be *identified*. Simple summary measures of overall performance offer almost no information about the relationship between forecasts and errors…"

Verification Data Set

The basic data for forecast verification is contained in the verification data set or sample (VDS), a sequence of matched (forecast, observation) pairs (together with any covariates that may be required).

In the case of temperature forecasts this is simply the sequence of forecast and observed temperatures, together with relevant dates and times. This is already available in the AIFS FVS for temperature.

A useful additional capability would be the ability to select subsets of the full VDS for verification, for example to verify occasions on which the observed or forecast temperatures were above or below a user-selected threshold

Recommendation: that the facility to extract subsets of the VDS on user-defined thresholds be provided.

The basic VDS should be compatible with other Bureau database systems, and downloadable to PC-based spreadsheet systems.

Joint Distribution

The VDS is summarised as a joint frequency (or relative frequency) distribution, in which the rows represent forecasts and the columns, observations. The elements of this joint distribution (JD) are sample estimates of the probability of that combination of forecast and observation, usually represented as p(f,x). The JD contains all the non time-dependent information relevant to assessment of forecast quality. In the case of temperature forecasts on a monthly or quarterly basis it would usually be necessary to group the data in 3C or 5C ranges to obtain adequate frequencies in each cell.

Recommendation: that the joint distribution of forecast and observed temperature be provided. for the actual forecasts and for the persistence/climatology baseline forecasts.

Factorisation of the joint distribution

Murphy & Winkler (1987) point out that the information in the joint distribution is more easily understood when factored into conditional and marginal distributions. For a two-dimensional JD there are two of these factorisations, specifically

$$p(f,x) = p(f).p(x|f), \text{ and} \tag{4}$$

$$p(f,x) = p(x).p(f|x). \tag{5}$$

Further discussion of the use of these quantities can be found in the appendix to this report and in the references cited above. For the present it is noted that all currently used summary measures of forecasting skill, including MSE and SS can be derived from the JD and the above factorisations (Murphy 1996).

Recommendation: That conditional and marginal distributions be available as an option for all temperature forecasts and for the persistence/climatology baseline forecasts.

Box plots and conditional quantile plots

DO verification has used some forms of graphic display that are still unfamiliar to many meteorologists. The main examples in temperature verification are box plots and conditional quantile plots.

Tukey (1977) invented box plots (sometimes called box-and-whisker plots) as a convenient means of summarising and comparing frequency distributions. Conventional practice in forecast verification usually extends only to presentation and comparison of means of distributions, occasionally with standard errors (implying symmetry). Box plots describe the central tendency, variability and symmetry or lack thereof by displaying five sample quantiles; the median (50 percentile), upper and lower quartiles (75 and 25 percentiles) and the upper and lower deciles (90 percentile and 10 percentile). Practice varies somewhat in details. Sometimes the upper and lower extreme values are also plotted, and a variant called the notched box plot also shows an estimate of the standard deviation of the median.

Box plots are more informative than conventional data summaries for distributions. They are an appropriate means of displaying the distributions of forecast and observed temperatures and of forecast errors. Examples of applications in meteorology can be found in Murphy et al (1989), Graedel and Kleiner (1985), Wilks (1995) and Brown et al, 1997.

Another useful form of display for temperature verification data is the conditional quantile plot. The conditional distributions p(x|f) and p(f|x) provide information about relationships between the forecasts and observations in terms of several dimensions of forecast quality. Conditional quantile plots of p(x|f) are graphs of observed temperature against forecast temperature in which the 10, 25, 50, 75 and 90 percentiles of the observed temperatures are plotted, rather than actual temperatures. In conditional quantile plots of p(f|x) the same quantiles are plotted but of the forecasts, against the observed temperatures. Examples can be found in Murphy et al (1989).

Use of these forms of display in AIFS may require acquisition or development of specialised software. Nevertheless, graphic displays can communicate the main features of complex data more effectively than tables of numbers, and it is considered that comprehension and hence utilisation of verification results would be enhanced by their use.

Recommendation: that the possible usefulness of box plots, conditional quantile plots, and other appropriate graphic displays for verification data, be investigated.

Applicability of SDT/ROC-based methods in temperature verification

ROC-based methods for assessment of skill are appropriate when the forecasts are for binary events, and are preferable to “traditional” methods in this case. The direct applicability of ROC-based methods is limited when the forecasts are for a continuous predictand like temperature. Nevertheless, some users of temperature forecasts have a particular interest in specific threshold temperatures. The most important of these may be the temperature at which frost forms, for protection of crops. There may be other temperatures of importance to other users. When temperature can be dichotomised in relation to a specified threshold, methods from signal detection theory are appropriate to assess skill in forecasting. The results of a series of forecasts can be summarised in the usual 2x2 joint distribution, and the indices d’ and β, or Az, can be calculated following procedures outlined in Appendix 9.4.4.

d’ and Az are alternative measures of pure forecasting skill, in the sense of an ability to discriminate between states of the predictand. d’ is the separation of the means of the underlying “noise alone” and “signal plus noise” distributions, in units of the “noise alone” distribution. Az is the area under the fitted ROC. Az is in general preferred, as it is appropriate for forecasts produced in any form. d’ is only appropriate for yes/no forecasts, as it assumes a specific value for a parameter that cannot be estimated from a single set of yes/no forecasts.

β is an index of implied decision threshold, the level of certainty at which the forecast is changed from non-occurrence to occurrence. β is a likelihood ratio, and can be converted to a threshold probability using Bayes’ rule in the odds form for binary events (see Mason 1982a). While the forecasts in this case are issued as temperatures and not explicitly as probabilities, any set of less than perfect forecasts that can be summarised as a 2x2 forecast/observed array implies a certain decision threshold on an underlying scale related to probability. The specific threshold is of some importance because the expected utility of the forecasts is optimised at a threshold that depends on the economics of the user’s decision situation. If the actual implied threshold is different from the optimal threshold then the forecasts are not achieving their full potential value.

Recommendation: That the measures d’,β and Az be available for temperature forecasts collapsed to yes/no forecasts by thresholding on a critical temperature. This temperature should be variable by the user.

Communication with users

Output Format Type A:

Diagnostic output for individual forecasters, forecasting teams, numerical modelers and the developers of Model Output Forecasts (MOF)

Information for forecasters and forecasting teams

Forecasters need real time information on the current performance of guidance forecasts, for the purpose of identifying consistent biases. Williams (1997) has shown that verification using simple linear regression can reveal biases which can be allowed for in issued forecasts, producing significant improvement in error statistics.

An appropriate form of display is, as used by Williams, a scatterplot of observed against forecast temperatures with the fitted regression line, the slope and intercept of the line, and the correlation coefficient. The period of record for the scatterplot should be variable, with the default being the past 30 days. The scatterplots should be available for all guidance forecasts, and for the official forecasts, with values of MAE for each.

Recommendation: That scatterplots of observed vs forecast temperatures should be available for all guidance forecasts, and for the official forecasts, with statistical parameters of the fitted line and values of MAE for each.

One of the main uses of verification information is as feedback to forecasters and forecasting teams on their own performance. Scatterplots as detailed above should also be available for individual forecasters, preferably at the forecaster's initial logon at start of shift. This should take the form of forecast and observed values for the individual's previous shift and scatterplots with fitted regression data and MAE for the previous 30 forecasts, and for other periods as an option. Management also has a need for this kind of information, as outlined below.

Recommendation: that verification information as detailed above for individual forecasters be available at logon at the start of each forecaster's shift.

The extent to which this information should be available other than to management and the individual forecaster requires further consideration.

Research

The detailed data on forecasting performance available from a DO verification is a rich source of information for research, and should be as widely available as possible. Ready access to verification data is likely to encourage its use.

Numerical modelers and developers of MOF and other types of guidance should have access to the full range of output from the FVS. Use of verification in research may require unusual types of analysis, and stratification of the VDS by other covariates. This would be facilitated by ensuring that the VDS is compatible with other Bureau database systems, and readily downloadable to common spreadsheet systems.

Output Format Type B:

Basic information for weather services users, the media, Bureau management and Government

Communicating with the public

As a single measure of the correspondence between forecast and observed temperature suitable for dissemination to the media and general public, mean absolute error (MAE) is recommended. MAE seems to correspond to the way the "ordinary reasonable person" thinks of accuracy. For the public this quantity could be referred to as "average accuracy of forecast temperatures" (rather than "average error…"). It is easy to explain as the average size of the difference between daily forecast and observed temperatures.

A second choice would be the skill score SS described above. This may be preferable from a theoretical point of view, as it shows the proportion or percent improvement over a zero-skill baseline and so shows the "value added" by the skill of the forecasts. However it is likely to be less comprehensible than MAE to a public in which mathematics, or even a high degree of interest, cannot be assumed. SS is not completely straightforward and is difficult to explain succinctly without distortion. Similar comments apply to RMSE

For public information MAE could be included on the Bureau's public web site in more or less the following form:

ACCURACY OF TEMPERATURE FORECASTS FOR CANBERRA IN OCTOBER 1998

Average accuracy of afternoon forecast for tomorrow's maximum:	3.5C
Average accuracy of afternoon forecast for tomorrow's minimum:	2.5C
Average accuracy of morning forecasts for today's maximum:	2.9C

with an explanation possibly in a hypertext link of the meaning of "average accuracy".

Recommendation: That MAE be adopted as a single basic measure of temperature forecast accuracy for public information throughout Australia, and be referred to as "average accuracy".

Information for managers and others

Users of weather services, management, Government and members of the public with a more than casual interest in temperature forecasts need a higher level of detail than MAE alone. They can not generally be expected to spend the time required to extract significant elements from the mass of information potentially available from the AIFS FVS.

At the next level of detail it is recommended that the error distribution be presented in addition to MAE, both for signed errors and absolute errors, and in cumulative form for absolute errors, for afternoon forecasts for maximum and minimum and morning forecast for maximums. These distributions should be available in both tabular and graphic form (e.g. as bar charts), accompanied in each case by the number of errors greater than 3C and 5C as at present, and the temperature intervals that contain 80% and 90% of the forecasts. This information could be on the public web site as an optional link for "more detail", updated regularly

Bureau management also has an interest in information on the performance of individual forecasters, as an aspect of monitoring the quality of the service. The most informative reasonably succinct summary for individuals is the scatterplot, together with fitted regression line and regression parameters, with the corresponding MAE.

Although verification of individual forecasters is a legitimate part of monitoring the overall quality of the Bureau's output, concerns have occasionally been expressed about the potential for unfairness in this practice. These concerns need to be addressed, but are beyond the scope of this report.

Annual Reports

The level of detail required in Annual Reports to satisfy accountability concerns about the accuracy of temperature forecasts is unclear; practice varies widely between Regions. An error distribution stratified by seasons for maximum and minimum temperature forecasts, together with the corresponding values of MAE, would probably be as much detail as most readers of Annual Reports would wish to cope with. A contact could be given for further information if required.

Quantitative rainfall verification (RAINV);

Pre-AIFS system

In the rain forecast verification system RAINV, forecasts are prepared twice daily for capital cities. One is prepared at 8pm for the 24-hour period starting at 9am the following day and the second at 6am for the same period, updating the first forecast. The forecasts are done in ranges defined as follows:

0	**no pptn**
1	**0.2 to 2.4mm**
2	**2.5 to 4.9mm**
3	**5.0 to 9.9mm**
4	**10.0 to 19.9mm**
5	**20.0 to 39.9mm**
6	**40.0 to 79mm**
7	**above 80mm**

The observations are the mean of a group of gauges around the city, at most cities. Some use only the single official city gauge. The average rainfall over the group of gauges is converted to a range as above.

The forecasts are not issued directly to any customer. They are done solely to assess skill in rain forecasting.

The output of the pre AIFS system consists of a table of the following statistics:

% correct (rain/no rain)
% correct (range)
% within one range
% errors > 3 ranges
forecast bias (range)
mean modulus of range error
HK skill score (rain/no rain)
extreme range error

Values for these statistics are calculated for the actual forecasts and for persistence at both issue times.

Other output consists of two 8x8 contingency tables of forecast against observed ranges, for the forecasts and for persistence, and two 2x2 contingency tables, collapsing the 8x8 tables into rain/no rain tables for the forecasts and for persistence.

AIFS system

The current AIFS system is based on RAINV, with extensions that take advantage of the capabilities of the AIFS platform. Details are in the document **FVS (Forecast Verification System) HELP Quantitative Rainfall Module**.

The following outlines briefly the main features.

The forecast types that can be verified are the official forecast, and GASM, MOS and NWP guidance systems.

AIFS provides near real-time feedback, in that verification can be performed as soon as the observations are available.

Forecasters can identify their own forecasts.

There are three data displays available. These are

1. A text display showing the forecast and observed rainfall ranges, and rainfall amounts in mm can also be shown.

2. Time sections of forecast and observed ranges on the same axes, and also the differences between observed and forecast ranges. A scatterplot option is also available, showing observed against forecast ranges.

3. Four contingency tables and two tables of scores. The contingency tables are two 8x8 tables of forecast against actual ranges for the forecasts and for persistence, and the same data collapsed into 2x2 (rain/no rain) tables. The scores are the same as those listed above for RAINV.

Comments on the AIFS system

The AIFS quantitative rain forecast verification module essentially implements RAINV in the AIFS environment, with the addition of the graphics time series displays and the capability of verifying guidance forecasts from several sources.

The text display is equivalent to the basic verification data set required in a full DO implementation, and the tables of forecast against observed ranges are the basic joint distribution (lacking only the marginal totals).

Verification measures currently provided in AIFS

AIFS at present provides essentially the same set of verification measures as RAINV.

There is a long historical record for the verification measures used in RAINV, and they should be maintained for continuity.

Recommendation: That the current verification measures used in the AIFS rain FVS be maintained for continuity.

Scores

Percent correct (PC) is calculated for the 8x8 contingency table and Hansen & Kuipers score (HK), for the 2x2 tables only. Interpretation of both PC and HK is made unclear by the fact that they depend on the decision threshold implied for selecting rain or no-rain as the forecast, and PC also depends on the climatological probability of rain.

While the forecasts are not in general formulated with a specific decision threshold in view, any set of less than perfect forecasts for a binary event does imply a certain decision threshold, and the location of this threshold affects values of PC and HK. PC is maximised at a threshold probability of 50% and HK is maximised when the decision threshold is equivalent to the climatological probability of the predictand. This issue is further discussed in section 8.3

Hansen and Kuipers' score (and also the Heidke score) can be calculated for multicategory contingency tables. Details can be found in Wilks (1995). While there is some attraction in a single number summary indicator of overall skill, a measure of performance of this kind cannot provide a satisfactory description of forecast quality. The basis for this statement is to be found in the concept of the dimension of verification problems.

In brief, the dimension of a verification problem is the number of parameters that must be determined in order to reconstruct the joint distribution p(f,x). In the case of an 8x8 JD, the dimension is 8*8-1 = 63 (Murphy 1997), that is a complete reconstruction of the 8x8 JD could require that as many as sixty-three joint (or marginal and conditional) probabilities be determined. Reducing the dimension from sixty-three to one inevitably results in loss of information about some aspects of forecast quality. Murphy (1997) states that "Considerations related to the dimensionality of forecast verification problems have generally been ignored in the traditional … approach to forecast evaluation. To provide a reasonably complete description of forecast quality the dimensionality of verification problems must be respected."

Elsewhere Murphy (1996) states that "It is now generally understood that no universally acceptable measure of performance for the kxk problem can be found."

Hence the multicategory forms of the Heidke and Hansen and Kuipers' scores are not recommended.

Other output

The "standard" output from RAINV includes a number of measures based directly on the distribution of errors in ranges, for example the percentage of occasions on which the error was zero, within ±1 category, greater than 3 categories, etc. It seems desirable in addition to provide the full distribution of errors, both signed and as the modulus of the error to show the cumulative distribution of errors of zero, 1 category or less, 2 categories or less, etc. This information can be displayed as histograms as well as in tabular form.

It would also be interesting to see distributions of actual rainfall amounts for each forecast category. In addition to the tabular output box plots would be an appropriate display for this kind of data, as it is likely to have a markedly skewed form.

A distributions-oriented framework for verification of categorical rain forecasts.

A general outline of the distributions-oriented (DO) approach to forecast verification is at section 8.1. In brief, DO verification operates on the basis of two data sets. These are the verification data sample (or set) (VDS) and the joint distribution of forecasts and observations (JD).

The verification data set

The VDS is simply a sequence of matched (forecast, observation) pairs, together with any covariates of interest. The VDS may be a time series of forecasts and observations, or it may be selected from a larger body of forecasts and observations on the basis of some criterion (for example all the occasions of anticyclonic southeasterly airflow at the surface).

The VDS for the RAINV rain forecasts is already available in the AIFS module, and in fact the AIFS module provides the further useful capability of listing the actual rain amounts in mm against each forecast, rather than just the observed category.

The joint distribution

The JD is a contingency table in which the rows correspond to forecasts and the columns to observations (or vice versa), and the elements are (sample estimates of) the joint probabilities p(f,x), where f represents the forecasts and x the observations. Murphy (1997) states that "Under the assumption that the bivariate time series of forecasts and observations is (i) serially independent and (ii) stationary in a statistical sense, the distribution p(f,x) contains all the relevant information in the verification data sample. *Forecast quality*, as defined here, is the totality of the statistical characteristics of the forecasts, the observations, and their relationships embodied in this distribution."

The JD is available from the AIFS module, as the 8x8 forecast/observed contingency table, except that it does not provide marginal totals. These should be provided, if only to save users the task of manually calculating them.

Factorisations of the joint distribution

Following Murphy and Winkler (1987), the joint distribution is factored into conditional and marginal distributions. Two such factorisations can be defined:

$p(f,x) = p(f).p(x|f)$, called the calibration/refinement or CR factorisation, and

$p(f,x) = p(x).p(f|x)$, called the likelihood/base rate or LBR factorisation.

These marginal and conditional probabilities are referred to as *performance characteristics*, to distinguish them from verification measures.

To implement the full DO framework in the AIFS module the conditional distributions should be provided. If the joint probabilities are denoted $p_{ij} = p(f=f_i,x=x_j)$ and the marginal distributions are $p_{.j} = p(x=x_j)$ and $p_{i.} = p(f=f_i)$, then the conditional probabilities for the CR factorisation, $p(x=x_i|f=f_j)$ are given by

$$p(x=x_i|f=f_j) = p_{ij}/p_{i.} \tag{6}$$

Similarly, the conditional probabilities for the LBR factorisation are given by

$$p(f=f_i|x=x_j) = p_{ij}/p_{.i} \tag{7}$$

The 2x2 rain/no rain JD can be factored in the same way as the 8x8 table, and these factors should also be available as an option.

Recommendation: that the full CR and LBR factorisations for the 8x8 and 2x2 JDs be available as an option.

Useful graphics displays in addition to those already provided could include histograms of the marginal and conditional distributions as described above, and plots of actual rainfall amounts against forecast categories.

SDT indices of skill and decision criterion

Methods from signal detection theory are applicable only to forecasts of binary events. While it may seem that this eliminates the possibility of using SDT to assess the skill of the forecasts represented by the 8x8 JD, it is possible to reduce the 8x8 distribution to a series of seven 2x2 distributions by thresholding on successive rainfall range boundaries. Two considerations suggest that this is an appropriate procedure:

(i) Thresholding on category boundaries corresponds with the way in which a decision-maker might use the forecasts. In general, someone with a decision to make on the basis of forecast rainfall would have a threshold amount such that he/she would do one thing if the forecast rainfall were less than this amount and another if it were greater. Hence the results of a series of decisions can be represented in a 2x2 table produced using the appropriate threshold, and the performance of the forecasts for this user can be represented by a joint distribution.

(ii) As noted above, there is no satisfactory single summary measure of skill for multicategory contingency tables. An adequate description of forecast quality requires that the dimensionality of verification problems be respected.

Methods based on signal detection theory are the only means currently available of deriving a reliable measure of forecasting skill for binary events. All other measures of skill or accuracy for forecasts summarised as a 2x2 JD are unreliable because they vary with either sample climate or decision threshold, or both (section 8.2).

There are several ways of describing forecasting skill for the 2x2 case using SDT. One of these is to give values of the indices d’ and β. d’ is a measure of discrimination capacity, in the sense of the propensity of the system to give different forecasts before different events. In terms of the SDT model (Appendix 2) it is the separation of the means of the “noise alone” and “signal plus noise” distributions, when these are Gaussian *and of equal variance*. β is an index of the location of the implied decision threshold used to generate the forecasts (see section 8.1.8).

Alternatively, the SDT area index Az can be calculated from d’, using the formula (for the equal variance case) $z(A) = d'/\sqrt{2}$, where z(A) is the normal deviate value corresponding to a cumulative probability of Az (Swets and Pickett 1982, p35).

The procedure recommended for the AIFS QPF verification is to generate the series of 2x2 JDs by thresholding on successively higher rainfall range boundaries, until one of the cell frequencies becomes zero. For each 2x2 JD with non-zero cell frequencies calculate d’, β and Az as outlined above.

Recommendation: that the SDT-based indices d’, β and Az be calculated for rain forecasts as outlined.

Communication with users

Output from RAINV and the AIFS rain verification module is used in some Regional Annual Reports, by Bureau management and forecasters, and may be used for research, for example for comparison with rain forecasts from numerical models. As far as is known there are no external users of these particular forecasts, apart from whatever use may be made of information in Annual Reports.

Output Format Type A:

Diagnostic output for individual forecasters, forecasting teams, numerical modelers and the developers of Model Output Forecasts (MOF)

Information for forecasters and forecasting teams

Forecasters and forecasting teams need to monitor their performance in real time. It is therefore desirable that the outcome of the most recent forecasts be available at the start of the forecaster's shift, as a tabulation of forecast against observed categories, and as a summary of performance over a time period long enough to give stable estimates of skill. This period should be at least 30 forecasts and preferably at least 50, but this may vary between locations. As a minimum, no cell of the 2x2 rain/no rain contingency table should have a frequency of less than five.

Measures of two aspects of forecasting performance should be provided. These are skill and implied decision threshold, ie the level of certainty at which the forecast changes from below to above a category boundary. In order to keep to a minimum the amount of information to be absorbed by forecasters in an operational setting, measures of these parameters should be presented only for the rain/no rain case, with values for other category thresholds available as an option.

An appropriate measure of skill in the 2x2 case is the SDT measure d' (see section 9.4(?) for details). d' is considered preferable to traditional measures such as percent correct, Hansen and Kuipers' score, etc, because it is relatively unaffected by either sample climate or decision threshold. d' ranges from zero, indicating completely uninformative forecasts, to +T. For practical purposes d'=3.0 probably represents the upper level of skill that is likely to be encountered in practice. Values around 1.0 to 2.0 are typical of 24-hour forecasts of rain.

As a measure of implied decision threshold the SDT measure β is appropriate. β is a likelihood ratio, and is described in section 9.4 (Appendix). Likelihood ratio may be an unfamiliar concept to forecasters, who in general are more comfortable with uncertainty expressed in terms of probability. β can be transformed to a probability using Bayes rule (section 9.4), but is theoretically preferable to probability because it does not depend on the (sample) climatological probability of rain.

Recommendation: That values of d' and β for individual forecasters most recent adequately large sample of forecasts be available at logon, together with the outcome (forecast and observed categories) of their previous forecast.

Information for numerical modelers and developers of MOF guidance

Where rain forecasts from numerical models or statistical guidance are produced as quantitative precipitation forecasts for point locations, DO analysis and SDT/ROC-based verification measures are appropriate.

In a research environment verification is often tailored to specific problems. Subject to the assumptions of stationarity and independence, the joint distribution of forecasts and observations contains all the information relevant to assessment of forecast quality, and can provide a common framework for diverse approaches to verification.

SDT/ROC-based measures of forecast quality are particularly appropriate when forecasts are compared for different locations or time periods, as these measures are independent of sample climate and decision threshold. Thus comparisons of RAINV forecasts with forecasts produced by other means should use d' and β or Az.

Output Format Type B:

Basic information for weather services users, the media, Bureau management and Government

Selection of a single summary parameter to communicate skill in QPF is not straightforward, for reasons of dimensionality outlined in section . There is no single quantity that adequately describes all aspects of skill.

The capacity to discriminate between rain and no rain is probably the most important aspect of skill in this case, for which the SDT measures d' and β have been recommended above. It seems likely, however, that d' and β might appear too technical for non-specialist users. An alternative is the ROC-based measure Az (section 9.4), which, as with d', is independent of sample climate and threshold probability. Az has some appeal in this context because it can be interpreted as expected proportion correct *when the sample climatological probability of rain is*

0.5. Az is readily calculated from d'. It could be described for non-specialists simply as a skill index, and expressed as a percentage. Users can "calibrate" Az with the information that 50% represents no skill and 100% perfect skill.

The actual percent (or proportion) correct (PC) is the measure which historically has probably been most used as a convenient summary of the accuracy of categorical forecasts, in spite of being frequently and cogently criticised by meteorologists and in other fields where assessment of skill in discrimination is required (for example see Murphy 1996 and Swets 1996, p4).

Characteristics of PC are described in section 8.2.2.

Annual Reports

The level of detail required in Annual Reports is unclear. In reporting the results of the RAINV forecasts as in other verification data there is wide variation between Regions.

It must be open to doubt whether the full output from RAINV is required for annual reporting purposes. One 8x8 table and the 2x2 rain/no rain table, for the forecasts alone, the error distribution in ranges, and the table of statistics already provided would probably be sufficient.

Management

Management needs to monitor the quality of forecasts, on a short- and long-term basis, so that corrective action can be taken on deficiencies and improvements can be locked in place. There is also a need to monitor the performance of individual forecasters for similar reasons; so those individuals with lower performance can be helped and to learn from the high performers.

For the purpose of monitoring trends in forecasting skill, it is desirable to maintain the "traditional" output of RAINV as described earlier in this section. Availability of time series output in AIFS should be particularly interesting in this regard.

For individuals, values of Az for the rain/no rain case calculated over at least 50 forecasts seems likely to give a reasonable estimate of skill. The number of forecasts may need to be increased with experience in different locations, the basic requirement being that stable estimates are needed of all four cells in the rain/no rain JD.

Further comments on RAINV

Operational issues

The non-operational nature of the forecasts may tend to lower their priority in the view of some forecasters, so there is a possibility that in times of heavy workload they may be forgotten or formulated with only brief consideration. Anecdotal evidence also suggests that there may be occasions when the forecast has been overlooked but done at a later time, sometimes well into the validity period of the forecast or even after it has ended. Forecasts prepared under these circumstances may not give a reliable indication of "state of the art" skill in quantitative precipitation forecasting. It is therefore recommended that the RAINV forecasting form in AIFS be incorporated in the operational OWR forms, rather than appear as a separate form, and if possible be made unavailable for changes after the start of the forecast validity period.

Recommendation: that the form for RAINV forecasts be incorporated into operational AIFS forms for standard OWR issue times, to reduce the risk that they will be overlooked or done after the start of the validity period of the forecast.

Sample size

A Regional meteorologist made the point that 30 years is considered the minimum period for stable climatological statistics for rainfall. It may be unrealistic to expect statistically stable results for verification of rain forecasts in a shorter period. It is therefore important that good quality verification data sets for quantitative rain forecasts be produced and archived with the same concern for quality control as the actual rainfall observations themselves.

Fire weather verification;

There are broadly four types of fire weather forecasts. These are

1. Detailed forecasts of temperature, dew point/humidity and wind usually issued twice daily to Government authorities with responsibilities for fire management,

2. Fire weather warnings issued as required when weather conditions are expected to be conducive to the rapid spread of fires,

3. Operational forecasts issued to fire authorities for going fires, and

4. Fire danger ratings issued with routine public weather forecasts as risk categories (low, moderate, high, and extreme).

Only the first two of these types of forecast are currently verified in AIFS and the second only insofar as the fire danger ratings are verified and warnings are issued when thresholds on fire danger ratings are exceeded. The others are verified manually and reported in fire weather reports prepared at the end of each fire season. Only AIFS verification is considered in this Report.

Pre-AIFS system.

Regions without AIFS verify routine fire weather forecasts using methods appropriate for the individual weather elements (temperature, dew point, wind speed and direction) and fire danger ratings. Verifying data is taken from official Bureau observing sites. At the end of the season some verification data are presented in reports, focussing on fire weather warnings. Detailed case histories of major fires are prepared.

While the processes have been automated to varying degrees in different regions, a considerable amount of manual data entry and analysis has been required.

There is considerable variation between Regions in the verification data provided in end-of-season Reports.

AIFS Fire Weather verification module

Detailed specifications for the AIFS fire weather verification system are available in the document **FVS (Fire Weather) Help: Fire Weather Verification Module**.

In outline, the AIFS system has five data displays. These show forecast and observed values for weather elements and fire danger ratings (grasslands and forests) in tabular and graphical formats, observations from METARs and as calculated FDRs in tabular and graphical formats, and verification measures in tabular format.

Access to forecasts and observations in the AIFS database enables rapid feedback. Verification can be available as soon as the relevant observations are in the database.

A raw data display is available showing forecast and observed values for each weather element and errors (F-O). Separate tables are provided for each element, for each station. Tables are also provided for grasslands and forest FDRs. The raw data can be displayed as scatterplots or as time series.

Verification statistics provided are bias, MAE, RMSE and a histogram of numbers of errors in three ranges.

Data for the previous day is available from METARs for each weather element and for FDRs.

Comments

The AIFS system as specified provides a sound foundation for verification of routine fire weather forecasts. A comprehensive distributions-oriented verification system could be developed on the basis of the current system, with the raw data displays forming the basic verification data sample (VDS).

Graphics displays

Inspection of graphics displays indicates that scatterplots in the Fire Weather FVS are sometimes difficult to read when fire danger indices are plotted, due to the highly skewed nature of the data. This tends to give a high density of data points at low values, and a few points at high values. The usual solution to this problem is to transform the data so that it is more or less normally distributed. The appropriate transformation is a matter for investigation. A good starting point would be the discussion of transformations to normality in Graedel & Kleiner (1985).

An alternative approach to display of skewed data is to use box plots, which may be preferred to use of transformations because box plots show the actual rather than transformed, data. In the case of FDRs the horizontal axis could represent the categorised forecast FDRs (low/moderate/high/extreme) and the corresponding box plots show the medians, upper and lower quartiles, etc for the observed FDRs for each forecast.

Recommendation: That use of transformations to normality and box plots be investigated to enhance the clarity of graphical displays of skewed data such as fire danger ratings.

When a linear model is fitted to scatterplots it is useful to provide the parameters of the model (slope and intercept). It is also desirable to provide the standard errors of these estimates, t values and the correlation coefficient, and if possible without degrading the clarity of the graph, the upper and lower 95% confidence bounds on the regression line.

Recommendation: That where a linear model is fitted to data, the parameters of the model and other statistical information as above be provided.

Statistical verification measures used in the AIFS Fire Weather verification output.

With regard to the specific verification measures used in the current AIFS, comments made in earlier sections of this Report and in the Appendix apply. The measures used are bias, RMSE and MAE, and the correlation coefficient is provided on scatterplots. Three-category distributions for errors are also provided.

These verification measures are widely used and should continue to be calculated. The properties of RMSE noted in the section on temperature verification need to be appreciated, particularly its dependence on bias and on the variance of the forecasts and observations. It is desirable to provide, in addition to RMSE, components of the partition of MSE as outlined in Murphy (1996). As with temperature forecasts, the Bayesian Correlation Score (Krzysztofowycz 1992) and the LEPS score (Potts et al 1996) warrant investigation when resources permit.

A distributions-oriented framework for fire weather verification

Following the outline of distributions-oriented verification in the Appendix to this Report, the basic data structures are the verification data set (VDS) and the joint distribution of forecasts and observations (JD) derived from the VDS.

The verification data set

In essence the VDS is a sequence of matched (forecast, observed) pairs, together with any associated covariates. The raw data displays provided in the AIFS fire weather verification system provide this sequence, and include the errors (F-O) and can be displayed as a time series.

In verification studies it is sometimes useful to be able to look at subsets of the VDS, selected on the basis of a threshold or some other kind of criterion on the data. For example, to select for a VDS only those occasions on which the FDR was greater than a given value, or to select the specific dates on which fire bans were issued. (This capability is available in spreadsheet programs such as Excel as “filters”.)

Recommendation: That the capability to select subsets of the VDS be provided.

A second useful facility would be to be able to download the VDS to standard spreadsheet programs, either in full or as a subset as described above. The range of analysis and display options would thereby be greatly increased.

Recommendation: That an option to download the VDS to standard spreadsheet programs be provided in the AIFS fire weather verification system.

The joint distribution

The JD is derived from the VDS as a contingency table in which the rows correspond to the forecasts and the columns to observations. If the forecasts are represented by f and the forecasts by x, then the elements of the table p(f,x) represent the joint distribution of forecasts and observations (Murphy and Winkler 1987). Under the assumption that the time series is stationary and independent the JD contains all the non time-dependent information that is relevant to the assessment of forecasting performance.

In a DO analysis, insight into forecasting performance is obtained by factoring the joint distribution p(f,x) into conditional and marginal distributions. These factorisations are

$p(f,x) = p(f).p(x|f)$ and

$p(f,x) = p(x).p(f|x)$.

In (i), referred to by Murphy & Winkler as the calibration/refinement or CR factorisation, p(f) is the marginal distribution of the forecasts and the p(x|f) are the distributions of observations for each forecast.

In (ii), the likelihood/base rate or LR factorisation, p(x) is the marginal or climatological distribution of the observations and the p(f|x) are the distributions of the forecasts for each observation.

Further discussion of the DO approach to verification is in the Appendix to this report and in Murphy & Winkler (1987), Murphy et al (1989), Murphy (1995, 1997), and Brooks & Doswell (1996)

Temperature and dew point

Derivation of the JD and its factorisations for temperature and dew point forecasts prepared for fire weather purposes should follow the process outlined in section 1 of this Report dealing with verification of temperature forecasts.

Wind: some problems

There are some particular problems in verification of wind forecasts. Among these are

The vector nature of wind.

Since users generally consider direction and speed separately, it seems preferable to verify these aspects separately, rather than attempt to treat wind explicitly as a vector.

Representativeness of observations.

Bushfires occur in a wide variety of terrain, in which local effects induced by the terrain and by the fire itself can cause great differences between wind at a fire and winds observed at standard observation sites. In verification of wind forecasts in AIFS it is implicitly assumed that the forecasts and verifying observations refer to the same site, in general an official Bureau observation site, so this issue will not be pursued further in this Report. However it is likely that forecasters seeking to optimise the usefulness of their forecasts to the fire authorities are biasing forecasts to some extent towards conditions considered more representative of the whole forecast district, or towards the “worst case”. There is little that can be done about this in a formal verification system. It contributes to an impression of “overforecasting” in fire weather forecasts.

Forecasts given as ranges of direction and speed, rather than as specific values.

This practice is a realistic response to the natural variability of wind, but presents a problem for categorisation of the forecasts in extracting the JD. For example, should forecasts of NW-NE winds, intended to indicate a trend with time from NW to NE during the forecast period or sometimes just an indication of a variable wind field, be verified as northerly (the mid point), or in some other way? Should wind speeds forecast as 15-25 km/h be verified as the mid-point of the range or one or the other end-points? There is no completely satisfactory solution to this problem. Whatever rule is used to “truncate” the forecasts should be explicit and consistently applied. A reasonable approach would be to use the mid-point for directions. The high end of the forecast range should be used for speeds, since in general fire authorities will consider the worst case as the forecast for planning purposes.

Forecast periods divided into sub-periods of variable duration.

Wind changes can arrive at any time, so the forecast period is often split into “before change” and “after change” sub-periods. The problems for verification in general are to incorporate the variability into the JD, and to take account of the fact that the change will not usually arrive at exactly the forecast time. The latter problem gives three periods to consider; before the forecast or actual time of arrival (depending on which is earlier), between the forecast and actual time of arrival, and after the forecast or actual time of arrival (whichever is later). All these times of course are variable. Again, there appears to be no satisfactory general approach to this problem, beyond case-by-case examination. For practical purposes in a general verification framework, the wind at the time of the forecast maximum temperature is probably most appropriate for verification in AIFS, but as with the other issues above, the underlying complexity needs to be borne in mind.

Wind direction.

The dimensionality of the JD for wind direction is unmanageable unless forecasts are categorised so as to reduce their number. The four major directions (N, S, E, W) plus intermediate points (SW, etc) plus calm gives nine possible forecasts. As noted above, forecasts are often given as a range (NW-W, etc), which gives seventeen possible forecasts if only adjacent directions and eight distinct points are allowed in the range The allowed wind direction forecasts would be N, N-NW, NW, NW-W, W, W-SW, SW, SW-S, S, S-SE, SE, SE-E, E, E-NE, NE, NE-N and calm.

Wind observations are normally given in degrees to the nearest 10, rather than as N, NW, etc, giving 36 possible values for observed wind direction, plus calm. With 17 possible categories for forecast direction, there are 629 cells in the JD. As the length of the fire weather season is usually about 100 days there will obviously not be enough data to estimate all the p(f,x).

To reduce the number of cells without losing too much detail the forecast and observed directions should have no more than four categories (plus calms).

A possible form for the JD for wind direction would then be:

Observed deg→ Forecast ↓	350-070	080-160	170-250	260-340	Calm or L&V	p(f)
N, N-NE, NE, NE-E						
E, E-SE, SE, SE-S						
S, S-SW, SW, SW-W			p(f,x)			
W, W-NW, NW, NW-N						
Calm or L&V						
p(x)						

N in the lower right cell is the total number of (f,x) pairs in the VDS, so that the absolute frequencies can be calculated for each cell in the JD if required.

The CR and LBR factors should also be available. These would take the form

CR factorisation

Observed deg→ Forecast ↓	350-070	080-160	170-250	260-340	Calm or L&V	p(f)
N, N-NE, NE, NE-E						
E, E-SE, SE, SE-S						
S, S-SW, SW, SW-W			p(x|f)			
W, W-NW, NW, NW-N						
Calm or L&V						

where p(x|f) = p(f,x)/p(f).

LBR factorisation

Observed deg→ Forecast ↓	350-070	080-160	170-250	260-340	Calm or L&V
N, N-NE, NE, NE-E					
E, E-SE, SE, SE-S					
S, S-SW, SW, SW-W			p(f|x)		
W, W-NW, NW, NW-N					
Calm or L&V					
p(x)					

Where p(f|x) = p(f,x)/p(x)

Wind speed

In the current AIFS fire weather verification system the distribution of wind speed errors is given in three categories. A full DO analysis requires the JD of forecast and observed values, plus the CR and LBR factorisations, as with wind direction.

It is understood that most if not all Regions provide fire authorities with forecasts of wind speed at the time of maximum temperature as a single value (rather than a range). If a range is given then the high end should be used for verification purposes.

The category ranges for wind speed may need to be varied in the light of experience. As a starting point 10 km/h ranges are suggested.

The JD would then take the form

Observed → Forecast ↓ km/h	≤10	11-20	21-30	31-40	≥41	p(f)
≤10						
11-20						
21-30			p(f,x)			
31-40						
≥41						
p(x)						N

CR factorisation

Observed → Forecast ↓ km/h	≤10	11-20	21-30	31-40	≥41	p(f)
≤10						
11-20						
21-30			p(x|f)			
31-40						
≥41						

LBR factorisation

Observed → Forecast ↓ km/h	≤10	11-20	21-30	31-40	≥41
≤10					
11-20					
21-30			p(f|x)		
31-40					
≥41					
p(x)					

Fire danger ratings.

The fire danger rating as calculated is a continuous variable which is related to the rate of spread of a fire and is provided to fire control authorities in this form. It is however issued to the public as a verbal rating of fire danger (low, moderate, high and extreme).

The numerical fire danger ratings should be verified as continuous variables, using scatter diagrams and linear regression.

The JD, CR and LBR distributions should also be provided for fire danger ratings. The category boundaries should be those corresponding to the verbal fire danger ratings.

These three distributions should be provided for both the grasslands and forest fire danger ratings as appropriate.

Verification measures for fire weather forecasts

Temperature and dew point

The verification measures currently provided in the AIFS fire weather verification system are bias, mean absolute error, root mean square error and three-category error distributions. These measures are widely used and accepted and should continue to be provided. Comments made on these measures in section 1 of this Report apply in this section also.

Where scatterplots are provided the correlation coefficient is available. As recommended above, the estimated parameters (slope and intercept) of the fitted linear model should also be available, together with standard errors of these estimates, t values, and if possible without degrading the clarity of the graph, the upper and lower 95% confidence bounds on the regression line.

Wind and fire danger ratings

For the numerical fire danger ratings suitable verification measures are MAE and those based on the linear model, supplemented with scatter diagrams.

With regard to summary measures of skill for multicategory forecast/observed contingency tables like those for wind speed and direction and the verbal ratings of fire danger, comments made in section 2 on the rain forecast verification system apply. In brief, there is no completely satisfactory single measure of skill for multicategory contingency tables (notwithstanding the occasional use of multicategory forms of the Hansen and Kuipers and Heidke scores). The reason lies in the multidimensional nature of forecast quality and the inevitable loss of significant detail when many dimensions are reduced to one (Murphy 1997).

As with the RAINV forecasts, it is recommended that symmetric kxk contingency tables be collapsed to a sequence of k-1 2x2 tables by thresholding on successively higher category boundaries, and the SDT indices (d',β) be calculated for each table.

This approach provides measures of the capacity of the forecasting system to discriminate between weather states above and below each selected threshold.

The more familiar measures POD, FAR, CSI, Hansen & Kuipers or Heidke scores (etc) are not recommended for reasons detailed in the appendix, essentially that they are subject to variation associated with either variations in sample climate or threshold probability, or both, and have unrealistic dependence on threshold probability. These scores are however widely used. Recognising that people are likely to want to calculate values for POD, FAR etc for comparison purposes, it is probably worthwhile providing them.

Recommendation: That the primary measure of skill for wind speed and direction and fire danger ratings be in the form of a table of values of d' and β for successive category boundaries on the joint distribution.

Communication with users

Output format type A:

Diagnostic output for individual forecasters, forecasting teams, numerical modelers and the developers of Model Output Forecasts (MOF).

Forecasters and forecasting teams need real time feedback on performance. An appropriate form of feedback for this group is the most recent forecast and observed fire danger ratings, and summary verification measures over the past 30 forecasts. MAE for the numerical fire danger ratings would be adequate as a basic “first look” statistic, with an option for display of forecast/observed scatterplots and the fitted linear model. This information should be available to individual forecaster either at logon or as part of the fire weather forecasting module in AIFS.

For other users in this group the full DO verification data set should be available for each of the forecast parameters.

Output Format Type B:

Simpler output for weather services users, the media, Bureau management and Government.

The most operationally significant aspect of fire weather forecasting performance is the level of skill in forecasting around the threshold for fire weather warnings. An appropriate basic display of verification data is the 2x2 JD as raw frequencies for this threshold, ie forecast and observed fire danger ratings above and below the warning threshold. As a single index of skill the SDT measure Az is recommended, presented as a “skill index” with the information the 50% represent no skill and 100% perfect skill.

If more detail is required forecast and observed numerical fire danger ratings plotted as a scatter diagram with details of the fitted linear model may be suitable.

TAF verification;

A number of different approaches have been tried to verification of Terminal Aerodrome Forecasts in different countries. Recent examples are Leigh (1995), Gordon (1989) and Stanski et al (1989) and references therein.

TAFs tend to be complex, attempting to provide a high degree of spatial and temporal detail. Adding to the complexity for verification purposes, the available observational data in METARs and SPECIs are not adequate to fully specify the fluctuating conditions described in forecasts by INTER and TEMPO.

In outline, this Report proposes implementation of a distributions-oriented approach to verification of TAFs, with flexibility in extracting subsets of interest as verification data samples, and proposes an analysis in which INTER and TEMPO appear as separate forecasts, in addition to the current AIFS analyses in which INTER and TEMPO are included in a 2x2 joint distribution. The specification of observations corresponding to TEMPO, INTER and PROB follows the scheme proposed by Keith and currently implemented in the AIFS TAF verification prototype.

Pre-AIFS practice

Verification of TAFs in the Bureau of Meteorology from 1973 to the advent of AIFS followed the system outlined by Shanahan (1973). It is done for capital city airports only, in six-hour blocks. The forecasts and observations are classified into “No alternate”, “Alternate”, INTER and TEMPO. A 4x4 joint distribution of forecasts vs observations is produced from which a number of measures are calculated. Data extraction is done manually.

There are a number of deficiencies in this system.

The need for manual processing has limited the number of locations and weather verified, and subjectivity can enter in deciding which of INTER. TEMPO, or alternate is an appropriate classification for the observations. There are significant irreplaceable missing data for some sites. It has not been possible to input pseudo data or ignore the missing data in the desktop computer program used to do the calculations, and significant further work would be required to change the program to cope with leap years and otherwise run effectively under unix.

Proposed system for AIFS

Details of the proposed system are set out in the document **Software Requirements Specification AIFS Aviation Verification System Date: 24/2/1998** (http://servb.ho.bom.gov.au/fvs/aviation/html/pages/). A prototype system exists but all features of the specification were not implemented at the time of writing.

The proposed AIFS system automates the generation of reports, accessing METAR/SPECI observations from the central climate archive (ADAM), and TAFs from the NMC AIFS database.

Initially, the following weather elements will be verified in AIFS:

- cloud amount and base,
- visibility
- thunder
- wind speed and direction

Output consists mainly of 2x2 contingency tables in which the forecasts and observations are collapsed into binary form, above and below user-specified thresholds for each of the above elements. The prototype includes a total distribution in which the category totals for cloud, visibility and thunder are summed, providing an indication of skill in forecasting operationally significant weather

Wind is treated separately.

The system will be expanded later to include the verification of other weather elements

The user will be able to

generate numerical reports,

generate graphical reports,

Specify the following parameters:

location of the aerodrome,

forecaster,

time period,

forecast hours (hours of the day),

forecast lead time, and

verification method.

threshold values for cloud, wind and visibility.

The system will be modular, providing for later addition of other verification methods.

As with the earlier method, the basic output is a 2x2 contingency table of forecast against observed weather, of the form of table 4.1

Table 5.1: General form of table output from AIFS prototype TAF verification system

FORECAST OBSERVED	BELOW MINIMUM	ABOVE MINIMUM
BELOW MINIMUM	X	Y
ABOVE MINIMUM	Z	W

Treatment of INTER, TEMPO and PROB in the AIFS prototype

Each hour has one unit to contribute to the contingency table. A profile of the weather conditions observed for each hour is built up, and the unit is allocated to cells of the table in proportions that are determined by the length of time that below minimum conditions were observed, and the forecast issued, whether INTER, TEMPO or PROB. The scheme is described in detail in section 3.4.6 of the specifications.

TTF

TTF are not verified in the current Specifications for the AIFS module.

Code Grey

Code Grey forecasts are not verified in the current Specifications.

Wind

Wind is verified separately from the other elements, and speed and direction are treated individually for both the actual forecasts and persistence. Statistics are generated on the basis of user-supplied thresholds and tolerances.

Visibility

For visibility the misses and false alarms (Y and Z respectively in table xx) are further stratified according to whether fog or mist was in fact observed, whether other weather was observed, or whether there were below minimum conditions from other causes.

Verification measures

Verification measures ("scores") proposed for the TAF verification module are, in terms of the elements of table xx,

POD ("probability of detection") = $X/(X+Y)$ (8)

FAR ("false alarm ratio") = $Z/(X+Z)$ (9)

HKS ("Hansen & Kuipers' score") = $X/(X+Y) - Z/(Z+W)$ (10)

HSS ("Heidke skill score") = $(PC-E)/(1-E)$, (11)

> where PC is proportion correct, equal to $(X+W)/N$, and E is the proportion correct expected by chance subject to the constraint that row and column totals are unchanged. Under this constraint E = (X+Z)(X+Y)/N^2 + (Z+W)(Y+W)/N^2.The HSS is a form of proportion correct adjusted for chance.

RF1 ("relative frequency of below minimum conditions") = $(X+Y)N$ (12)

Bias = $(X+Z)/(X+Y)$, (13)

ie the number of forecasts of below minimum conditions per actual occurrence.

Some comments on verification measures

Critical success index

POD, FAR, bias and CSI ("critical success index") often appear as a group, particularly in verifications of forecasts of rare events, because they can all be calculated without using the frequency W in Table 5.1. CSI should probably be included in the scores calculated, because it is widely used and values may be desired for comparisons. CSI is essentially a sample estimate of the probability that the event was both forecast and observed given that it was either forecast or observed. The formula in terms of table 5.1 is

CSI = $X/(X+Y+Z)$

Mason (1989) examined the properties of CSI in some detail. Some caution should be exercised when comparing different forecast sets using this score, and most others, because its value is affected by sample climate and decision threshold.

Recommendation: that values for CSI be calculated for the 2x2 JD.

Problems with current verification measures

While the verification measures proposed for TAFs are entrenched in current practice, there are problems with interpretation of some due to dependencies on decision threshold and sample climate.

POD, FAR, CSI, HK and HSS all depend on decision threshold. POD and FAR can vary through their whole range (0,1) as a consequence of variations in decision threshold alone. HK and HSS have maxima at specific values of threshold probability. In the case of HK this is the climatological probability of the event. The optimising threshold probability for HSS and CSI depends on the current value of the score. In general it lies between the sample climate and 0.5 for both these scores

FAR, CSI and HSS also depend on sample climate. FAR decreases as the sample frequency of the event to be forecast increases (so long as decision threshold and overall skill remain constant), because there are fewer non-occurrences to be incorrectly forecast as occurrences ("false alarms"). CSI increases with sample climate (Mason 1989). HSS in general has a maximum at a particular value of the sample climatological probability of the event, when other factors are constant.

These considerations make interpretation of absolute values or differences in these scores somewhat problematic unless sample climate and the relevant threshold probability are known.

Verification measures based on SDT/ROC are more reliable and are described in section 5.5 below.

Confidence intervals

Seaman et al (1996) noted that estimation of sampling variability is rarely attempted for verification measures, making it impossible to assess the significance of differences, and proposed that confidence intervals for scores should be estimated wherever possible. The methods proposed to calculate confidence intervals assumed that a stationary process generates the sequence of values, and that the variability is due only to random sampling. In view of the usual presence of the uncontrolled sources of variation noted in the preceding paragraph, it may be questionable whether standard confidence intervals on these scores are likely to be realistic.

Nevertheless, on the general principle that sample estimates calculated from empirical data should be accompanied by estimates of the likely error or sampling variability, values of the scores presented in AIFS should be accompanied by confidence intervals using the methods discussed in Seaman et al (1996).

Recommendation: that values of the scores presented in AIFS should be accompanied by confidence intervals using the methods discussed in Seaman et al (1996).

Comments

General

The specifications provide for a sound basic system that will produce valid results. Inspection of output from the prototype suggests that there may be a few bugs in the current implementation. Also some aspects of the specifications are not fully implemented in the prototype, for example the capacity to verify subsets of the TAF validity period. In view of the importance of the results for the Bureau and the aviation industry, and the novelty of the points allocation system, it is most important that thorough testing be undertaken before it is made available as an ‘official’ system. Detection of significant shortcomings by the industry after release would have a bad effect on the Bureau’s credibility.

Recommendation: that the AIFS system be exhaustively tested on validated data sets before use as the Bureau’s official system for TAF verification.

Some extensions are proposed below to produce information that has been requested by the industry, and to implement a distributions-oriented verification framework.

TTF and Code Grey

The accuracy of TTF is of interest to the airline operators, and it is recommended that a system verifying TTF be developed, for the purpose of monitoring the quality of a product of importance to customers.

Similarly, Code Grey forecasts have been introduced to provide an additional service to airline operators, and it is desirable that the quality of this service be assessed regularly.

Recommendation: That a system for verifying TTF and Code Grey forecasts be developed.

Treatment of INTER and TEMPO

One of the difficulties in verifying TAFs has been treatment of the terms INTER and TEMPO. These terms are used to indicate significant variations of a temporary or intermittent nature.

Full definitions can be found in the Australian Aeronautical Services Handbook, Ch 6, pars 41-46.

The METAR and SPECI observations do not have a sufficient time resolution to determine which of TEMPO or INTER was correct. Leigh (1995) removed these terms from consideration altogether on the grounds that they are used relatively infrequently. Leigh considered that the resulting error introduced into expected cost to airline operators was less than 10%. Another approach that has been used in New Zealand (Gordon 1989) and in Canada (Stanski et al 1989) is to treat INTER and TEMPO as probabilities and use the Brier score or ranked probability score as the main verification measure.

Neither of these approaches is entirely satisfactory. With regard to Leigh's approach, INTER and TEMPO do have significant cost and safety implications for operators when they are used, and an error of 10% in costs may not be insignificant. They can not simply be discarded.

Converting TAFs to probability forecasts has some appeal, as it facilitates use of verification methods that are well developed for probability forecasts (although neither of the above papers uses the full power of these methods). It seems desirable, however, to verify the forecasts as they were understood by forecasters, as nearly as possible. If forecasters had known that INTER, for example, would be taken to mean the same as PROB40 they may have forecast differently.

Both Gordon and Stanski et al assigned probabilities to these terms a priori, that is, without any empirical study of the corresponding relative frequency of the forecast event.

The method proposed for AIFS by Keith and described in section 3.4.6 of the Specifications appears to be appropriate and is considered to be closer to the way forecasters would think of these terms. Details of the specific proportional allocation of points could possibly be criticised but it is not unreasonable, and fine tuning details would probably make little difference in the aggregate. The effect of different ways of allocating points could perhaps be investigated further if resources permit.

Perhaps a more serious criticism is that aggregation of INTER and TEMPO with unequivocal forecasts of above and below minimum conditions loses some significant detail; specifically the distributions of actual weather when these kinds of forecasts are issued. It is considered that contingency tables should be available showing the distribution of actual cloud base (>4/8), visibility, thunderstorm occurrence and wind speed and direction when INTER and TEMPO have been used, and a table showing the frequencies with which below minimum conditions are observed for any reason during an hour for which INTER or TEMPO were used. This concept is further developed in section 5.5 below.

Aggregation of contingency tables

It appears that the process of aggregation of contingency tables for cloud base, visibility and thunderstorms by simple addition into a single table to represent operationally significant events may occasionally lead to double-counting. There will be occasions on which two or all three of these elements are below the minimum threshold. If any one element is below its threshold then there are operational implications, regardless of the other elements.

Recommendation: That the possibility of double counting in combining contingency tables be investigated and eliminated if found.

Persistence

The standard of persistence defined in the Specifications is “”the observation immediately prior to or at the commencement of the current TAFs validity period will be taken as the forecast for the current hour”. This appears inappropriate. The point of using persistence is to provide a no-skill forecast as a baseline for comparison with the actual forecasts. TAFs are usually issued more than an hour before the start of their validity period, and forecasting the observation at the start of the period would sometimes require significant, at least non-zero, skill.

A preferable definition would be the latest observation *before the issue time* of the TAF. This is not a particularly stringent no-skill standard for the later hours of the TAF but provides useful information about the quality of the first few hours, where persistence has been found hard to beat.

Averaging skill over lead times

Verifying the whole 24 hours of a TAF as a single unit provides an overall assessment of skill over the whole period. However, skill varies over the 24 hours, raising a question as to the meaning of the average figure. Also, the variation in skill may be non-linear, so the average may not even be a good estimate of the skill at the mid-point.

While the 24-hour average may be required as a summary statistic, the change in skill with time is one of the interesting outcomes of forecast verification. It therefore seems desirable for the verification results to be averaged over periods short enough that the change in skill over the period is negligible. The length of this period is a matter for investigation, but three hours might provide a reasonable starting point.

Recommendation: That verification results be available for the validity period of the TAF in 3-hour blocks as a routine option (in addition to the currently proposed variable forecast age).

Other forecast quantities.

The current specifications do not provide for verification of the temperature or QNH forecasts.

While these quantities are less important to operators than cloud, visibility, thunderstorms and wind, automation makes it possible to verify them with little additional effort. On the general principle that all forecasts should be verified, it is desirable that TAF forecasts of T and QNH be verified.

Real time verification of T and QNH forecasts (at least) would be useful in ensuring that deviations larger than amendment criteria are brought to the forecaster’s attention, and could be included in the AIFS Alerts system.

Recommendation: That the next revision of the TAF verification module includes verification of temperature and QNH, and consideration be given to real time verification of at least these quantities.

Verification of PROB forecasts

PROB forecasts are included in the scheme for allocating points to the contingency table (section 5.2.1 above), but this does not give any indication of the reliability of these forecasts as probabilities. Events stated to have a probability of p% should occur on p% of occasions. It therefore appears desirable to assess the reliability of PROB30 and PROB40 forecasts by extracting the number of occasions on which they were, and were not, followed by the corresponding event.

Recommendation: That the reliability of PROB30 and PROB40 forecasts be assessed by extracting the relative frequencies of corresponding forecast events.

Distributions-oriented verification for TAF

Verification data sample

The verification module as it stands derives joint distributions of the forecasts and observations directly from stored TAFs and METAR/SPECI data. There is an option to display the list of

TAFs and METAR/SPECIs from which the joint distribution (JD) is drawn, which is useful and should be retained. There appears however to be no provision for display of the verification data set as such, as there is an other AIFS FVS modules. Explicit availability of the VDS is desirable for the following reasons:

- Inspection of the VDS makes it easier to examine individual pairs of forecasts and observations. This can reveal features that are lost in the JD, for example near misses and 'justified' false alarms, when the weather parameter is close to the operational threshold.
- INTER, TEMPO and PROB can be displayed as distinct forecasts in the VDS, which should facilitate re-analysis if there is an interest in the operational implications of these terms for the weather.
- It facilitates transfer of the data to other systems.
- It facilitates identification of outliers in the data.
- It facilitates examination of trends and serial dependence in the time series.

The industry is interested in the details of occasions when below-minimum conditions were either forecast or observed, so there should be a capacity for selection of hours to go into the VDS on the basis of thresholds set by the user on the forecasts and the observations. Operationally significant thresholds vary between locations, between aircraft types and possibly for other reasons, so the criterion for selection needs to be set by the user. The criterion would be a boolean expression something like 'select if *either* the forecast cloud is greater than 4/8 below 1400 ft *or* forecast visibility is less than 2000m *or* crosswind component is greater than 15 kts *or* there is a thunderstorm within 5km, *or* any of these thresholds are met in the METARs or SPECIs'. Specific values for the thresholds should be input by the user.

In the absence of some such selection facility the great majority of cases in the VDS will be correct forecasts of above minimum conditions, typically 99% or more. From the industry's point of view this simply loads the data with a lot of 'fair weather' cases which make the Bureau's statistics look better but do not help airlines make an assessment of the confidence to be placed in the forecasts for risk management purposes in marginal situations. (From the point of view of assessment of pure skill these occasions are important, because every non-occurrence of below minimum conditions is an opportunity to commit a false alarm, and the proportion of these opportunities that are taken up is interesting. They should not simply be eliminated at the outset.)

Structure of the VDS

The basic VDS should be comprehensive, to facilitate different kinds of analysis at a later stage. The following proposed structure for a TAF VDS should be regarded as tentative, to be modified in the light of experience. It has been based to some extent on the Canadian model as described in Canada's submission to the CAeM (2-11 March 1999).

The TAF VDS is inevitably much larger and more complex that for most other types of forecasts due to the variety and level of detail in both forecasts and observations. Among aspects of TAFs which are or may at some stage be of interest for verification are

- time of issue (as distinct from start of validity period) (are morning issues less accurate than others?);
- period of validity in relation to diurnal cycle (derivable from UT times of validity period);
- lead time; whether the hour being verified is early or late in the TAF validity period;
- weather elements (cloud base, visibility, wind, TS etc);
- specific values for these elements;
- start and/or finish times of INTER, TEMPO, FM, PROB and PROB INTER qualifiers;

- operational implications of INTER, TEMPO, FM, PROB and PROB INTER qualifiers;
- nature and values of weather elements qualified by INTER etc statements;
- whether trends forecast or observed are deterioration, fluctuation about an operational threshold, or improvement;
- whether allowance should be made for “near misses” as against major errors.

This list is not intended to be comprehensive. Other issues will be identified, but these serve to illustrate the need for a comprehensive TAF VDS.

A proposed structure

For a specific location and for each standard TAF validity period (which may vary between locations and between Standard Time and Daylight Saving Time) the TAFs, METARs and SPECIs should be decoded and merged into a listing ordered by time of validity. Some TAF qualifiers can have overlapping validity times with the main TAF (eg INTER, TEMPO, PROB and PROB INTER), so both the main TAF and the qualified parameters must be available for the period of overlap.

The above should provide a raw VDS which contains all the available data relevant to TAF verification.

Recommendation: That the raw verification data set be available as an option, for both the actual and persistence forecasts, and for subsets to be selectable on user-defined criteria.

The joint distribution

Output from the present AIFS prototype system provides 2x2 forecast/observed contingency tables, in which the elements are the numbers of hours in which the relevant combination of forecast and observation occurred, and INTER, TEMPO and PROB are treated following Keith’s algorithms. Tables of this kind are provided for the actual forecasts and for forecasts of persistence. The data are categorised on the basis of user-supplied threshold values that will usually be the operational minima for the terminal. These kinds of tables are referred to as joint distributions in DO verification, and should continue to be provided.

The forecasts are verified hour by hour, and the results aggregated into a distribution for the full TAF validity period. While this aggregate JD is useful as a summary, there is likely to be some interest in the deterioration in accuracy from the beginning to the end of the validity period, so an option should be available to provide shorter periods and variable lead times at the user’s discretion. For example, it should be possible to extract a JD for the fifteenth hour of the 0606 TAF, or hours 22, 23 & 24 of the 1818 TAF, etc.

DO verification makes use of the marginal distributions, so marginal totals should be provided in these tables.

Most forecast verification studies using DO methods present this table in terms of relative rather than absolute frequencies (numbers), but this is optional. If the elements are presented in relative frequencies then the total number of forecasts verified must also be specified, so that the original frequencies can be recovered. The present system provides joint frequencies in a list above the contingency table headed Report Totals. There is enough space to include these numbers on the table itself, which seems more convenient.

A suggested format is table 5.3 below

Table 5.3: suggested format for joint frequency distribution

FORECAST	BELOW	ABOVE	
OBSERVED	MINIMUM	MINIMUM	TOTAL

BELOW MINIMUM	X X/N	Y Y/N	X+Y (X+Y)/N
ABOVE MINIMUM	Z Z/N	W W/N	Z+W (Z+W)/N
TOTAL	X+Z (X+Z)/N	Y+W (Y+W)/N	N

Recommendation: That contingency tables include row and column totals, plus the joint relative frequencies.

Preserving dimensionality: an expanded JD

The 2x2 contingency table 5.3 has dimension 3 when the elements are expressed as relative frequencies. That is, the minimum number of quantities required to reconstruct a 2x2 JD is 3. This is a fairly radical compression of the dimensionality of a full JD for a TAF verification. If only five different possible forecasts are considered (above threshold, INTER, TEMPO, PROB, below threshold) then the dimension of a full JD would be 24. There is inevitably some loss of information when 24 dimensions are reduced to 3. (The “true” dimensionality is of course greater than 24, because INTER, TEMPO and PROB can all be qualified by time and/or probability values.) It therefore seems desirable to expand the JD to capture some of this lost detail.

A second reason for seeking a more detailed JD is the interest of airlines in the operational implications of, in particular, INTER and TEMPO statements. The quantity of operational significance is the probability of encountering sub-threshold conditions given that the forecast included INTER or TEMPO.

An expanded JD of the following form might meet these concerns to some degree without a great increase in the complexity of the situation, by showing INTER and TEMPO as separate forecasts but without regard to the specific forecast values following these qualifiers.

The relevant observations are also classified only as above or below user-defined threshold values, so the JD has the form of table 5.4 below.

Table 5.4: Joint distribution of forecasts and observations for TAFs, expanded to include forecasts of INTER and TEMPO.

OBSERVATION→ FORECAST ↓	BELOW THRESHOLD	ABOVE THRESHOLD	TOTAL
BELOW THRESHOLD	a a/N	b b/N	a+b (a+b)/N
TEMPO	c c/N	d d/N	c+d (c+d)/N
INTER	e e/N	f f/N	e+f (e+f)/N
ABOVE THRESHOLD	g g/N	h h/N	g+h (g+h)/N
TOTAL	a+c+e+g (a+c+e+g)/N	b+d+f+h (b+d+f+h)/N	N

Recommendation: That 4 (forecasts) x 2 (observation) contingency tables as described in the text be available as an option.

Factorisations of the JD

DO verification focuses on the two groups of marginal and conditional distributions resulting from factorisation of the JD. Many important aspects of the forecasts, observations, and their relationship can be assessed from these distributions, and they have a direct relationship with forecast value (Appendix). Murphy (1997) lists ten basic aspects of forecast quality and their relationship to the joint, marginal and conditional distributions. These distributions are referred to as performance characteristics.

It is desirable that the marginal and conditional distributions be available, for both the forecasts and persistence. The following outline is given in terms of the 2x2 JD for simplicity.

If the basic contingency table is represented as in table 5.1, following the Specifications document, then using Murphy's notation (and ignoring the distinction between sample estimates and probabilities as is usual in this field),

$X/N = p(f=1,x=1)$ (15

$Y/N = p(f=1,x=0)$ (16)

$Z/N = p(f=0,x=1)$ (17)

$W/N = p(f=0,x=0)$ (18)

There are two ways of factoring the joint distribution into marginal and conditional distributions, referred to in the verification literature as the calibration/refinement (CR) factorisation and the likelihood/base rate (LBR) factorisation (see ***Appendix*** for an explanation of these terms)

CR factorisation

The possible CR factorisations in the 2x2 case, with expressions in terms of table 5.1, are

$p(f=1,x=1) = p(f=1).p(x=1|f=1) = [(X+Z)/N].[X/(X+Z)]$ (19)

$p(f=1,x=0) = p(f=1).p(x=0|f=1) = [(X+Z)/N].[Z/(X+Z)]$ (20)

$p(f=0,x=0) = p(f=0).p(x=0|f=0), = [(Y+W)/N].[W/(Y+W)]$ (21)

$p(f=0,x=1) = p(f=0).p(x=1|f=0) =[(Y+W)/N].[Y/(Y+W)]$ (22)

Only three of the four relationships above are independent, since
$p(x=1|f=1) = 1 - p(x=0|f=1)$, $p(x=0|f=0) = 1 - p(x=1|f=0)$, and
$p(f=1) = 1 - p(f=0)$. It is useful in practice to have all values displayed, to avoid manual arithmetic.

The elements of the CR factorisation can be displayed in tabular form as in table 5.5.

Table 5.5: Elements of CR factorisation, in terms of the elements of table 1 and the conditional probabilities $p(x|f)$

FORECAST OBSERVED	BELOW MINIMUM (f=1)	ABOVE MINIMUM (f=0)
BELOW MINIMUM (x=1)	X/(X+Z) $p(x=1\mid f=1)$	Y/(W+Y) $p(x=1\mid f=0)$,
ABOVE MINIMUM (x=0)	Z/(X+Z) $p(x=0\mid f=1)$,	W/(Y+W) $p(x=0\mid f=0)$,
TOTAL	(X+Z)/N	(W+Y)/N

	p(f=1)	p(f=0)

LBR factorisation

The possible LBR factorisations in the 2x2 case are

p(f=1,x=1) = p(x=1).p(f=1|x=1) = [(X+Y)/N].[X/(X+Y)] (23)

p(f=1,x=0) = p(x=0).p(f=1|x=0) = [(W+Z)/N].[Z/(W+Z)] (24)

p(f=0,x=0) = p(x=0).p(f=0|x=0) = [(W+Z)/N].[W/(W+Z)] (25)

p(f=0,x=1) = p(x=1).p(f=0|x=1) = [(X+Y)/N].[Y/(X+Y)] (26)

The LBR factorisation can be displayed as in table 5.6:

Table 5.6: Elements of the LBR factorisation, in terms of the elements of table 1 and the conditional probabilities p(f|x)

FORECAST OBSERVED	BELOW MINIMUM (f=1)	ABOVE MINIMUM (f=0)	TOTAL
BELOW MINIMUM (x=1)	X/(X+Y) p(f=1|x=1)	Y/(X+Y) p(f=0|x=1)	(X+Y)/N p(x=1)
ABOVE MINIMUM (x=0)	Z/(W+Z) p(f=1|x=0)	W/(W+Z) p(f=0|x=0)	(W+Z)/N p(x=0)

Factorisations of the JD in the 4x2 case

The following table shows the CR factorisation in the 4x2 case.

Table 5.7: CR factors of the expanded joint distribution for TAFs. B = ”below threshold”, A = ”above threshold”, T = “TEMPO” and I = “INTER”.

OBSERVATION→ FORECAST ↓	BELOW THRESHOLD	ABOVE THRESHOLD	TOTAL
BELOW THRESHOLD (B)	p(x=B|f=B) a/(a+b)	p(x=A|f=B) b/(a+b)	p(f=B) (a+b)/N
TEMPO (T)	p(x=B|f=T) c/(c+d)	p(x=A|f=T) d/(c+d)	p(f=T) (c+d)/N
INTER (I)	p(x=B|f=I) e/(e+f)	p(x=A|f=I) f/(e+f)	p(f=I) (e+f)/N
ABOVE THRESHOLD(A)	p(x=B|f=A) g/(g+h)	p(x=A|f=A) h/(g+h)	p(f=A) (g+h)/N

The following table shows the LBR factorisation in the 4x2 case

Table 5.8: LBR factors of the expanded joint distribution for TAFs. BM = ”below minimum”, AM = ”above minimum”, T = “TEMPO” and I = “INTER”.

OBSERVATION → FORECAST ↓	BELOW THRESHOLD	ABOVE THRESHOLD
BELOW THRESHOLD	p(f=B|x=B) a/ N_B	p(f=B|x=A) b/N_A
TEMPO	p(f=T|x=B) c/ N_B	p(f=T|x=A) d/N_A
INTER	p(f=I|x=B) e/ N_B	p(f=I|x=A) f/N_A
ABOVE THRESHOLD	p(f=A|x=B) g/ N_B	p(f=A|x=A) h/N_A
TOTAL	p(x=B) N_B/N a+c+e+g=N_B	p(x=A) N_A/N b+d+f+h=N_A

Recommendation: That the components of the CR and LBR factorisations be available as an option for both the actual forecasts and persistence, in both the 2x2 and 4x2 forms of the joint distribution.

Verification measures

Current verification measures

Verification measures ("scores") proposed for the TAF verification module are listed above (section 5.3) and discussed in some detail in Appendix 8.2.

Note that these scores are defined in terms of the 2x2 JD. Except for RF1, the sample relative frequency of occurrence of below minimum conditions, they are not defined for the 4x2 table (or any asymmetric table). Summary measures of skill suitable for a 4x2 table are described in the nextsection.

Verification measures from SDT

There are measures of skill based on SDT and the ROC which do not have the problems noted above for POD, FAR, etc. Some definitions are in the appendix and more detail in references by Mason (1982a) and Swets (1997). A practical approach is in Macmillan and Creelman (1991). An outline of computation of ROC/SDT measures of skill for the 2x2 case is in Appendix 9.1.5.1. More detail can be found in the references.

If only a single 2x2 JD is available then the recommended indices are d', the separation of the means of the underlying "noise" and "signal plus noise" distributions in units of the common standard deviation, accompanied by β, a measure of the implied decision threshold expressed as a likelihood ratio. d' is preferable to other skill measures, because the variation of d' with threshold and climatology is small.

The measure d' is not completely satisfactory because of an underlying assumption that a certain parameter of the SDT model (the slope of the ROC on "binormal" axes) is equal to unity (Swets 1986). Satisfactory assessment of skill for forecasts of a binary event requires that forecasts be made at at least two different decision thresholds (preferably more), so that two parameters of the forecasting system's ROC can be estimated. Probabilistic forecasts are an example of such forecasts, or forecasts done as risk ratings (eg nil, low, moderate, high risk). If forecasts at more than one threshold are available, then two ROC-based measures of skill are recommended The first of these is (Δm,s), where Δm is the separation of the means of the underlying distributions in the SDT model and s is the ratio of the variance of the non-event distribution to that of the event distribution. The second is Az, the area under the fitted ROC on probability axes. Az is recommended by Swets as the best available single number index of skill. In the 2x2 case a simple relationship connects d' and Az (Swets and Pickett, 1982, p35).

Following a suggestion by Williams (personal communication) indices of skill analogous to the SDT measures can be calculated for the 4x2 JD described above (section 5.2.2) by taking INTER and TEMPO as implying levels of risk of below minimum conditions. This is similar to the approach criticised earlier in this Report of assigning arbitrary probability values to INTER and TEMPO, but differs in that it is not necessary to assign specific probabilities, only an order in terms of an implied probability of below minimum conditions. The calculations are essentially identical to those described in Mason (1982a) for probabilistic forecasts. The measure recommended in this case is Az, the area under the fitted ROC on linear probability axes.

Computer programs produced by C.E. Metz and others to perform the necessary calculations are available by anonymous ftp from ftp://random.bsd.uchicago.edu/roc/. These programs also give maximum likelihood estimates of the variance of fitted parameters from which confidence intervals can be found.

Recommendation: That values of d', β and Az be calculated for all 2x2 forecast observed contingency tables and presented with estimated confidence intervals, and that Az be calculated for the 4x2 JD.

Communication with users

Output format type A:

Diagnostic output for individual forecasters, forecasting teams, numerical modelers and the developers of Model Output Forecasts (MOF);

This group may have an interest in a wide variety of TAF verification data. A “top level” set of output which could be available as the default option comprises the 2x2 and 4x2 JDs for all TAF periods and elements aggregated, for the actual forecasts and for persistence, plus wind verification as in the current AIFS prototype. Verification measures should include POD, FAR, bias, CSI, HK, HSS and RF1 plus d’ and β for the 2x2 tables and Az for the 4x2 table. Thresholds to be set as defaults for extraction of the JDs should be determined for each station.

As a single summary measure of skill Az for the 4x2 table is recommended.

For individual forecasters aggregated 2x2 and 4x2 JDs and verification measures for the individual forecasters most recent 30 (say) shifts should be available at logon. The optimal number of shifts may need adjustment with experience.

This basic set of JDs and measures should be available as a default without requiring entry of parameters for time periods, etc. In a busy office the need to do this is a disincentive to use the system. An option to go to the full TAF verification system should be available if required.

Output Format Type B:

Simpler output for weather services users, the media, Bureau management and Government.

The aviation industry

The requirement is for a simple presentation that nevertheless carries enough detail to be meaningful. As suggested by the preceding pages, AIFS creates the potential for a deluge of information. The most appropriate mix will be a matter for development.

Following discussions with industry representatives, it is clear that a major concern is the reliability of forecasts on occasions on which weather below alternate minima is either forecast or observed, and there is a need to examine these situations in some detail.

Output proposed for the industry, at least as an initial product for discussion, is, at the “top level”, sample estimates for four quantities. These are

- the probability of sub-threshold conditions for any hour for which above-threshold conditions are forecast,
- the probability of sub-threshold conditions during any hour for which INTER is forecast,

- the probability of sub-threshold conditions during any hour for which TEMPO is forecast, and
- the probability of sub-threshold conditions for any hour for which sub-threshold conditions are forecast.

These are available from the CR factorisation of the 4x2 JD, as described above.

It may be necessary to provide this set of probabilities for a number of different thresholds at some airports, and for different lead times. It is understood that the longer TAF lead times are important for planning purposes for international flights.

At the next level of detail, the following information should be available.

- A listing hour by hour of occasions on which below-threshold conditions are either forecast or observed, in the form of the verification data sample described earlier in this Report. As noted there, the threshold should be variable by the user.
- The joint distribution corresponding to this verification data sample, in both the 2x2 and 4x2 forms described above.
- Verification measures for each of the JDs produced.
- JDs for each 3-hour interval (non-overlapping) of the validity period of each standard TAF issue for user-defined periods, together with verification measures in each case, so that changes in skill with lead time can be assessed.

Management.

AS TAFs are provided as a service to the aviation industry it seems appropriate that management receive as a default the same data listed for the industry above.

As a single summary measure of skill Az calculated for the overall aggregated 4x2 JD is recommended.

As familiarity with the capabilities of the AIFS system and the breadth and flexibility of the DO framework develops other kinds of output can be provided. Availability of a comprehensive basic verification data set as described in 5.5 above should facilitate the development of new forms of output.

As with Category A output, the basic set of information should be available as a default without the need to enter parameters. Suitable time periods and thresholds will need to be determined on an individual basis.

Government

As a single summary measure of skill Az calculated for the aggregated 4x2 JD is recommended.

An adequate explanation of the meaning of Az at this level would be that it is a measure of skill in forecasting operationally significant weather, and ranges from 50% representing no skill to 100% representing perfect skill.

Verification of 7-day forecasts

Recognition that there is some skill in numerical models at seven days suggests the possibility of providing useful weather forecasts out to this lead-time. The media are interested in this service and some outlets already broadcast 7-day temperature forecasts in which at least the sixth and seventh day's forecasts are produced by non-Bureau forecasters. Reliable forecasts of temperature and rainfall at longer ranges would have economic value to the community.

Temperature forecasts with lead times out to seven days have recently begun to be prepared by the Bureau of Meteorology at a number of capital cities (Melbourne, Sydney, Canberra and Brisbane).

At this stage there appear to be no plans to produce quantitative forecasts for variables other than temperature out to seven days. In general rainfall is of more significance than temperature, so in anticipation of longer period QPF some comments are made below on verification of this type of forecast.

The main theoretical interest in longer period forecasts at present is in the limit to skill; is it possible to consistently show useful skill at lead-times beyond 4 days? This section therefore focusses on suitable measures of skill for this purpose.

Present practice

AIFS

Verification of 7-day temperature forecasts in AIFS is undertaken for maximum and minimum forecasts prepared at Melbourne, Sydney, Canberra and Brisbane in the AIFS temperature FVS.

The methods and measures used are those used in the AIFS FVS temperature verification module (section 2), and comments in that section apply equally here.

Similarly, it is expected that longer period QPF will be prepared and verified in the same form as RAINV at present (section 3).

Limits to skill

Recent studies

There have been several recent studies of longer period forecasts in Australia that are very briefly reviewed below.

Noone and Stern

Noone and Stern (1995) verified 100 days of QPF from the GASP model for Melbourne, Perth and Brisbane, out to 7 days. Verification measures used were PC, HK, RMSE and the correlation coefficient. Skill was assessed in relation to climatology and persistence. Forecasts were verified against observations from a single rain gauge.

They found that, overall, the forecasts had some skill out to four days.

Stern

Stern (1998) reported an experiment to establish the limits of the Bureau's predictive capability. Forecasters in the Victorian Regional Office prepared quantitative forecasts out to 14 days, based on the official 4-day outlooks for days 1-4 and on subjective interpretation of NCEP predictions of MSLP, 500 hPa height and 1000-500 hPa thickness for days 5-14. He used a number of verification measures, mostly based on RMSE as the measure of correspondence between forecasts and observations, and calculated skill scores based on monthly mean climatology.

Minimum temperature forecasts and yes/no rain forecasts were found to have some skill compared with climatology out to day 4. Maximum temperature forecasts had some skill over climatology out to 6 days. QPFs prepared in the RAINV format fell to a level of skill no better than climatology at day 4.

Stern colcluded that the level of skill (for Melbourne) was such that "**routinely** providing or utilising day-to-day forecasts beyond day 4 would be inappropriate at present", although useful forecasts might be possible for some elements, in some seasons, and in some situations, out to about six days. It might also be possible to make useful statements about the expected average weather conditions over the 10-day period from day 5 to day 14.

Davis

Davis (1999) verified two years of 5-day maximum and minimum temperature forecasts prepared in Canberra as a commercial service. Verification measures were the Priestly skill score (an MSE-based skill score; Davis used long-run daily mean temperatures as the climatological forecasts), and percentage of forecasts with smaller error than the climatological forecast.

He found positive skill over climatology for both maximum and minimum forecasts out to day five, the amount of skill being dependent on season. Linear extrapolation of the trend in skill indicated that overall the limit of skill over climatology was reached at around day 6 for minimum temperatures and day 7 for maximums.

Mason

Mason (1999) verified ten years of probabilistic forecasts for rain occurrence (yes/no) in Canberra at some short lead times out to 24 hours, verified at a single gauge. The verification measure was the SDT measure Az.

The decline in Az with lead-time was close to linear, and extrapolation indicated that zero skill would be reached at about 3½ days

Williams

Williams (http://www.ozemail.com.au/~wbc/forecast_verify.html) verified maximum and minimum temperature forecasts with lead-times out to 4 days for a large number of locations around Australia, and some 7 day forecasts by Weather News International (WNI) for Channel 7 in Sydney. The main verification measure was the Priestly skill score using daily climatology.

Overall, he found positive skill at 4 days. The rate of decline of skill with lead-time suggests limits beyond 7 days for both maximums and minimums, although this would be highly dependent on location.

The WNI maximum temperature forecasts for Sydney had positive skill at 7 days, albeit small. This was not the case for other cities, but Williams considered that WNI's problems updating their web site contributed to an apparently poor performance..

Overseas

There appears to be little recent published work on limits to skill in forecasting specific weather parameters. Eckel and Walters (1998) reported on a year of probabilistic QPFs produced from ensemble runs of a numerical model, at lead times out to 16 days. The grid scale for the forecasts was 2.5º and forecast validity period 24 hours. Verification measures were the ranked probability skill score and the Brier skill score (essentially the same as the Priestly skill score).

They found that skill declined to equivalent to climatology for yes/no precipitation forecasts at a lead time of six days. It was less for thresholds at higher rainfall amounts. They considered that because of the coarse grid scale and validity period, it is likely that these limits are optimistic. Forecasts for smaller scales would be less skillful.

Eckel and Walters make the interesting comment that, in using probabilistic QPF (PQPF) products, “forecasters should be aware of the limits of predictability. For example, PQPF for CAT4[1] is available for out to 9 days, but this research showed that … such a forecast is of only of value out to about 2 days”. This kind of information can only come from verification.

Comments

There is a degree of consistency between the results of the above studies, regardless of the format of the forecasts or the particular measure of skill. The average limit to skill for temperature forecasts is around 6-7 days and for forecasts of simple occurrence or non-occurrence of rain, between 3 and 6 days, probably nearer to the low than to the high end of this range. These are overall figures, and considerable variation could be expected between locations.

Assessment of skill

Measures of skill, accuracy and value

Baselines for skill

Since interest in longer lead-time forecasts focusses on the time at which zero skill is reached, the definition of zero skill is of some importance.

Climatology

The most common baseline for zero skill is climatology, which in this context means a constant forecast of the mean (or occasionally the median) value of the predictand. Practice varies in whether to use the sample mean over the verification data set or the long-term mean and if the latter whether to use annual, monthly or daily means.

In day-to-day temperature forecasting for the public the appropriate value is the long-run mean daily temperature. The long-term mean should be used since this would be known exactly to the forecaster in advance whereas the sample mean in general would not. Use of a long-run mean also gives the forecasting system credit for recognising deviations of the sample climate from the long-run climate. Use of daily means rather than means of longer periods avoids discontinuities, for example between the last day of one month and the first day of the next. The annual mean is just too easy to beat and gives an unrealistic impression of skill.

The daily means should be a smoothed time series, rather than simple averages for each day.

Similar considerations apply to precipitation.

Persistence

Another zero-skill forecast sometimes used is persistence, essentially a forecast that the most recent relevant observation available at forecast issue time will persist through the forecast period. In the case of temperature forecasting this would usually mean taking the maximum or minimum temperature on day N as the forecast for day N+1.

Persistence is often surprisingly accurate over short lead times, as a result of the existence of useful serial correlation in some meteorological time series. It is very difficult to beat consistently over lead times less than three hours. Beyond 24 hours however it is usually a weak standard for skill.

Climatology and persistence

Murphy (1992, 1996) pointed out that a linear combination of climatology and persistence may out-perform either singly as a standard for skill. The optimal weightings are determined by the

[1] Greater than 1 inch (25.4mm)

correlation of the persistence forecast, although the optimality of a linear combination like this is not particularly sensitive to the precise values of the weights (eg Dawes, 1979).

On the principle that the best “naïve” standard for skill should be applied (Williams 1997), the use of a combination of climatology and persistence as a basline in skill scores should be investigated.

Measures-oriented approach

In a traditional measures-oriented approach to verification it is usual to calculate a measure of correspondence between forecasts and observations for both the actual forecasts and for the zero-skill forecasts, and form the proportional difference between the two. The resulting proportion is referred to as skill over the baseline forecasts. The Priestly skill score used in the AIFS FVS temperature verification is a score of this type.

There is a place for skill scores calculated in this way, as broad summary measures of the “value added” by the forecasting system to simple forecasts that could be provided at no cost above that of the observations. Several points need to be borne in mind when interpreting skill scores. Very briefly, these are

- Values of single measures of forecast quality do not respect the dimensionality of most verification problems (Murphy 1996).
- Skill scores can be partitioned in various ways, revealing multiple aspects of the forecasts, the observations and their relationships and raising complex issues in interpretation. Superiority in terms of a measure of “overall skill” is no guarantee of superiority in terms of all aspects of forecast quality (Murphy 1996)
- Apparent skill is dependent on the baseline for zero skill.

With these caveats, the Priestly skill score using long-run daily climatology is a satisfactory summary indicator of forecasting skill for 7-day temperature outlooks.

Recommendation: That the Priestly skill score be used as a summary measure of skill for 7-day outlooks. Smoothed daily means should be used as the climatology.

Assessment the skill of rain forecasts in a measures-oriented framework similarly involves calculation of scores for the actual forecasts and for climatology or persistence as reference forecasts. None of the currently used scores is entirely satisfactory, due to dependence on either sample climate or decision threshold or both (section 9.3). As discussed in relation to the RAINV forecasts (section 3), the best summary measures of skill available at present are those based on SDT/ROC methods, either d’ or Az.

Recommendation: That (d’, β) be used as summary indicators of skill for 7-day rain forecasts prepared as yes/no forecasts. For probabilistic forecasts Az should be used.

A distributions-oriented approach

Discussions of skill in a DO framework have usually involved, as in the measures-oriented approach, comparison of the accuracy of the forecasts with the accuracy of reference forecasts, either climatology or persistence (eg Murphy 1993). A different view is possible within the DO framework, which does not require explicit definition of a no-skill baseline forecast.

It seems reasonable to say that a forecasting system has no skill if learning the forecast can not change a user’s pre-existing opinion about the weather, whatever the forecast. It is assumed that the user has enough information about the performance of the system to make this judgement on a rational basis, and that the performance of the system is stationary. In the simplest case, suppose the forecasts are for a two-state event E Y {0,1}, and the allowable forecasts are F Y $\{f_i\}$, i=1,..,N, for example probabilistic forecasts of precipitation occurrence, where the f_i would be probabilities. Then the forecasts have no skill if

$$P(E=1|F=f_i) = P(E=1) \qquad (27)$$

for all forecasts f_i.

Eqn 27 could be used as it stands to assess skill, by comparing the conditional probability on the LHS with the unconditional (climatological) probability on the RHS for each forecast. Alternatively, Bayes' formula in the odds form gives

$$P(E=1|F=f_i)/P(E=0|F=f_i) = [P(E=1)/P(E=0)].[P(F=f_i|E=1)/P(F=f_i|E=0)] \quad (28)$$

The second term in square brackets on the RHS of 28 is the ratio of the ordinates of the two basic distributions of SDT, the distribution of the forecasts before occurrence of the event and the same before non-occurrence. These distributions may be derived from the LBR factorisation of the JD. If 27 holds then this ratio is unity for all i and the ROC of the forecasting system plots on the diagonal, corresponding to d'=0.0 or Az =0.5. This (much abbreviated) argument shows that the SDT indices are measuring skill over a non-skill baseline represented by eqn 27.

A full DO analysis of temperature and rain forecasts should follow the frameworks outlined in sections 2 and 3 of this Report.

Communication with users

Output format type A:

Diagnostic output for individual forecasters, forecasting teams, numerical modelers and the developers of Model Output Forecasts (MOF)

The main quantity of interest in 7-day forecasts is the lead-time at which skill falls to zero. As noted in section 3.1.7 above, the consistency of empirical studies suggests that the particular measure of skill used does not have a major effect on this parameter. Limits to skill for temeprature forecasts come out at 6-7 days and for rainfall 3-6 days, for a variety of measures.

As noted above, for temperature forecasts the Priestly skill score is a satisfactory overall indicator of skill over a baseline forecast, subject to careful interpretation as outlined in 6.3.2 above.

For rain forecasts the SDT/ROC indices (d', β) or Az are appropriate.

At a higher level of detail, the full VDS and JD, together with CR and LBR factorisations and verification measures as detailed in sections 2 and 3 above, should be available.

Output Format Type B:

Simpler output for weather services users, the media, Bureau management and Government.

In 2.9.2.1 above MAE was recommended for communicating the accuracy of temperature forecasts to non-specialists, referred to as "average accuracy". It is possible to form a skill score from MAE in the same way as MSE, but this loses the simplicity of interpretation.

Following Davis (1999) a suitable measure for temperature forecasts of skill over climatology at this level might be the percentage of forecasts with a smaller absolute error than climatology (long-term daily). Zero skill would be indicated by 50%.

Graphs of these two quantities against forecast lead-time would probably provide as much detail about 7-day forecasts of temperature as most users at this level would want.

In the case of rainfall, verification measures based on SDT/ROC methods are the best available at present. Az for the rain/no rain threshold, expressed as a percentage and referred to as a "skill index" may be suitable, presented with the information that 50% represents no skill and 100% perfect skill.

Verification of model output forecasts (MOF)

Model Output Forecasts (MOF) are a development of the Model Output Statistics (MOS) approach to obtaining forecasts of weather elements from numerical model output (Glahn and Lowry, 1972). Forecasts are produced by a linear multiple regression process in which observed values of predictands are related to predictors selected by screening regression from numerical model output.

Issues of interest in verification of MOF forecasts are the need for real time verification, the large amount of data potentially available, and the rate of decline in skill with lead time.

The forecasts

The current set of MOF forecasts provides, for all Australian stations with sufficient data, predictions out to 60 hours from the initialising observations, at 3-hour intervals where appropriate and possible. Predictions are made from both the 00UTC and 12UTC model runs.

Quantities forecast are

Minimum temperature*

Maximum temperature*

Temperature at 3-hour intervals*

Precipitation: 24-hour totals to 0900K*

Precipitation: 3-hour totals*

Terrestrial minimum temperature*

Dew point at 3-hour intervals*

Wet bulb temperature at 3-hour intervals*

Relative humidity at 3-hour intervals

Wind direction at 3-hour intervals*

Wind speed at 3-hour intervals*

Total cloud at 3-hour intervals

Low cloud amount at 3-hour intervals

Hours of sunshine

Visibility at 3-hour intervals

Evaporation: 24-hour total to 0900K

MSLP at 3-hour intervals

Not all stations have forecasts for all these variables at all times.

Current verification practice

MOF output is or will be verified within the appropriate AIFS FVS module. Temperature forecasts will be verified by the AIFS temperature FVS, precipitation in the rain FVS.

All the verification measures and graphical displays used in these modules are available for analysis of the relevant MOF forecasts.

Forecasts marked by an asterix in the above list are those for which the discussions and recommendations in this Report under temperature, rain or, in the case of wind, the Fire Weather FVS are relevant, and these discussions should be referred to.

Comments

As noted in earlier sections of this Report, the current AIFS modules provide a sound basic system which can be extended to provide a full DO verification, and this is equally true when the modules are applied to MOF forecasts.

In view of the operational nature of MOF guidance it should be verified routinely in real time. This suggests that MOF should be verified in a separate module from other similar forecasts, and that summary verification measures for the most recent 30 or so forecasts for which data is available should accompany each issue of MOF.

Recommendation: That MOF be verified in a separated AIFS FVS module and the results for the previous 30 forecasts be transmitted with each operational issue.

There are no fundamentally novel problems in verification of MOF output. The basic verification process and DO approach should be followed.

The verification process

Verification Data Set and Joint Distribution

In all cases the VDS should include lead time from the observational data on which the forecasts are based, and be flexible enough to include other covariates if required.

The basic JD and CR and LBR factorisations as detailed elsewhere in this Report should be available for each location, each forecast variable and each lead time, with an option to consolidate locations and lead times, in order to reduce the quantity of data.

The facility to display verification data in map form, as provided in the AIFS temperature FVS, is useful and should be available in the MOF module.

Graphical displays of the distributions of forecasts against observations and vice versa should be available, and the facility to plot verification measures as time series and against lead time.

Verification Measures

In view of the large amount of verification data potentially available from the MOF module it is desirable to select a small number of summary verification measures that can be used as Category B output, and internally for a quick “first look” at performance. Some recommended measures for this purpose follow.

Temperature-related

Temperature forecasts from MOF are

Minimum temperature

Maximum temperature

Temperature at 3-hour intervals

Terrestrial minimum temperature

Dew point at 3-hour intervals

Wet bulb temperature at 3-hour intervals

For consistency with recommended practice in the temperature section of this Report, MAE should be used as a “top level” summary statistic for temperature forecasts.

For forecasters bias (μ_f-μ_x) is an important statistic. Beyond this, the correlation coefficient is a good single-number measure of association for variables that are linearly related and more or less normally distributed. Scatter diagrams and details of the fitted linear model should also be provided in this case. Williams (1997) is a good example of the use of linear regression in operational temperature verification.

MAE, bias and the correlation coefficient should be available for each of the 3-hour steps where these are available, and as an aggregate over the full forecast period.

Recommendation: That MAE and bias be used as summary measures of skill for MOF temperature forecasts.

Rain

Rain forecasts provided by MOF are:

Precipitation: 24-hour totals to 0900K

Precipitation: 3-hour totals.

As a summary indictor of skill d' and β for the rain/no rain threshold, as discussed in section 3.5.2, are recommended. β can be converted to a threshold probability if required. The option should be available to set different threshold values to derive the JD. When and if MOF precipitation forecasts are produced as probabilities, Az is a preferable measure.

Since both forecasts and observations are available as quasi-continuous variables (mm, rather than categorised as in RAINV), the correlation coefficient appears to be a suitable measure for internal users, possibly after transformation of forecasts and observations to normality (eg Graedel and Kleiner 1985). Where correlation coefficients are used in verification it is desirable that details of the fitted linear model also be available, and scatter diagrams. The correlation coefficient is unreliable as measure of association if the relationship is non-linear, so the means of assessing linearity should be available. Bias should also be available.

(d'.β), bias and the correlation coefficient should be available for each of the 3-hour steps where these are available, and as an aggregate over the full forecast period.

Recommendation: That (d'.β) for the rain/no rain threshold be used as a summary indictor of skill for MOF rain forecasts.

Wind

Wind forecasts provided by MOF are:

Wind direction at 3-hour intervals

Wind speed at 3-hour intervals.

Wind direction is in general of less interest than wind speed, so for simplicity it is recommended that summary statistics for wind speed only be provided at this level.

For non-meteorologists MAE is probably the most comprehensible measure.

A suitable summary statistic for wind speed for internal users is the correlation coefficient, possibly after transformation of forecasts and observations to normality. As noted above, details of the fitted linear model and scatter diagrams should be available. Bias should also be available.

If a single-number measure of accuracy for wind direction is required, MAE for direction in degrees is suitable.

MAE, bias and the correlation coefficient should be available for each of the 3-hour steps where these are available, and as an aggregate over the full forecast period.

Recommendation: That MAE and bias for wind speed be used as summary indictors of skill for MOF wind forecasts.

Other

Other variables forecast by MOF are:

Relative humidity at 3-hour intervals

Total cloud at 3-hour intervals

Low cloud amount at 3-hour intervals

Hours of sunshine

Visibility at 3-hour intervals

Evaporation: 24-hour total to 0900K

MSLP at 3-hour intervals

All these variables except MSLP have limited ranges and some have odd climatological distributions, for example cloud amount which tends to be U-shaped. These factors raise problems in the selection of a single-number index of skill. Nevertheless, it would be convenient to use the same measure of association as far as possible for all MOF forecasts. Therefore MAE should be used as a “quick look” measure of accuracy for non-meteorologists, possibly with some explanatory comment about the scale and range of each variable. For internal users bias (μ_f-μ_x) together with the correlation coefficient is suitable, with optional access to the full details of the fitted linear model and scatter diagrams.

MAE, bias and the correlation coefficient should be available for each of the 3-hour steps where these are available, and as an aggregate over the full forecast period.

Recommendation: That MAE and bias be used a summary measures of performance for the MOF forecasts listed above.

Communication with users

Output Format Type A:

Diagnostic output for individual forecasters, forecasting teams, numerical modelers and the developers of Model Output Forecasts (MOF)

The information required on the performance of MOF by individual forecasters and forecasting teams is essentially bias, so that adjustments can be made in real time.

At the next level of detail for output format Type A, the summary statistics detailed in the previous section plus the full DO analysis should be available for each of the forecast variables for each station, at each 3-hour interval where appropriate.

Output Format Type B:

Basic information for weather services users, the media, Bureau management and Government

The measures recommended in sectionx.3.2 are suitable for this group, essentially MAE for non-specialists, plus bias and the correlation coefficient, or Az where appropriate, for users with some background in meteorology.

Appendices

Outline of a structured approach to forecast verification

This section is intended to provide an introduction to terms and concepts used in the Report. It is not intended to be an exhaustive treatment of forecast verification. For further detail the references should be consulted.

In order to provide some structure for individual modules in this Report, forecast verification is modelled as a three-stage process.

These stages are

- data collection,
- analysis and
- communication with stakeholders.

The outcome of the *data collection* stage is the verification data sample (or samples) and joint distribution of forecasts and observations. These are the basic data structures for any verification problem.

In a distributions-oriented framework the *analysis* of verification data starts with factorisation of the joint distribution into marginal and conditional probabilities, broadly referred to as performance characteristics,

Further analysis involves calculation of verification measures, which summarise significant aspects of forecast quality, and proceeds to making inferences about the absolute and/or relative performance of forecasting systems on the basis of the performance characteristics and verification measures.

Communication with stakeholders involves selection of appropriate measures and forms of communication. This aspect is not often considered explicitly in verification studies but is clearly important, as the whole effort is lost if communication between verifier and user fails. Methods and measures appropriate in a research environment are not in general suitable for decision-makers in management or government, and the selection of simple but not over-simplified measures of forecasting performance is not a trivial exercise.

Splitting the verification process conceptually into three stages provides a useful structure and makes it easier to localise difficulties. For example, in yes/no forecasts of rare events the frequency of correct forecasts of non-occurrence in the 2x2 contingency table is often several orders of magnitude larger than the other frequencies, and this can render interpretation of common measures of skill unclear. This particular problem has usually been regarded as a problem with the measures, the second stage of the process above (see, for example, Doswell et al 1990, and Schaefer, 1990). It is more appropriately seen as a problem with stage 1, definition of the basic verification data set, specifically with what is, and is not, to be counted as a forecast and observation of non-occurrence of the event of interest.

These components of the verification process are not carried out in isolation from each other. Development of the verification data set clearly needs to be done with knowledge of the quantities to be calculated from it (stage 2), and of the likely recipients of the information (stage 3). Nevertheless, it is not always possible to foresee all the uses to which verification data may be put so as much flexibility and generality as possible should be built in.

The next sections consider each of these stages in more detail.

Data collection

Verification data set (or sample)

In its simplest form, the VDS is a set of matched pairs of forecasts and observations. Following Murphy (1997), the VDS is represented by $\{(f_i,x_i), i=1,..,n\}$ where n is the sample size. A typical example is a listing of daily forecast and observed maximum temperatures.

Original and derived VDS

When an underlying continuous (or quasi-continuous) variable is to be categorised, for example in assigning rainfall in mm to one of the RAINV categories, or converting cloud base in feet to a dichotomous variable, above or below the aerodrome minimum, there are effectively two VDSs which are referred to here as *original* and *derived*.

The original VDS contains the raw forecasts and observations, and the derived VDS contains the categorised forecasts and observations. For example, the procedure recommended in relation to TAF verification in the body of this Report envisages two VDSs, one containing the forecasts and observations in as close as possible to their raw form, and the other containing forecasts and observations categorised according to user-supplied thresholds. It is desirable that the original VDS be retained so that different thresholds can be applied to extract derived VDS as required.

Some other issues in relation to the VDS

It is necessary to have unequivocal definitions of the quantity being forecast, and to ensure that the verifying observations correspond as closely as possible to the definition of the forecast quantity. The event being verified should be the same event that the forecaster was trying to forecast.

For example, when a forecaster issues a maximum temperature forecast, is the appropriate verifying observation the reading on the max temp thermometer at 9.00 am, or is it the mid-afternoon maximum temperature (the 9 am max problem)? If the forecaster is assessing the chance of thunderstorms in a defined time and area, are marginal reports (one GPATS report in a sparsely inhabited corner of the area) treated as an occurrence of a storm in the same way as unambiguous reports (widespread severe storms)? The point for present purposes is that these decisions are part of the first stage of verification, development of the VDS. The criteria for classification of verifying data must be clear and explicit, and fair to the forecasting system

Sometimes the observational data is not adequate to fully determine whether a forecast was correct or not, for example in the case of INTER and TEMPO in TAFs. In these cases some explicit rule must be devised to determine the class to which the observation is to be assigned (for example Keith's rules implemented in the prototype FVS for TAFs), or the cases discarded.

Stationarity and serial dependence

When the VDS is a time series it is usually assumed, often without stating it, that it is stationary in the sense that there is no change in the relevant climatological means and variances, and also that successive terms are statistically independent. These assumptions are made to justify the statement that the joint distribution of forecasts and observations contains all the information from the VDS relevant to assessment of forecast quality (Murphy 1997).

Both these assumptions should ideally be tested. An overall measure of skill calculated for a VDS within which there are significant trends or discontinuities in skill is clearly of limited significance. Trends in verification data are of interest in themselves, suggesting improvement or deterioration in the forecasting system. Discontinuities are similarly of interest, and may suggest that the data set be split before making an assessment of forecast quality.

Serial dependence reduces the effective sample size for statistical tests. There has been little or no examination of its effect in forecast verification data sets, largely because efforts to use standard statistics to assess the significance of differences in verification measures have been rare (see Seaman et al 1996). A recent discussion of time series in meteorology is in Wilks (1995).

Joint Distribution

The joint distribution of forecasts and observations is a contingency table summarising the performance of the forecasts.

If there are n^f distinct forecasts and n^x distinct observations, and n_{ij} denotes the joint frequency of forecast f_i and observation x_j, then the n^f-by-n^x matrix $\{n_{ij}\}$ contains the joint frequencies of all the (f,x) pairs in the VDS. The joint distribution is the matrix of the corresponding joint relative frequencies (p_{ij}), where $p_{ij} = n_{ij}/n$. These joint relative frequencies are usually treated as joint probabilities, ie $p_{ij} = \Pr[f=f_i,x=x_j]$, often written p(f,x).

In fact the p_{ij} are *sample estimates* of the joint probabilities, but it is customary in forecast verification to ignore this (Murphy 1997).

The joint distribution of forecasts and observations is the basic data structure for the DO approach to forecast verification. The JD contains all the non-time-dependent information that is relevant to assessment of forecast quality, where forecast quality is defined as the totality of the statistical characteristics of the forecasts, the observations and their relationships embodied in this distribution (Murphy 1997). Forecast verification consists of describing and summarising the statistical characteristics of the JD.

Data analysis

The second stage of the verification process proceeds in three parts,

- factorisation of the joint distribution into marginal and conditional probabilities, broadly referred to as performance characteristics,
- calculation of verification measures, which summarise significant aspects of forecast quality, and
- making inferences about the absolute and/or relative performance of forecasting systems on the basis of the performance characteristics and verification measures.

Factorisations of the JD

Insight into forecasting performance can be gained by factoring the joint distribution into marginal and conditional distributions. For a two-dimensional distribution there are two such factorisations. These are the calibration/refinement (CR) factorisation

$$p(f,x) = p(x|f).p(f), \text{ and} \tag{A1}$$

the likelihood/base rate (LBR) factorisation.

$$p(f|x) = p(f|x).p(x) \tag{A2}$$

The CR factorisation

In the CR factorisation, p(x|f) represents the distributions of observations given the forecasts and p(f) is the marginal or unconditional distribution of the forecasts. The terminology reflects the history of the concept in analysis of probabilistic forecasts. In this situation p(x|f) indicates the *calibration* or reliability of the forecast probabilities; the correspondence between forecasts and sub-sample relative frequencies.

In general the p(x|f) describe characteristics of the forecasts related to communication between the forecasting system and users. The p(x|f) are the distributions of weather x that users should expect given the forecasts f. “Reliable” or in the probabilistic context well-calibrated forecasts have the property that, at least on the average, the forecasts really mean what they seem to mean; the average observation for each forecast is the same as the forecast.

The second term in the CR factorisation, p(f), is the distribution of the forecasts, regardless of the verifying observations. For probabilistic forecasts of binary events p(f) measures the ‘refinement’ or ‘sharpness’ of the forecasts, their tendency to be concentrated at the extremes of the probability range, or not.

The LBR factorisation

In the LBR factorisation the p(f|x) represent the distributions of forecasts given the observations, and p(x) is the marginal or unconditional distribution of the observations. The p(f|x) distributions are analogous to likelihoods in statistics, which are probabilities for data conditional on hypotheses. The p(f|x) are closely related to the intrinsic capacity of the forecasting system to discriminate between the events to be forecast, regardless of whether the forecasts are presented in a “reliable” form. Discrimination refers to the propensity of the forecasting system to issue different forecasts before different weather events. A system shows perfect discrimination if the p(f|x) do not overlap at all for different observations x.

The SDT indices d’ and Az measure discrimination in this sense.

The second term of the LBR factorisation, the marginal distribution p(x), is often referred to as the sample climate. In the binary case p(x=1) may be referred to as the base rate of the event. p(x) is a characteristic of the environment of the forecasting system rather than of the system itself, but does affect many verification measures.

Murphy (1997) has identified ten aspects of forecast quality related to the joint and marginal distributions described above.

The dimension of verification problems

The dimensionality of a verification problem is defined as the number of probabilities or parameters that must be estimated to reconstruct the basic joint distribution p(f,x) (Murphy, 1991).

In the 2x2 case, a minimum of three values is required to reconstruct the joint distribution. These could be the joint probabilities for three of the four cells, the fourth being fixed by the constraint that the sum equals unity. Alternatively, POD, POFD and the sample relative frequency of the event are sufficient. However, POD, FAR, bias and CSI together are not.

For nxm joint distributions, the dimensionality is in general equal to mn-1, and grows quite rapidly as m and n increase. For a 6x6 JD the dimensionality is 35; 35 parameters are required to completely reconstruct the distribution.

Murphy comments that verification problems generally suffer from the “curse of dimensionality”. That is, most problems require specification of a relatively large number of quantities to describe forecast quality completely. Nevertheless, if the dimensionality of verification problems is not respected, important aspects of forecast quality may be overlooked and misleading results obtained.

Much effort has gone into the search for a single-number summary measure of forecasting skill, and continues (eg Potts et al, 1996). Consideration of the dimensionality of verification problems suggests that there is no completely satisfactory single measure of this kind, even in the simplest case of non-probabilistic forecasts for a binary event. Forecast quality is inherently multidimensional, and its multifaceted nature must be accommodated if misleading results are to be avoided. Murphy (1996) stated, in relation to verification measures for multidimensional contingency tables, “It is now generally understood that no universally acceptable measure of performance for the kxk problem can be found”, and this appears to be true for all k≥2.

There have been some investigations of the possible use of statistical models to reduce the dimensionality of verification problems. Parameters of the models might then provide useful verification measures. Murphy (1997) describes some such studies. Measures based on SDT such as d' fall into this class.

The salient message of Murphy's investigations of dimensionality is that no single-number index of forecast quality can be adequate, even in the simplest case. Forecast quality is intrinsically multidimensional, and verification must take this into account.

Treatment of covariates

There is often interest in other quantities related to the forecasts and observations of primary concern. Among these are the date, season, time of issue of the forecast and lead time to the verifying observation, and other variables that may be used to stratify the data, for example synoptic map types, individual forecaster's identification, SOI, Total totals index, and so on. The VDS may include these, and the database system should be flexible enough to be expanded to include new variables if the need for them becomes apparent later in the study.

Murphy (1995) proposed an extension of the distributions-oriented framework to provide a coherent method of addressing stratification of data in this context. In brief, this approach involves the introduction of a new variable, Z, to identify some quantity that it is thought might have significant implications for forecasting performance. This might be a classification of weather regimes or of degrees of forecast difficulty however arrived at. If Z takes values from the set $\{z_j\}$, $j=1,..,n_j$, the joint distribution can then be expressed

$$p(f,x) = \Sigma_j\ p(f,x|z_j).\Pr[z_j] \qquad (A3)$$

and the CR and LBR factorisations become

$$p(f,x|z_j) = p(x|f,z_j)p(f|z_j) \qquad (A4)$$

and

$$p(f,x|z_j) = p(f|x,z_j)p(x|z_j) \qquad (A5)$$

For details of the suggested use of these factorisations in verification Murphy's paper should be consulted. He states that "at a minimum, ..., the augmented framework appears to provide a reasonably general and potentially useful structural setting within which alternative verification or forecast-quality studies can be designed and/or their results evaluated".

Verification measures, performance measures, scoring rules

The second stage of *analysis* of verification data involves calculation of summary measures of aspects of the joint, marginal and conditional distributions and their relationships.

A *verification measure* is defined by Murphy (1997) as any function of the forecasts, to observations, or their relationship. Examples are means, medians, variances, covariances, etc.

A *performance measure* is a verification measure that focusses on the correspondence between forecasts and observations. Examples are MSE and the Heidke score.

A *scoring rule* is a performance measure that is defined for individual pairs of forecasts and observations. Examples are MSE or the log skill score. Scoring rules are typically used as day-to day feedback on performance, and should be "proper" (Murphy 1997).

There is an extensive literature on verification measures and scoring rules (see references in Wilks, 1995). Some of those familiar to Australian meteorologists are discussed in Appendix 8.2, and measures from signal detection theory in Appendix 8.1.4.

Communication

The need for appropriate communication of the results of verification to users has had little explicit discussion in the verification literature. Nevertheless, if this communication is not effective the whole effort will be lost.

A full DO analysis of a forecasting system’s performance can generate a large amount of information, much of which requires significant background in meteorology and statistics for its comprehension. The challenge in communicating verification results is to select measures and forms of display which are soundly based in theory and convey the essential meaning of the data without being confusing or misleading.

Where forecast users form an identifiable group, for example the aviation industry, or emergency services in the case of bushfires, or within the Bureau operational forecasters, they may have views on useful forms of verification for their own purposes. These views should be ascertained where possible.

A comprehensive basic verification data set facilitates the production of different forms of verification output for different users.

An outline of methods from signal detection theory in weather forecast verification

This section is an outline of an approach to forecast verification using methods from the mathematical theory of detection of signals in noise (SDT, for signal detection theory), and the use of the relative operating characteristic (ROC). It is based on a paper given at the Forecast and Warning Verification Workshop in 1997.

Methods based on SDT and the ROC are a useful addition to techniques currently used in forecast verification. They facilitate the assessment of pure forecasting skill, as distinct from an ability to communicate with users of forecasts (reliability or calibration). They also provide measures of skill that can be calculated for forecasts issued in any form, enabling valid comparisons between the skill of simple yes/no forecasts and those issued as risk ratings or probabilities, or in any other way.

ROC/SDT methods are compatible with and complement the distributions-oriented approach developed by Murphy and Winkler (1987).

These methods also provide criteria for good measures of forecasting skill and reveal deficiencies in current measures (Mason 1982b, 1989; Swets, 1986).

The methods described were originally developed by psychologists studying human sensory discrimination, including the ability of human observers to detect specific signals on radar screens in a military context, and also drew on statistical decision theory and detection of electromagnetic signals in noise. Swets (1973) gives a detailed account of the historical development of the field to that time, updated in 1988 (Swets, 1988). In a meteorological context, Mason (1980, 1982a,b, 1989) has discussed the applicability of the methods to forecast verification. Other presentations for meteorologists include Levi (1985), Harvey et al (1992) and Buizza et al (1999). Some good basic texts are Egan (1975), Swets and Pickett (1982), Macmillan and Creelman (1991) and Swets (1996).

This presentation is rather more discursive and heuristic than it would be in a formal setting, as the ideas are still unfamiliar to most meteorologists.

The signal detection model

There are many situations of practical importance in which it is necessary to decide among alternative courses of action on the basis of information which does not provide absolute certainty. A meteorologist deciding whether to forecast rain on the basis of the usual synoptic data is usually an example of such a situation. Another is the manager of a weather-sensitive business deciding on the basis of a weather forecast whether to take precautions against adverse weather. Others, among very many, include medical professionals making a diagnosis, engineers seeking metal fatigue in aircraft, police using polygraph lie detectors, and research scientists seeking to ascertain whether a particular experimental manipulation did or did not have a significant effect.

A common feature of these and many other formally similar situations is that the available information provides only a certain “weight of evidence” for occurrence of an event of interest. This weight of evidence varies from one occasion to the next and is usually not sufficient for certainty. A decision is made by comparing the current weight of evidence with a pre-determined decision threshold. The system (forecaster, doctor, engineer, etc) asserts a positive result when the evidence exceeds the threshold, and negative when it is less. Figure 1 illustrates this situation.

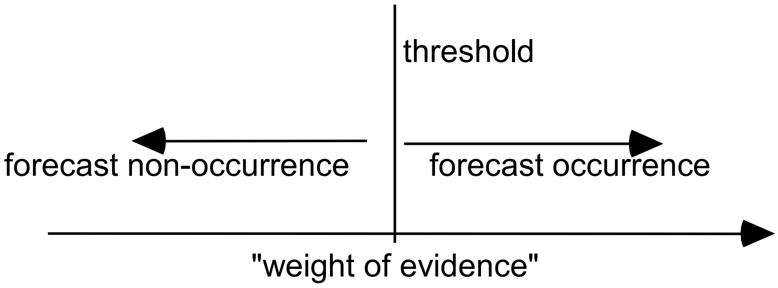


Figure 1: Forecasting a two-state event under uncertainty

There is a similarity to statistical hypothesis testing. The weight of evidence for a hypothesis is represented by a function of the data, usually something like Student’s t or chi squared. Above a critical value a null hypothesis is rejected, and below it, accepted.

The model illustrated in Fig.1 is developed further by supposing that the weight of evidence, which is assumed to be represented by a scalar quantity X, has a certain fixed and known probability density when the event of interest does not occur, denoted $f_0(x)$, and a different distribution $f_1(x)$ when the event does occur, as shown in figure 2.

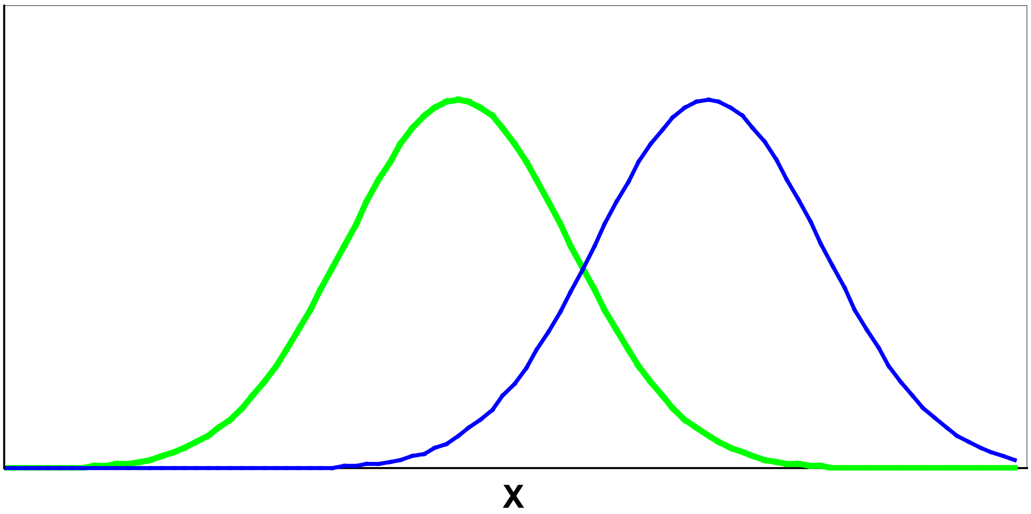


Figure 2: Probability distributions for the “weight of evidence” X prior to non-occurrence, $f_0(x)$, and occurrence, $f_1(x)$, of the event to be forecast. The diagonally hatched area represents the probability of a forecast of occurrence given that the event does not occur, and the horizontally hatched area the probability of a forecast of occurrence given that the event occurs.

The decision threshold is indicated by x*. The form of the distributions does not need to be specified at this stage.

We are in the familiar territory of statistical hypothesis testing, where $f_0(x)$ would be the distribution of a test statistic under the null hypothesis and $f_1(x)$ the distribution of the statistic under a (simple) alternative hypothesis. When the model is applied to signal detection, $f_0(x)$ represents the distribution of incoming data when noise alone is present, and $f_1(x)$ the distribution when a signal is present in addition to noise. In the case of weather forecasting $f_0(x)$ is the distribution of the weight of evidence for a weather event prior to non-occurrence, and $f_1(x)$ prior to occurrence. The event is forecast when the weight of evidence is greater than x*.(X could be something like the Total Totals index in the case of thunderstorms, but in general is simply a scalar variable monotonically related to the likelihood of the event.)

Given the distributions $f_0(x)$ and $f_1(x)$, the location of the decision threshold x* determines some interesting probabilities. The area under $f_1(x)$ to the right of x* represents the probability of a forecast of occurrence given that the event does occur, or a hit or true positive (TP). This area is equivalent to Probability of Detection, POD. The area under $f_0(x)$ to the right of x* represents the conditional probability of a forecast of occurrence given that the event does not occur, ie the probability of a false alarm, sometimes referred to as a false positive (FP). This area is equivalent to Probability of False Detection (POFD). The other two areas, under $f_0(x)$ and $f_1(x)$ to the left of x*, are the probabilities of true negatives (TN) and of false negatives (FN, or misses), respectively. POFD is analogous to the probability of a type 1 error, and POD to the power of a statistical test, or 1 – the probability of a type 2 error.

The probability of each of these four combinations of forecast and event (TN, FN, TP, FP) changes as x* moves along the weight of evidence axis. When the decision threshold is low, ie x* is toward the left end of the X continuum, then the event will be forecast relatively often, practically all of the occurrences will be correctly forecast (POD near 1.0) but there will be many false positives (POFD also approaches 1.0). Just how many false positives is determined by x* and the nature and relative separation of the distributions. Conversely, when x* is towards the right end of the X axis, POFD will be low but so will POD; there will be many misses.

(The forecasts have been assumed to be unequivocal assertions that the event will or will not occur. In the case of forecasts issued as probabilities there are K-1 thresholds on the “weight of evidence” axis, corresponding to the K allowed values for the probabilities (eg 0, 2%, 5%, 10%, ... 100%).)

X is related to the probability of the event. Given the distributional form of $f_0(x)$ and $f_1(x)$, and the climatological (prior) probability of the event, p_c, a numerical value for the threshold x* can be converted into a threshold probability p* by Bayes’ rule:

$$p^* = \Pr\{\text{event}|X=x^*\} = R/(1+R)(1) \qquad \text{(A6)}$$

$$\text{where } R = [p_c/(1-p_c)][f_1(x^*)/f_0(x^*)] \qquad \text{(A7)}$$

There is an optimal location for x* or p* for decision-making which maximises the expected value of the forecasts and which is determined by the relative benefits and costs of the four possible outcomes referred to above. In the well known cost-loss decision model (Thompson & Brier, 1955; Murphy, 1977), p*=C/L, where C is the cost of precautions against the event, and L is the loss if the event occurs and no precautions have been taken.

Calculation of the parameters of the SDT model for yes/no forecasts

If a single set of verified yes/no forecasts is available then, subject to assumptions about the form of f_0 and f_1, the separation of the means of these distributions can be calculated, and also the implied location of the decision threshold.

We assume that f_0 and f_1 are Gaussian in form with equal variances and means separated by d’. (anticipating a little, the assumption that the distributions are Gaussian, or at least Gaussian to within a monotonic transformation, is well supported by data. The variances are not usually exactly equal, but this is disregarded for the present). These assumptions make it possible to make the x scale quantitative, and derive a specific value for the threshold x*, simply by looking up POD and POFD in tables of the standard normal deviate. For example, suppose the verification array is

	event	
	no	yes

forecast	no	247	8
	yes	49	66
			370

Then POD = 66/(8+66) = 0.892 and POFD = 49/(247+49) = 0.166

The standard normal deviates corresponding to these taken as areas under f_1 & f_0 can be found from any table of areas under the standard normal curve. They are -1.237 and +0.970 respectively, in units of the common standard deviation of f_1 and f_0.

Thus the decision threshold is 1.237 units to the left of the mean of f_1 and 0.970 units to the right of the mean of f_0 Hence the separation of the means is 2.207 Figure 3 illustrates what is going on.

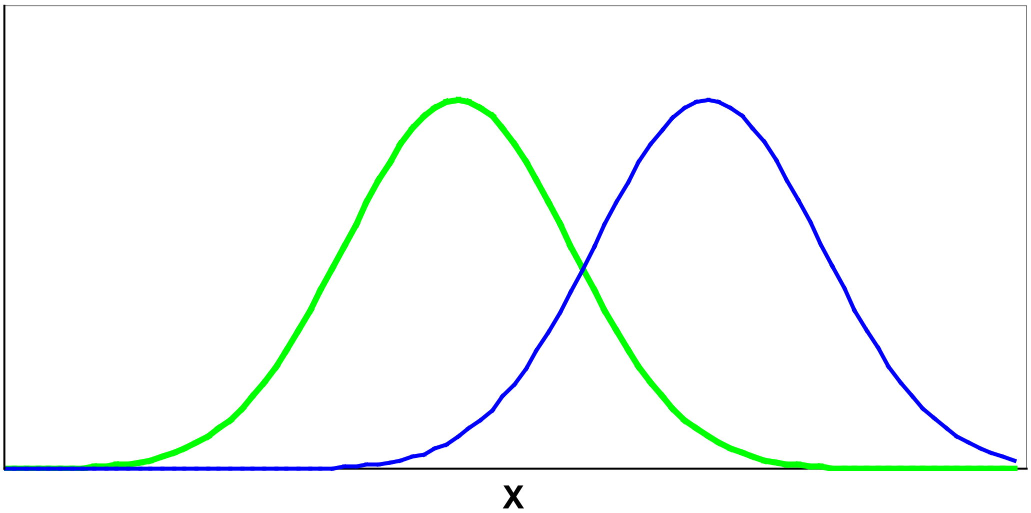


Figure 3: Calculation of d'. $f_0(x)$ is the distribution of X prior to non-occurrences with mean μ_0 and $f_1(x)$ prior to occurrences with mean μ_1. The common standard deviation is 1.0. The implied decision threshold is 1.237 standard deviations to the left of μ_1 and 0.970 standard deviations to the right of μ_0, so the separation of the means, d', is 2.207.

The separation of the means, conventionally denoted d' when the variances are equal, is used as an index of the intrinsic discrimination capacity of a forecasting system. If d' is small then POD and POFD are only slightly different, ie a forecast of occurrence is only slightly more likely before an occurrence than it is before a nonoccurrence. If d' is large then POD»POFD and the system is showing a high level of discrimination capacity.

The location of the threshold, x*, is usually indexed by a likelihood ratio β=f1(x*)/f0(x*), the ratio of the ordinates of the distributions at x*. β can be expressed as a threshold probability p* via the odds form of bayes formula, which gives

$$p^* = \Omega/(1+\Omega), \qquad \text{(A8)}$$

where

$$\Omega = \Omega_O \cdot \beta \qquad \text{(A9)}$$

and

$$\Omega_O = \Pr[E=1]/(1-\Pr[E=1]) \qquad \text{(A10)}$$

(the prior odds on an occurrence).

It is interesting to note that all scores expressible in terms of the elements of the 2x2 verification array can be expressed in terms of POD, POFD and pr[E=1] = pc, the (sample) climatological probability of the event.

The relative operating characteristic

A single set of yes/no forecasts is not sufficient to determine the performance of a forecasting system for all thresholds, due to the need to assume equality of the standard deviations of the underlying distributions. An adequate description of performance requires specification of the ratio of these standard deviations, which in turn requires knowledge of the performance of the system at different thresholds.

The overall performance of a forecasting system for any threshold is determined by the nature and parameters of f_0 and f_1, and can be described empirically for real forecasts by graphing the variation of POD with POFD as x*, or p*, varies, using forecasts for discrete events (eg rain/no rain) issued as ratings of risk or probabilities. The threshold probability is stepped through the range of forecast probabilities used, and values for POD and POFD calculated at each step. Table 1 shows the process, for some rain forecasts for Canberra.

Table 1: Calculation of POD and POFD as functions of threshold probability, p*, for some forecasts of rain in Canberra. N1 is the number of forecasts of the corresponding probability followed by occurrences of rain, N0 is the number followed by no rain. N(N1>p*) is the number of forecasts of rain probability greater than or equal to the corresponding forecast followed by rain. N(N0>p*) is the corresponding quantity for forecasts followed by no rain. POD is the proportion of occurrences of rain preceded by a forecast greater than or equal to the probability in the left column, and POFD the corresponding quantity for no rain.

FCST PROB	N1	N0	N(N1>p*)	N(N0>p*)	POD	POFD
0.00	0	3	82	282	1.000	1.000
0.02	1	47	82	279	1.000	0.989
0.05	5	89	81	232	0.988	0.823
0.10	13	64	76	143	0.927	0.507
0.20	12	38	63	79	0.768	0.280
0.30	16	17	51	41	0.622	0.145
0.40	12	12	35	24	0.427	0.085
0.50	10	8	23	12	0.280	0.043
0.60	4	2	13	4	0.159	0.014
0.70	6	1	9	2	0.110	0.007
0.80	2	1	3	1	0.037	0.004
0.90	1	0	1	0	0.012	0.000
0.95	0	0	0	0	0.000	0.000
1.00	0	0	0	0	0.000	0.000
TOTALS	82	282				

The ROC is a graph of POD (Y axis) against POFD (X axis) for all values of p*. Figure 3 shows the data of table 1 plotted in this way. The form of the resulting curve is purely empirical, determined by the data. While the possible usefulness of these axes is suggested by the SDT model, there is no modelling involved in calculation of the quantities plotted, and no assumptions about underlying distributions.

To orient ourselves on Figure 3(?), it is useful to note several of its properties.

Firstly, the major diagonal represents forecasts which have no skill. In this case learning the forecast does not change one’s opinion about the event. This can be shown using Bayes’ rule, which shows how new information changes the probability of an event. Recalling that POD = Pr{forecast>=p*|event} and

POFD = Pr{forecast>=p*|no event}, and putting p=Pr{event|forecast>=p*} and p_0 = Pr{event} ie the probability of the event before getting the forecast, Bayes' rule in the odds form provides

p/(1-p) = [p_0/(1-p_0)]*(POD/POFD) (A11)

Hence if POD = POFD, the probability of the event is the same after getting the forecast as it was before; the forecasts make no difference to the user's opinion about the probability of rain. It therefore seems reasonable to say that a forecasting system that produces forecasts which plot on the major diagonal of the ROC has shown no skill.

Second, the further the ROC-point is from the major diagonal, the more skilful the forecasts. This is evident from the fact that moving up or to the left either increases POD or reduces POFD, or both.

Perfect skill is indicated by an ROC from 0,0 to 0,1 to 1,1.

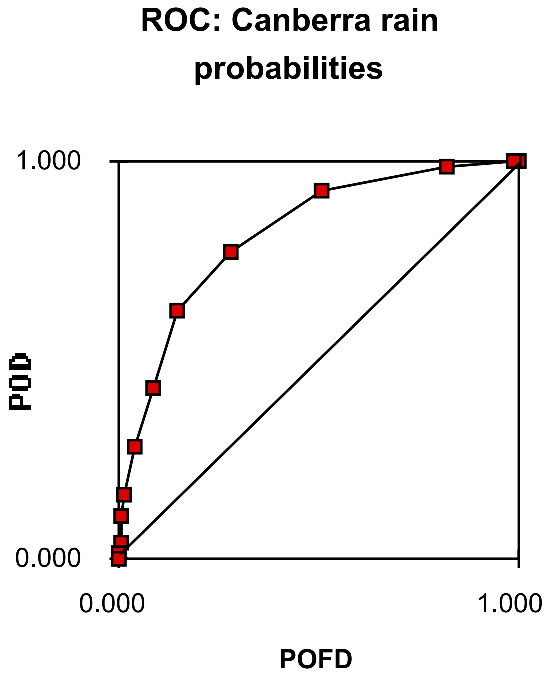


Figure 3: Empirical ROC for probability forecasts for rain issued in Canberra at 06.00 for 12.00 to 18.00 local time, 1987-1995.

The overall performance of the forecasting system, for all threshold probabilities, is indicated by the whole curve. A measure of this is the area under the empirical ROC, when the points are connected by straight lines, sometimes denoted PA. PA ranges from 0.5 for forecasts with no skill, to 1.0 for forecasts with perfect skill.

The bi-normal ROC

Referring to the model in fig 2, it is possible to compare ROCs generated by specific distributions with empirical data like figure 3.

Figure 4 shows a family of ROC curves generated by a moving threshold x* when the distributions are Gaussian with equal variances. The four curves correspond to distributions whose means are separated by 0.5, 1.0, 1.5 and 2.0 standard deviations. The similarity of form with the data of figure 3 is evident.

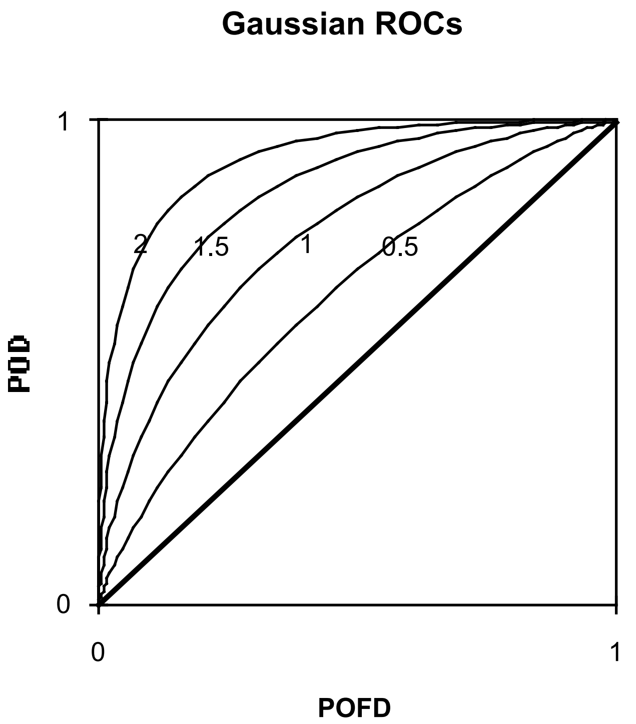


Figure 4: ROCs generated by moving a threshold through equal variance Gaussian distributions with means separated by 05., 1.0, 1.5 and 2.0 standard deviations.

ROCs generated by Gaussian distributions can be linearised by plotting on double probability axes or, equivalently, axes linear in the standard normal deviate corresponding to the probabilities POD & POFD.

Table 2 shows POD and POFD from table 1, with two additional columns showing the transformation to these deviates.

Figure 4 shows the data of fig 3, plotted on axes transformed in this way.

POD	POFD	Z(POD)	Z(POFD)
1.000	1.000		
1.000	0.989		-2.303
0.988	0.823	-2.251	-0.926
0.927	0.507	-1.453	-0.018
0.768	0.280	-0.733	0.582
0.622	0.145	-0.311	1.056
0.427	0.085	0.184	1.372
0.280	0.043	0.581	1.722
0.159	0.014	1.000	2.192
0.110	0.007	1.228	2.453
0.037	0.004	1.792	2.692
0.012	0.000	2.251	
0.000	0.000		
0.000	0.000		

Table 2: Calculation of standard normal deviated of POD and POFD

Figure 4: Data points represent the same data as figure 3, plotted on axes transformed to the standard normal deviates of those in figure 3. The straight line is the ROC generated by the SDT model in which the means are separated by 1.232 (units sd of POFD distribution), and the ratio of the sds of the distributions is 1.097.

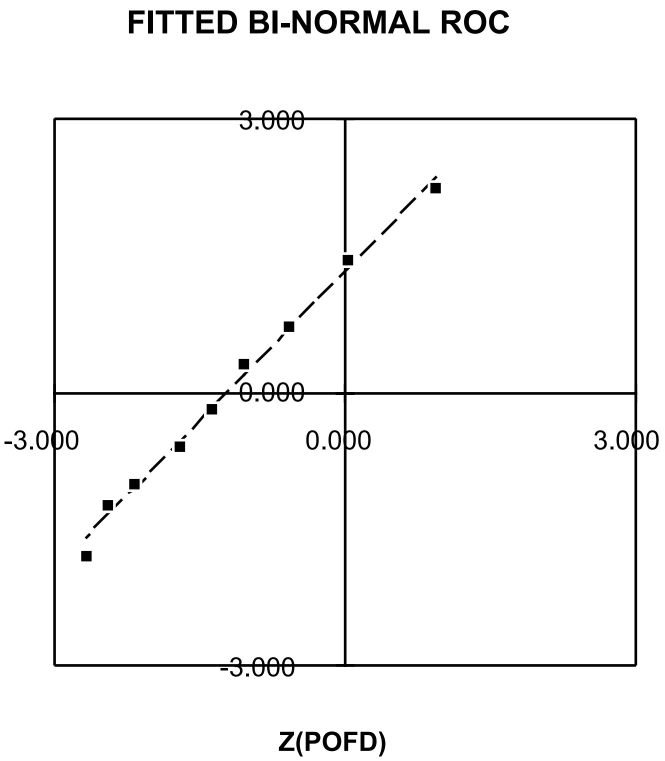


Figure 4: Data points represent the same data as figure 3, plotted on axes transformed to the standard normal deviates of those in figure 3. The straight line is the ROC generated by the SDT model in which the means are separated by 1.232 (units sd of POFD distribution), and the ratio of the sds of the distributions is 1.097.

Figure 4 illustrates one of the few robust empirical findings in the field of forecast verification, that the relationship between POD and POFD as decision threshold changes is very close to linear when plotted on double probability axes. The coefficient of linear correlation for the data in fig 3 is 0.9968. Mason (1982) showed that “bi-normal” ROCs for a wide variety of meteorological predictands follow this linear model. Some ROCs have a slight degree of curvature, but even for these, linearity is a very good first approximation. ROCs from other fields are also generally linear or at worst close enough to make linearity an acceptable practical approximation (Swets, 1986)

The linearity of empirical ROCs plotted in this way supports the use of Gaussian distributions in the SDT model. The dotted line in fig 3 is the ROC generated by the SDT model with Gaussian distributions whose means are separated by 1.232, in units of the standard deviation of the POFD distribution, and in which the ratio of the standard deviation of the POFD to that of the POD distribution is 1.097.

Strictly speaking, the linearity of bi-normal ROCs implies only that the underlying distributions can be transformed to Gaussian form by a monotonic transformation.

Computer programs are available to fit the SDT model to empirical data. There is a FORTRAN listing of one such program, RSCORE, in a text by Swets and Pickett (1982), and the same program can be downloaded from ftp://random.bsd.uchicago.edu//roc/. A significant benefit of RSCORE is that it provides variances for the parameter estimates. Seaman et al (1996) have recently commented on the importance of this for assessment of the significance of apparent differences in skill. The absence of significance tests has been a weakness of most published comparative forecast verification.

Summary measures of skill based on the ROC

A satisfactory description of forecasting skill in the SDT model requires specification of both the slope and intercept of the line representing the system's performance on the bi-normal ROC or in terms of the SDT model the separation of the means of the f_0 and f_1 distributions and the ratio of their variances.

The separation of the means when equality of variances is assumed is denoted d'. When this is not assumed, the separation of the means is usually denoted Δm, and is given by the X-intercept in units of the sd of the f_0 distribution. The slope, s, of the line provides the ratio of the standard deviation of the f_0 distribution to that of the f_1 distribution. Thus a complete description of the system's intrinsic skill is given by the pair of numbers (Δm,s). Given these parameters the ROC can be reconstructed and the system's performance specified for any and all decision thresholds.

There are occasions when a single-number index of skill is desirable. A set of yes/no forecasts only indicates the system's performance at one decision threshold, and hence provides only one point on the ROC. In this case it is necessary to assume a slope s=1.0 for the ROC, and the separation of the means d' provides the index of skill. Empirically, ROC slopes are usually between about 0.7 and 1.3.

Another single number measure of skill, which is recommended by Swets (1986) is the area under the fitted bi-normal ROC transformed back to axes linear in probability. This area is denoted A_z. Figure 5 shows the bi-normal ROC of fig 4 transformed in this way.

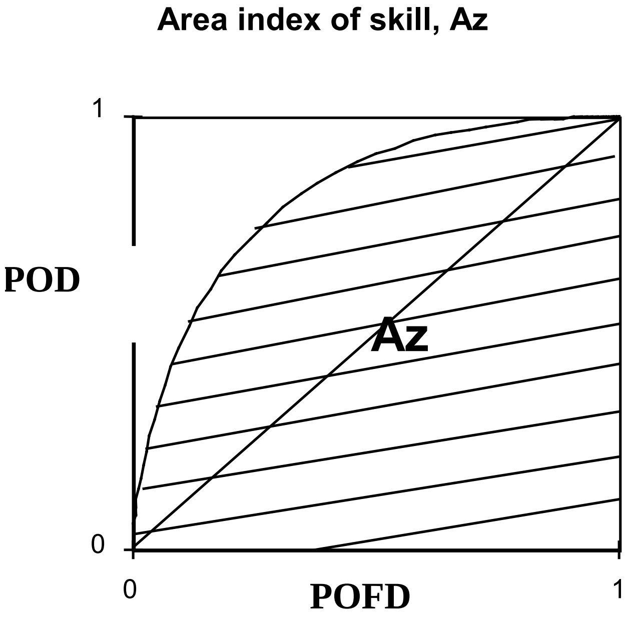


Figure 5: The smooth curve from 0,0 to 1,1 is the straight line of fig 4, generated by moving a threshold through an SDT model with Δm = 1.232 and s = 1.097, transformed to axes linear in probability. The hatched area is Az, a recommended index of forecasting skill

Using the fitted curve minimises random sampling variability, and variability due to differing spacing of data points on different ROCs. Az can be found for any set of data that can be plotted on ROC axes, and thus facilitates the comparison of different kinds of forecasts. An estimate of the variance of estimates of Az is provided by the program RSCORE referred to above.

Az has an interpretation as expected proportion correct in a particular kind of discrimination task known in psychology as a two alternative, forced choice task. In weather forecasting this task would involve presenting a forecaster or forecasting system with a series of paired data sets, one of which was followed by the predictand of interest and the other not, the task being to decide which was which. The experimental advantage of this design is that the decision threshold must correspond to a constant probability of 50%, which eliminates that source of variability.

It is still the case, however, that a complete description of the intrinsic skill of a set of forecasts for a two-state predictand requires both parameters of the ROC (slope and intercept), so a single number index of skill must generally lose some information and Az is no exception to this rule.

Comments and conclusions

The process of formulation of a forecast for a binary weather event can be modelled as a statistical decision. The variation of quantities analogous to the probability of a type 1 error and to the power of a statistical test, derived from verified forecasts, follow closely a model based on the classical theory of statistical hypothesis testing with underlying Gaussian distributions. Application of this model to assessment of the skill of diagnostic systems was developed by psychologists and engineers seeking to measure the capacity of human and electronic observation systems to detect signals in noise.

The ROC, in the weather forecasting context a graph of POD against POFD as decision threshold varies, is a useful way of assessing pure meteorological skill, and provides either a two-parameter description of skill for all thresholds (Δm,s), or a single number index of skill, d’ or Az.

Use of methods from SDT provides a powerful methodology for the assessment of pure skill, in the sense of discrimination capacity, for predictions made in a variety of formats and uncontaminated by variations in the reliability or calibration of forecast probabilities.

Computer programs are available to fit the model, and provide estimates of the variance of fitted parameters, making possible statistical assessment of the significance of differences in skill.

There is a firm basis in classical statistical theory and in numerous empirical studies for the validity of the SDT model in forecast verification, and the power and generality of the results obtainable suggest the time required to become familiar with the use of these methods is likely to be well repaid.

Theoretical and mathematical properties of some measures of forecast quality

Traditional measures of forecasting skill have been criticised in recent years from two theoretical viewpoints. These are Murphy's diagnostic or distributions-oriented framework, and the signal detection approach.

This section outlines some criteria that have been proposed for valid measures of forecast quality from both these points of view, and considers a number of widely used measures in the light of these criteria.

The treatment is not exhaustive. For a thorough review in a DO setting, Murphy (1997) is a good start. Mason (1982b, 1989) and Swets (1986) provide the SDT viewpoint.

Metaverification: assessment of measures of forecast quality

Murphy (1997) identified four criteria that can be used to screen alternative verification measures. His criteria relate to properties or characteristics considered desirable in verification measures. Particular measures may or may not possess any of these properties.

In brief, these criteria are as follows.

Sufficiency

The sufficiency relation (Ehrendorfer and Murphy, 1988) identifies conditions under which one forecasting system can be unambiguously identified as better than another, ie of greater decision-making value to all users. Under certain conditions this relation can be used to screen verification measures. Verification measures that rank forecasting systems in the same order as the sufficiency relation are preferred.

Only one such measure has so far been identified, Krzysztofowicz' (1992) Bayesian Correlation Score (BCS), designed for non-probabilistic forecasts for a continuous variable in which forecasts and observations can be related by a standard linear model. As noted in section 2 of this Report, it should be investigated for applicability to temperature forecasts.

In view of the lack of other such measures, Murphy comments that "this relation (is) of limited use as a screening criterion for verification measures".

ROC analysis suggests that scalar measures consistent with the sufficiency relation may exist only under quite restrictive conditions. One forecasting system (A) is sufficient for another (B) if A's ROC dominates B's in the sense that A's ROC is everywhere above B's. Two parameters for each ROC are required to establish dominance, typically the slopes and intercepts of the ROC's on bi-normal axes. (Levi, 1985). Az and d' are consistent with the sufficiency relation when the slopes of competing ROCs are equal. If the slopes are not equal then the ROCs may cross, implying that one system is preferable at lower thresholds and the other at higher thresholds. The only satisfactory way to compare forecasting systems for binary events is to plot the full ROC.

In general, values for common yes/no scores do not establish sufficiency.

Propriety

Scoring rules are said to be "proper" if their expected value is maximised when the issued forecast is the same as the forecaster's true judgement, and "strictly proper" if this is the only forecast that maximises the expected score (Murphy, 1997).

This criterion applies most directly to scores for subjectively formulated probability forecasts, in situations in which the score is used as feedback to forecasters. The Brier score is strictly proper, and there are some others which are rarely used.

Scores for probabilistic forecasts used in an operational setting should be strictly proper.

Consistency

Consistency is a weak form of the propriety principle applied to non-probabilistic forecasts.

It is based on the idea that, in formulating a non-probabilistic forecast for a continuous variable, forecasters follow an (implied or explicit) rule for collapsing their judgmental probability distribution onto a single value. This might for example be either to forecast the mean or the median of their judgmental distribution.

The consistency principle is that the primary measure used to verify the forecasts should be consistent with the rule used by the forecasters. Forecasting the mean minimises MSE, while forecasting the median minimises MAE.

In general there will not be a lot of difference between these alternatives. Anecdotal evidence suggests that many forecasters actually follow a minimax rule in this situation ("minimise the maximum possible error"). The consistent verification measure in this case is not known, but where the judgmental distribution is more or less symmetrical and bell-shaped the differences between mean, median and minimax forecasts will be small. The particular measure should not have a major impact.

The consistency principle in relation to yes/no forecasts implies that scores which are known to have a strong dependence on decision threshold should be optimised at the same threshold as was used by forecasters making the forecasts. Percent correct, for example, is optimised at a threshold probability of 50%. If the forecaster's threshold probability for issuing a forecast of occurrence is 20%, or indeed any threshold other than 50%, percent correct is an inconsistent score.

Equitability

The equitability criterion applies to non-probabilistic forecasts of discrete variables, and has been mainly applied to scores for yes/no forecasts. The underlying principle is that constant forecasts of any event, or forecasts produced at random, should receive the same expected score (Gandin and Murphy, 1992).

In the case of yes/no forecasts, equitability means that constant forecasts of non-occurrence should get the same score as constant forecasts of occurrence. Hansen and Kuipers' score is equitable. CSI, the Heidke score and percent correct are not.

In the ROC framework any set of constant or random forecasts plots on the diagonal, indicating zero skill (d'=0.0 or Az=0.5), so these measures are equitable.

Equitability has been widely accepted as a criterion for scores for yes/no forecasts but may need further consideration. In the case of rare events it is not obvious that a constant forecast of occurrence, which will almost always be wrong, should get the same score, however low, as a constant forecast of non-occurrence, which will almost always be right.

Scoring rules on the ROC

Mason (1982b) suggested that the ROC provides a criterion for evaluation of scoring rules for yes/no forecasts, and this idea was developed by Swets (1986) using the concept of a "regular" ROC. Mason (1989) illustrated an application to CSI.

All indices of forecasting performance for yes/no forecasts are functions of elements of the 2x2 JD. Since each element of this JD can be expressed in terms of the ROC variables POD and POFD together with the sample probability of the event, all scores can be expressed in terms of these variables. Hence a constant value for a score implies an isopleth on ROC axes.

Equal vales for a score should imply equal skill, so isopleths of scores on ROC axes should correspond with the curves that are known to represent constant skill, that is with so-called "regular" ROCs (Swets 1986). A regular ROC passes through (0,0) and (1,1) and has a monotonically decreasing slope over this range. All empirical ROCs have this form.

All currently popular scores for yes/no forecasts fail this test. Isopleths of POD, FAR, CSI, bias, Hansen and Kuipers' score, the Heidke score and percent correct all plot as straight lines on

ROC axes. Equal values of these measures cannot be regarded as indicative of equal skill, and different values do not necessarily indicate differing levels of skill

A few indices of skill for yes/no forecasts do produce regular ROCs. One of these is Yules' Q (Yule, 1912). Others are the log odds ratio and a measure identified as η , discussed by Swets (1986). These are all consistent with a SDT model using equal variance logistic distributions, and for a single set of yes/no forecasts are monotonically related to d'. They are deficient as general measures of the skill of a forecasting system because of the equal variance property. The slope of the bi-normal ROC, equal to the ratio of the variances of f_0 and f_1 in the SDT model, is observed to vary, so satisfactory assessment of skill must involve a variable parameter equivalent to the slope of the bi-normal ROC.

Scores for yes/no forecasts

Scores for unequivocal predictions of a binary event ("yes/no" forecasts) have a long history of controversy in weather forecasting, starting with Finley's tornado forecasts (Finley 1884) and recently reviewed by Murphy (1996).

The frequency distribution for a set of verified yes/no forecasts is represented by the following table.

		Event		
		No	**Yes**	**Total**
Forecast	**No**	a	b	a+b
	Yes	c	d	c+d
	Total	a+c	b+d	N

Table 1: Frequency distribution for yes/no forecasts. a, b, c, and d are the numbers in each category, and N=a+b+c+d.

Several basic quantities are defined as follows.

Sample relative frequency of the event, r ("sample climate"):

$$r = (b+d)/N \qquad \text{(A12)}$$

It will occasionally be convenient to use the climatological odds, $\omega_0 = r/(1-r)$.

The quantities plotted as X and Y axes on the ROC are here identified as h and f for simplicity of notation. h is equal to POD and f to POFD. In terms of the elements of table x.1 they are given by

$$h = d/(b+d) \qquad \text{(A13)}$$

and

$$f = c/(a+c) \qquad \text{(A14)}$$

POD

POD is a sample estimate of the probability of a forecast of occurrence given that the event did in fact occur. In terms of table 1,

$$POD = d/(b+d) \qquad \text{(A15)}$$

and in terms of ROC coordinates

$$POD = h \qquad \text{(A16)}$$

Thus isopleths of POD on ROC axes are horizontal straight lines. A single value of POD can correspond to a very high level of skill, near the left edge of the ROC square, to zero skill where the isopleth of POD crosses the diagonal, or to “negative skill” near the right hand edge.

POD is sometimes used as a primary indicator of skill. It is unreliable for this purpose because it can vary through its whole range from 0.0 to 1.0 as a result of variations in decision threshold alone, with no change in forecasting skill. There is usually no evidence as to the constancy of decision thresholds. It seems quite likely that this factor varies to some degree between individual forecasters and possibly in individuals at different times. It is rarely considered explicitly.

In terms of Murphy’s (1997) criteria for verification measures, POD is inequitable. Constant forecasts of non-occurrence score zero and constant forecasts of occurrence score one. Random forecasts score somewhere between zero and one, depending on the proportions of events to non-events in the forecasts

FAR

FAR is a sample estimate of the probability of a non-occurrence of the event given that an occurrence was forecast.

In terms of elements of table 1,

$$FAR = c/(c+d) \tag{A17}$$

and in terms of ROC coordinates

$$FAR = 1/(1 + \omega_0.(h/f)) \tag{A18}$$

Isopleths of FAR on ROC axes have the equation

$$h = (1/\omega_0)((1-FAR)/FAR).f \tag{A19}$$

Thus they are a family of straight lines passing through the origin of the ROC, with slope equal to $(1/\omega_0)((1-FAR)/FAR)$. Zero skill, for which h = f, corresponds to FAR = 1-r, and this is the maximum value of FAR. The minimum (best) value of FAR is zero, which may indicate perfect skill or may simply indicate that non-occurrence was never forecast.

FAR is usually quoted in conjunction with POD, in recognition of the fact that an increase in POD may not indicate increasing skill if it is accompanied by an increase in FAR, as is often the case. However, POD does not have a dependence on the sample climate, whereas FAR does, indicated by the appearance of ω_0 in eqn A18. Hence their covariation is not a reliable indication of changes in either skill or decision threshold unless sample climate is unchanged. FAR also has a dependence on decision threshold as noted for POD above.

FAR is inequitable. If non-occurrence is always forecast then c=d=0 and FAR is undefined (eqn x.6), but it can be shown that FAR tends to zero as the threshold probability tends to 1.0.(A threshold probability of 1.0 implies no forecasts of occurrence.) Constant forecasts of occurrence imply h=f=1.0 and FAR = 1-r by eqn A18 above.

Bias

Bias is an indicator of over- or under-forecasting, rather than a measure of skill. It is the number of forecasts of occurrence per actual occurrence.

In terms of table 1

$$bias = (c+d)/(b+d) \tag{A20}$$

and in terms of ROC coordinates

$$bias = (f/\omega_0) + h \tag{A21}$$

Thus isopleths of bias on ROC axes are given by

$$h = bias - f/\omega_0 \tag{A22}$$

so isopleths of bias are straight lines with slope $-1/\omega_0$, always negative, and intercept equal to bias itself.

It is sometimes stated that bias should be near one, ie the event should be forecast with the same frequency as it occurs, but this is not necessarily the case. There is an optimal bias for any specific operation that depends on the optimal decision threshold for that operation. If the economics of the operation indicate that misses are much more expensive than false alarms then a bias above one may be optimal. If false alarms are costly but misses are not so serious then the optimal bias may be below one.

There may be an argument for bias=1.0 in the case of public weather forecasts, where users are assumed to have a wide range of optimal thresholds, but even in this case the dependence of bias on sample climate makes the optimality of bias=1.0 unclear. If the decision threshold is set at the climatological probability, a plausible value for public weather forecasts, then h is approximately equal to 1-f, This is because (assuming equal variances in the SDT model) a decision threshold equal to the prior probability implies a point on the ROC where it intersects the diagonal from (0,1) to (1,0), ie h=1-f. Setting h=1-f=k in eqn A21 gives bias = $(1-k(1-\omega_0))/\omega_0$, which is not in general equal to 1.

Bias is inequitable as a verification measure, is in general not consistent, and assertions that its optimal value is 1.0 should be viewed with suspicion. It is an interesting and useful descriptive statistic.

The dependence of bias on sample climate indicates that caution is required in comparing biases between data sets in which the event occurs at different rates.

Critical success index, CSI

CSI can be regarded as a sample estimate of the probability of a correct forecast of occurrence given that the event was either forecast or observed.

In terms of table 1,

$$\mathrm{CSI} = d/(b+c+d) \tag{A23}$$

and in ROC terms

$$\mathrm{CSI} = h/(1+f/\omega_0) \tag{A24}$$

Thus isopleths of CSI on ROC axes have the form

$$h = \mathrm{CSI} + (\mathrm{CSI}/\omega_0).f \tag{A25}$$

which is a family of straight lines passing through the point h=0, f=$-\omega_0$ with slope CSI/ω_0 and intercept on the h axis equal to CSI itself.

CSI is sometimes presented with POD, FAR and bias in assessment of forecasts of rare events, because these measures can be all calculated without knowledge of the not forecast, not observed frequency, a in table 1. For rare events a can be several orders of magnitude larger than the other elements of table 1, and may be indeterminate since non-occurrence of rare events is not usually forecast explicitly. Mason (1989) discussed this and other problems with CSI.

CSI is problematic is a measure of forecasting skill because it has a strong dependence on both sample climate and decision threshold. Mason (1989) showed that at a constant level of skill as indicated by d’, and at a constant threshold probability, CSI varies through its whole range from 0 to 1.0 as r increases from 0 to 1.0.

If r (climate) and d’ (skill) are constant, CSI varies with threshold probability, and has a maximum at a threshold probability equal to CSI*/(1+CSI*), where CSI* is the optimal value (Mason 1989). Thus CSI is only consistent in Murphy’s sense if the forecaster’s decision threshold is the same as the optimising threshold for CSI, which will in general only happen by chance.

CSI is inequitable. For a constant forecast of occurrence CSI = r, and for a constant forecast of non-occurrence CSI = 0.0. Random forecasts will produce a value of CSI somewhere between these values.

Using CSI to assess changes in skill or to compare forecast sets is not advisable unless its dependencies on sample climate and threshold probability are taken into account.

Hansen and Kuipers' score

In terms of table 1, Hansen and Kuipers' score (HK) is

$$HK = (ad-bc)/((b+d)(a+c)) \qquad (A26)$$

and in ROC terms

$$HK = h-f \qquad (A27)$$

Isopleths of HK on the ROC are thus

$$h = HK + f \qquad (A28)$$

which is a family of straight lines with constant slope equal to 1.0 and intercept equal to HK.

HK does not depend on sample climate, but has a strong dependence on decision threshold, in a similar (but not identical) manner to the dependence described for CSI by Mason (1989). Mason (1979) showed that HK is maximised, at a constant level of skill, by setting the decision threshold at the climatological probability of the event. Unless this source of variability is recognised and allowed for, HK is not a reliable indicator of skill.

HK is not a consistent score.

HK is equitable. Any constant forecast or random forecasts score zero.

The Heidke score

The Heidke score (HD) is proportion correct reduced to account for the proportion expected correct by chance. The adjustment is based on the assumption that the chance forecasts have the same proportion of forecasts of occurrence to non-occurrence as the original forecasts.

Thus

$$HD = (PC–E)/(1-E) \qquad (A29)$$

where

$$PC = (a+d)/N \qquad (A30)$$

and

$$E = ((b+d)/N)((c+d)/N) + ((a+c)/N)((a+b)/N) \qquad (A31)$$

It is difficult to express HD succinctly in terms of ROC axes. One expression is

$$HD = 2r(1-r)(h-f)/(r-r(2r-1)h-(1-r)(2r-1)f) \qquad (A32)$$

so that isopleths of HD are given by

$$h = [(2-(1-1/\omega_0)HD)/(2-(1-\omega_0)HD)].f + (HD/r)/((2/\omega_0)+(1-1/\omega_0)HD) \qquad (A33)$$

Thus isopleths of HD are straight lines with slope equal to the coefficient of f in eqn A33 and intercept equal to the second term.

HD depends on both sample climate and decision threshold, in a similar (but not identical) manner to CSI. It is maximised at a decision threshold between climatology and 0.5, the exact value depending on r and the level of skill (Mason 1979). It is therefore not a reliable indicator of forecasting skill unless both sample climate and decision threshold are constant.

HD is equitable, because constant forecasts and all random forecasts score zero. This is shown by eqn A32. Any forecast for which h=f has HD=0.0.

HD is in general not consistent, for the same reason as CSI and HK. It is optimised at a specific threshold probability which is only fortuitously the same as that used by the forecasting system.

Proportion correct

Proportion correct (PC) is an intuitively appealing measure of performance and has been widely used. PC has however been repeatedly criticised from its first appearance in 1884 (Murphy 1996).

In terms of table 1,

$$PC = (a+d)/N \tag{A34}$$

On ROC axes,

$$PC = (1-r)(1-f) - rh \tag{A35}$$

so that isopleths of PC are given by

$$h = f/\omega_0 + PC/r - 1/\omega_0 \tag{A36}$$

Isopleths of PC are straight lines with slope $1/\omega_0$ and intercept $PC/r - 1/\omega_0$.

PC is maximised at a threshold probability of 50% (Mason 1979). Thus PC is not consistent, except by chance, since a forecaster's decision threshold will not usually be 50%.

PC has a strong dependence on sample climate, shown by eqn x.24. With no skill at all, PC can be as high as 1-r. If r is small (rare event) then a high value for PC can be obtained simply by forecasting non-occurrence all the time.

PC is not equitable. Always forecasting occurrence gives PC = r, and a random forecast can give a value anywhere between r and (1-r).

Measures from SDT

Measures from SDT are discussed at some length in section 8.1. They are fundamentally different from the familiar measures above because SDT indices are based on a model for the process of selection of a forecast.

This model implies that there are two major dimensions to forecasting skill. These are firstly the intrinsic capacity of the forecaster or forecasting system to discriminate between alternative states of the event to be forecast and secondly the decision threshold, that is the level of certainty at which the forecast changes from one state to another.

The systems intrinsic discrimination capacity is assessed using the ROC. The location of a system's ROC may be described by the parameters of the signal detection model (typically the slope and intercept of the linearised ROC on bi-normal axes), or by the area under the ROC on untransformed axes. In the case of yes/no forecasts the slope must be assumed to be unity, so skill is indexed by the single parameter d' (section 8.1), or the area index Az calculated for the model with unit slope.

Decision threshold may be described in several ways. In "traditional" SDT studies it has been indexed by β, the likelihood ratio $f_1(x^*)/f_0(x^*)$ at the decision threshold x* on the decision axis, where f_1 and f_0 are the relevant probability densities. Occasionally x* itself may be quoted. In weather forecasting studies the threshold probability (p*) has been used. All these are equivalent and can be calculated given the parameters of the SDT model. Calculation of threshold probability requires in addition the climatological probability.

Some general comments on scores for yes/no forecasts

There is no completely satisfactory single-number index of skill for yes/no forecasts. The reason is that the dimension of the 2x2 JD is 3; a minimum of 3 quantities is required for a full description of forecast quality even in this simple case. If this dimensionality is not taken account of, variation in the unrecognised parameters may affect measures of skill.

Common indices of forecasting performance are inadequate because all depend on either sample climate or decision threshold.

SDT measures of discrimination capacity (d' or Az) and decision threshold (β or p*) are the least unsatisfactory, as they describe the main dimensions of forecasting performance, have no hidden dependence on climate of decision threshold, and together with the climatological probability satisfy dimensionality considerations. They are still deficient to the extent that they require the assumption that the slope of the bi-normal ROC is unity, but this is a problem with yes/no forecasts rather than with the indices.

The fundamental problem with yes/no forecasts from a verification point of view is that a single set of such forecasts, implicitly done at a single decision threshold, is not sufficient to fully describe the performance of the system at all possible thresholds. This is another aspect of the operational deficiency of non-probabilistic forecasts described by Thompson (1952).

Recommendations

Temperature

Recommendation: that the tabulated values presented for TEMPV be maintained for continuity.

Recommendation: That error distributions be provided for signed and absolute errors, and in cumulative form for absolute errors.

Recommendation: That a combination of persistence and climatology be investigated for use as a baseline for the skill score.

Recommendation: that the scatterplot display include values for slope and intercept of the fitted regression line.

Recommendation: that the facility to extract subsets of the VDS on user-defined thresholds be provided.

Recommendation: that the joint distribution of forecast and observed temperature be provided. for the actual forecasts and for the persistence/climatology baseline forecasts.

Recommendation: That conditional and marginal distributions be available as an option for all temperature forecasts and for the persistence/climatology baseline forecasts.

Recommendation: that the possible usefulness of box plots, conditional quantile plots, and other appropriate graphic displays for verification data, be investigated.

Recommendation: That the measures d',β and Az be available for temperature forecasts collapsed to yes/no forecasts by thresholding on a critical temperature. This temperature should be variable by the user.

Recommendation: That scatterplots of observed vs forecast temperatures should be available for all guidance forecasts, and for the official forecasts, with statistical parameters of the fitted line and values of MAE for each.

Recommendation: that verification information as detailed above for individual forecasters be available at logon at the start of each forecaster's shift.

Recommendation: That MAE be adopted as a single basic measure of temperature forecast accuracy for public information throughout Australia, and be referred to as "average accuracy".

Rain

Recommendation: That the current verification measures used in the AIFS rain FVS be maintained for continuity.

Recommendation: that the full CR and LBR factorisations for the 8x8 and 2x2 JDs be available as an option.

Recommendation: that the SDT-based indices d', β and Az be calculated for rain forecasts as outlined.

Recommendation: That values of d' and β for individual forecasters most recent adequately large sample of forecasts be available at logon, together with the outcome (forecast and observed categories) of their previous forecast.

Recommendation: that the form for RAINV forecasts be incorporated into operational AIFS forms for standard OWR issue times, to reduce the risk that they will be overlooked or done after the start of the validity period of the forecast.

Fire Weather

Recommendation: That use of transformations to normality and box plots be investigated to enhance the clarity of graphical displays of skewed data such as fire danger ratings.

Recommendation: That where a linear model is fitted to data, the parameters of the model and other statistical information as above be provided.

Recommendation: That the capability to select subsets of the VDS be provided.

Recommendation: That an option to download the VDS to standard spreadsheet programs be provided in the AIFS fire weather verification system.

Recommendation: That the primary measure of skill for wind speed and direction and fire danger ratings be in the form of a table of values of d' and β for successive category boundaries on the joint distribution.

TAF verification

Recommendation: that values for CSI be calculated for the 2x2 JD.

Recommendation: that values of the scores presented in AIFS should be accompanied by confidence intervals using the methods discussed in Seaman et al (1996).

Recommendation: that the AIFS system be exhaustively tested on validated data sets before use as the Bureau's official system for TAF verification.

Recommendation: That a system for verifying TTF and Code Grey forecasts be developed.

Recommendation: That the possibility of double counting in combining contingency tables be investigated and eliminated if found.

Recommendation: That verification results be available for the validity period of the TAF in 3-hour blocks as a routine option (in addition to the currently proposed variable forecast age).

Recommendation: That the next revision of the TAF verification module includes verification of temperature and QNH, and consideration be given to real time verification of at least these quantities.

Recommendation: That the reliability of PROB30 and PROB40 forecasts be assessed by extracting the relative frequencies of corresponding forecast events.

Recommendation: That the raw verification data set be available as an option, for both the actual and persistence forecasts, and for subsets to be selectable on user-defined criteria.

Recommendation: That contingency tables include row and column totals, plus the joint relative frequencies.

Recommendation: That 4 (forecasts) x 2 (observation) contingency tables as described in the text be available as an option.

Recommendation: That the components of the CR and LBR factorisations be available as an option for both the actual forecasts and persistence, in both the 2x2 and 4x2 forms of the joint distribution.

Recommendation: That values of d', β and Az be calculated for all 2x2 forecast observed contingency tables and presented with estimated confidence intervals, and that Az be calculated for the 4x2 JD.

7-day forecasts

Recommendation: That the Priestly skill score be used as a summary measure of skill for 7-day outlooks. Smoothed daily means should be used as the climatology.

Recommendation: That (d', β) be used as summary indicators of skill for 7-day rain forecasts prepared as yes/no forecasts. For probabilistic forecasts Az should be used.

MOF

Recommendation: That MOF be verified in a separated AIFS FVS module and summary measure of accuracy for the previous 30 issues be available for each operational issue.

Recommendation: That MAE and bias be used as summary measures of skill for MOF temperature forecasts.

Recommendation: That (d'.β) for the rain/no rain threshold be used as a summary indictor of skill for MOF rain forecasts.

Recommendation: That MAE and bias for wind speed be used as summary indictors of skill for MOF wind forecasts.

Recommendation: That MAE and bias be used a summary measures of performance for the MOF forecasts listed in the text.

References

Barnston, A. G. 1992. Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Weather and Forecasting, 7*, 699-709.

Brooks, H.E. and Charles A. Doswell III, 1996. A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Weather and Forecasting, 11*, 288-303.

Brown, Barbara G., Gregory Thompson, Roelof T. Bruintjes, Randy Bullock and Tressa Kane 1997. Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *Weather and Forecasting, 12*, 890-914.

Buizza, R., T. Petroliagis, T. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi. 1998. Impact of model resolution and ensemble size on the performance of an Ensemble Prediction System. *Quarterly Journal of the Royal Meteorological Society, 124*, 1935-1960.

Centor, R.M. 1991. Signal detectability: the use of ROC curves and their analyses. *Medical Decision Making, 11*, 102-106.

Davis, C.J., 1999. Verification of the extended period temperature forecasts in Canberra. *Sixth National Australian Meteorological Society Conference, 8-11 February 1999, Canberra, Australia*.

Dawes, Robyn M., 1979. The robust beauty of improper linear models in decision making. *American Psychologist, 34*. 571-582.

Doswell, Charles A III, Robert Davies-Jones and David Keller 1990. On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting, 5*, 576-585.

Eckel, F. Anthony and Michael K. Walters, 1998. Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Weather and Forecasting, 13*. 1132-1147.

Ehrendorfer, M. and A.H. Murphy, 1988. Comparative evaluation of weather forecasting systems: sufficiency, quality and accuracy. *Monthly Weather Review, 116*, 1757-1770.

Finley, J.P. 1884. Tornado predictions. *American Meteorological Journal, 1*, 85-88.

Gandin, L.S., and A.H. Murphy, 1992. Equitable skill scores for categorical forecasts. *Monthly Weather Review, 120*, 361-370.

Glahn, H.R., and D.A Lowry, 1972. The use of model output statistics in objective weather forecasting. *Journal of Applied Meteorology, 11*, 1202-1211.

Goodman, L.A and W.H. Kruskal 1959. Measures of association for cross classifications:2. Further discussion and references. *Journal of the American Statistical Association, 54*, 126-163.

Gordon, N. 1993. Verification of terminal forecasts. Fifth AMS International Conference on Aviation Weather Systems. Vienna, Virginia, USA. August 1993.

Graedel, Thomas E. and Beat Kleiner 1985. Exploratory analysis of atmospheric data. In *Probability, Statistics and Decision Making in the Atmospheric Sciences* (Allan H. Murphy and Richard W. Katz, Eds), Westview Press, Boulder, Colorado. 1-38.

Green, D.M. and J.A. Swets, 1966. *Signal Detection Theory and Psychophysics*. Reprinted 1974 Robert E. Kreiger New York. 479pp.

Harvey, Lewis O. Jr, Kenneth R. Hammond, Cynthia M. Lusk, and Ernest F. Mross 1992The application of signal detection theory to weather forecasting behaviour. *Monthly Weather Review, 120*, 863-883.

Krsysztofowicz, Roman 1992. Bayesian correlation score: a utilitarian measure of forecasting skill. *Monthly Weather Review, 120*, 208-219.

Leigh, R.J. 1995. Economic benefits of terminal aerodrome forecasts (TAFs) for Sydney Airport, Australia. *Meteorological Applications*, *2*, 239-247.

Levi, Keith 1985. A signal detection framework for the evaluation of probabilistic forecasts. *Organizational behaviour and human decision processes, 36*, 143-166.

Macmillan, N.A. and C.D. Creelman 1991. Detection Theory: A User's Guide. *Cambridge University Press, Cambridge*.

Mason, I.B. 1979. On reducing probability forecasts to yes/no forecasts. *Monthly Weather Review, 107*, 207-211.

Mason, I.B. 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.B. 1982a. A model for assessment of weather forecasts*. Australian Meteorological Magazine, 30*, 291-303.

Mason, I.B. 1982b. On scores for yes/no forecasts. *Preprints of papers delivered at the Ninth AMS Conference on Weather Forecasting and Analysis*, Seattle, Washington, 169-174.

Mason, I.B. 1989. Dependence of the Critical Success Index on sample climate and threshold probability. *Australian Meteorological Magazine, 37*, 75-81.

Mason, I.B. 1999. Verification of Canberra rain probability forecasts. *Sixth National Australian Meteorological Society Conference, 8-11 February 1999, Canberra, Australia*.

McCoy, Mary Cairns 1986. Severe-storm-forecast results from the PROFS 1983 forecast experiment. *Bulletin American Meteorological Society*, *67*, 155-164.

Murphy, A. H. 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review, 116*, 2417-2424.

Murphy, A. H. 1991. Forecast verification: its complexity and dimensionality. *Monthly Weather Review, 119*, 1590-1601.

Murphy, A. H. 1992. Climatology, persistence and their linear combination as standards of reference in skill scores. *Weather and Forecasting, 7*, 692-698.

Murphy, A. H. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting, 8*, 281-293.

Murphy, A. H. 1995. A coherent method of stratification within a general framework for forecast verification. *Monthly Weather Review, 123*, 1582-1588.

Murphy, A. H. 1996. General decompositions of MSE-based skill scores: measures of some basic aspects of forecast quality. *Monthly Weather Review, 124*, 2353-2369.

Murphy, A. H. 1996. The Finley affair: a signal event in the history of forecast verification. *Weather and Forecasting, 11*, 3-20.

Murphy, A. H. 1997. Forecast verification. In Katz, Richard W. and Allan H. Murphy (eds), The Economic Value of Weather and Climate Forecasts. *Cambridge University Press, Cambridge.*

Murphy, A. H. and Winkler, R.L., 1987. A general framework for forecast verification. *Monthly Weather Review, 115*, 1330-1338.

Murphy, A. H. and Winkler, R.L., 1992. Diagnostic verification of probability forecasts. *International Journal of Forecasting, 7*, 435-455.

Murphy, A. H., 1996b. Forecast verification: a diagnostic approach. *Proceedings of Workshop on Evaluation of Space Weather Forecasts.* Boulder, Colorado; 19-21 June 1996.

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

Murphy, A.H., 1996. The Finley affair: a signal event in the history of forecast verification. *Weather and Forecasting, 11*, 3-20.

Pollack, I. And D.A. Norman 1964. A nonparametric analysis of recognition experiments. *Psychonomic Science, 1*, 125-126.

Popper, Karl, 1968. Conjectures and refutations: the growth of scientific knowledge. Harper, New York, NY. 417pp.

Potts, J.M, C.K. Folland, I.T. Jolliffe and D. Sexton 1996. Revised “LEPS” scores for assessing climate model simulations and long-range forecasts. *Journal of Climate, 34*, 34-53.

Schaefer, Joseph T. 1990.The critical success index as an indicator of forecasting skill. *Weather and Forecasting, 5*, 570-575.

Scurfield, B. K. (1996) Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology 40*, 253-269.

Seaman, R., I. Mason and F. Woodcock, 1996. Confidence intervals for some performance measures of yes/no forecasts. *Australian Meteorological Magazine, 45*, 49-53.

Shanahan, B. 1973. Verification of terminal aerodrome forecasts. Technical Report No. 2, Australian Bureau of Meteorology.

Stanski, H.R., L.J. Wilson and W.R. Burroughs, 1989. Survey of common verification methods in meteorology. Research Report No. 89-5, 114pp, Toronto: Canadian Atmospheric Environment Service.

Stern, Harvey. 1998. An experiment to establish the limits of our predictive capability. 14th Conference on Probability and Statistics / 16th Conference on Weather Forecasting and Analysis. *American Meteorological Society, Phoenix, Arizona, 11-16 January 1998.*

Swets, J.A. 1973. The relative operating characteristic in psychology. *Science, 182*, 990-1000.

Swets, J.A., and R.M. Pickett 1982. Evaluation of diagnostic systems: methods from signal detection theory. *Academic Press, New York*.

Swets, John A. 1986. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin, 99*, 100-117.

Swets, John A. 1996. Signal detection theory and ROC analysis in psychology and diagnostics: collected papers. Lawrence Erlbaum Associates Inc, pp 308.

Thompson, J.C., 1952. On the operational deficiencies in categorical weather forecasts. *Bulletin of the American Meteorological Society, 33*, 223-226.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 688pp.

Ward, M.N., and C.K. Folland 1991. Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperature. *International Journal Of Climatology, 11*, 711-743.

Wilks, Daniel S. 1995. Statistical methods in the atmospheric sciences : an introduction. *Academic Press*, pp 467.

Williams, Dave, 1997. "Naïve" forecasts and their use in deriving temperature skill scores. *Preprints of papers presented at the Forecast and Warning Verification Workshop*, 15-17 September 1997, Melbourne, Australia.

Williams, Dave, 1997. Why wait for AIFS. *Preprints of papers presented at the Forecast and Warning Verification Workshop*, 15-17 September 1997, Melbourne, Australia.

Winkler, R.L. and A.H. Murphy 1985. Decision Analysis. In *Probability, Statistics and Decision Making in the Atmospheric Sciences*, Allan H. Murphy and R.W. Katz, Eds. Westview Press, Boulder and London.

Winterfeldt, Detlof von, and W. Edwards 1986. *Decision Analysis and Behavioural Research*. Cambridge University Press, 604pp.

Woodcock, F. 1976. The evaluation of yes/no forecasts for scientific and administrative purposes. *Monthly Weather Review, 104*, 1209-1214.

Yule, G.U., 1912. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75*, 579-642.

Ian B. Mason

Integrated verification procedures for forecasts and warnings

Report on Consultancy Task 3

Task 3

Make recommendations on the design (*theoretical and mathematical basis* and *output format*) of future AIFS modules for the verification of the following types of forecasts:

- qualitative forecasts of precipitation for capital and provincial cities; and
- wind forecasts and warnings.

Introduction

Background

This is part 3 of a report on integrated verification procedures for forecasts and warnings prepared for the Australian Bureau of Meteorology.

Parts 1 and 2 relate to specific existing and planned verification modules in the Australian Integrated Forecast System (AIFS) and to output formats from AIFS verification modules.

Part 3 is concerned with the design of future AIFS modules for verification of qualitative precipitation forecasts for capital and provincial cities, and verification of wind forecasts and warnings.

Following discussion with Services Policy Branch, the section on qualitative rain forecasts focusses on probabilistic forecasts for simple occurrence/non-occurrence of rain. Categorical quantitative precipitation forecasts in the RAINV format were discussed in section 3 of the first Report.

This Report

Section 2 of this Report briefly reviews the general structure for forecast verification detailed in the first Report. Section 3 is concerned with probabilistic forecasts for occurrence of rain and section 4 with wind forecasts and warnings. Much of the discussion concerns theoretical and mathematical issues. Section 5 contains recommendations for the design of future AIFS modules.

A structured approach to forecast verification

The approach to verification in these Reports is based on Murphy and Winkler's (1987) general framework, with some minor extensions to make it more practical in an operational setting, and to include methods and measures from signal detection theory.

The first Report (section 8) outlined a structure for the process of forecast verification consisting of three stages. These stages are

- data collection
- analysis, and
- communication with stakeholders.

The outcome of the *data collection* stage is the verification data set (VDS), in its simplest form a sequence of matched pairs of forecasts with the corresponding observations. In practice the basic VDS may include other variables of interest in the verification exercise, for example the identity of the forecaster, synoptic type, etc. Derived VDSs may be produced from the basic VDS by selection of cases or transformation of the variables, for example by converting continuous to qualitative variables.

Following Murphy and Winkler (1987) the VDS generates the joint distribution (JD) of forecasts and observations, which is the fundamental data structure for distributions-oriented (DO) verification. The joint distribution is essentially a contingency table in which the elements are the joint probabilities of each combination of forecast and observation. If the VDS is stationary and serially independent in the time series sense then the JD contains all the information relevant to forecast verification.

Analysis in the DO framework proceeds by factoring the JD into conditional and marginal distributions, and producing various measures and graphical displays based on these distributions. Examples are means, variances and covariances or correlations, mean square error or mean absolute error, reliability diagrams and the relative operating characteristic.

In the *communication* stage appropriate verification measures and displays are selected for communication with the intended audience. There are broadly two kinds of user of verification information, administrative and scientific, but within these groups there are subsets with different requirements. For example the administrative group might include the general public, Bureau management, and the Minister. The scientific group includes bench forecasters who need frequent, readily comprehensible feedback on performance, and researchers interested in fundamental issues who can assimilate a high level of detail.

In the consultancy brief two types of output format are envisaged for AIFS. These are

Output format type A: Diagnostic output for individual forecasters, forecasting teams, numerical modelers and the developers of Model Output Forecasts (MOF); and

Output Format Type B: Simpler output for weather services users, the media, Bureau management and Government.

This classification is followed in this Report.

A more detailed discussion of these issues and references to the wider literature is in the first Report.

Qualitative rain forecasts

This section addresses the verification of probabilistic forecasts of occurrence of rain. The methods are applicable to probabilistic forecasts for any two-state event.

The most detailed analysis of probabilistic forecasts using the DO framework is Murphy and Winkler (1992). Comprehensive presentations of the theory of DO verification are in Murphy and Winkler (1987) and Murphy (1997). This Report draws heavily on these papers. Applications of methods from signal detection theory to probabilistic weather forecasts include Mason (1982), McCoy (1986), Levi (1985), Harvey et al (1992), Buizza et al (1998), Palmer et al (1999) and Hamill et al (1999).

This Report canvasses issues related to data collection, analysis and communication with users. Recommendations for design of a verification system for probabilistic rain forecasts are is section 5.

Data collection: the verification data set and joint distribution

The verification data set (VDS) in its simplest form is a list of matched pairs of forecasts and observations, plus any relevant covariates.

It is sometimes convenient to regard the VDS as *derived from* a more comprehensive data set, here referred to as the basic VDS

In general, the basic VDS is as close as possible to the raw data, and may contain values for other variables, for example lead time of the forecasts to the verifying observation, synoptic types, forecaster identity, etc. A theoretical framework has been developed to handle covariates within the DO approach (Murphy 1995), but this aspect will not be pursued in detail in this Report.

Derived VDSs may be obtained as subsets of the basic VDS, by selecting cases on the basis of user-supplied criteria. For example, Harvey et al (1992) examined the effect of forecaster stress on forecasting skill by splitting the VDS into high-stress and low stress samples on the basis of a definition of stress in terms of the amount of weather activity.

A derived VDS may also be obtained by categorising a continuous variable by setting a threshold on the variable, for example by converting observed rainfall to a 0/1 variables to indicate simple occurrence.

It is useful to be able to inspect both the basic VDS and any derived VDS in a verification exercise, for quality control and to assess serial dependence and possible non-stationarity. It would also facilitate analysis if the VDS were downloadable to standard spreadsheet or database systems.

For probabilistic rain forecasts produced by human forecasters the identity of the forecaster should be included in the basic VDS, so that feedback can be provided to individuals on their performance. Significant differences have been found between individuals in the calibration of their probabilities and in forecasting skill (Mason 1997).

Observations of rain should be stored as the actual amount in mm, rather than as a 0/1 variable indicating simple occurrence or non-occurrence, although this is the event being forecast. The distribution of amount of rain with forecast probability may be of interest, in addition to whether or not rain actually occurred. It appears likely that that higher forecast probabilities are associated with larger amounts of rain, as well as with a higher likelihood of occurrence.

It should be possible to select subsets of the basic data set on the basis of user-supplied criteria, for example to extract a derived VDS for specific individuals. It should also be possible to produce a derived data set in which the rain amounts are coded as 0/1.

The joint distribution of forecasts and observations

The joint distribution of forecasts and observations (JD) is essentially a contingency table in which the rows correspond to each allowed forecast and the columns to the observations. For probability forecasts the forecasts are usually restricted to {0%, 5%, 10%,… 90%, 95%, 100%}.

If the forecasts are denoted by f and the observations by x then the elements of the JD are the observed relative frequencies of each combination of forecast and observation, denoted p(f,x). The JD is the starting point for distributions-oriented verification.

Under the assumptions that the time series of forecasts and observations is serially independent and stationary, the JD contains all the information that is relevant to assessment of forecast quality.

As suggested by the notation the relative frequencies p(f,x) are taken as probabilities. They are more correctly regarded as sample estimates of the probabilities, with associated sampling uncertainties. This issue is usually ignored in the verification literature (eg Murphy 1997, p23) and will not be pursued in this Report.

As an example, table 1 is the JD for a set of rain probabilities produced in Canberra between 1987 and 1995. The forecasts were subjective estimates of the probability of occurrence of rain at Canberra airport between 0600 and 1200, prepared by the duty forecaster at 0530 daily.

The elements within the double line are the joint relative frequencies p(f,x). The marginal distribution p(f) is the unconditional distribution of the forecasts and p(x) is the climatological distribution of the observations.

Table 1: Joint and marginal distributions of forecasts and observations. Rain probabilities for Canberra issued 0530K for 0600-1200 1987-1995. N=3287.

	X		
forecast	**0**	**1**	**p(f)**
0	0.120	0.001	0.121
0.02	0.177	0.000	0.177
0.05	0.230	0.002	0.232
0.1	0.218	0.008	0.226
0.2	0.075	0.019	0.094
0.3	0.039	0.012	0.051
0.4	0.015	0.012	0.026
0.5	0.010	0.005	0.015
0.6	0.004	0.009	0.013
0.7	0.002	0.010	0.012
0.8	0.001	0.008	0.009
0.9	0.001	0.006	0.006
0.95	0.001	0.002	0.003
1	0.000	0.014	0.014
p(x)	0.891	0.109	**1.000**

Analysis of forecasting performance

This is the second stage of verification and in the DO framework involves factorisation of the JD and calculation of various verification measures.

Factorising the joint distribution

The information in the JD is more accessible when it is factored into conditional and marginal distributions. There are two ways of factoring a two dimensional JD. These are known in forecast verification as the calibration-refinement (CR) factorisation and the likelihood-base rate (LBR) factorisation.

The calibration-refinement (CR) factorisation

The CR factorisation splits the elements p(f,x) of the JD into conditional distributions of the observations given the forecasts and the marginal distribution of the forecasts,

$$p(f|x) = p(x|f)p(f) \tag{1}$$

The conditional distributions p(x|f) are related to the calibration or reliability of probabilistic forecasts. They provide an answer to the question "Do these forecasts mean what they seem to mean?". If the forecasts are completely reliable or well-calibrated then

$$p(x|f) = f \tag{2}$$

for all f. In the case of probabilistic forecasts the relevant observations are the sub-sample relative frequencies corresponding to each forecast. For well calibrated forecasts the best estimate of the probability of rain given the forecast probability is just that forecast probability. Calibration is important for communication with users of the forecasts, who may base commercial decision strategies on the assumption that the forecasts are well calibrated.

The marginal distribution p(f) shows how often different forecasts are used, regardless of the outcome. p(f) is used to assess the "sharpness" of the forecasts. More skillful forecasts will be more concentrated at the extremes of the range, ie sharper, although the converse is not necessarily true; just because the forecasts are concentrated at the extremes they are not necessarily better calibrated, or more skillful in discriminating between occurrence on non-occurrence of rain. p(f) has no necessary relationship with the event unless assumptions are made about calibration and skill.

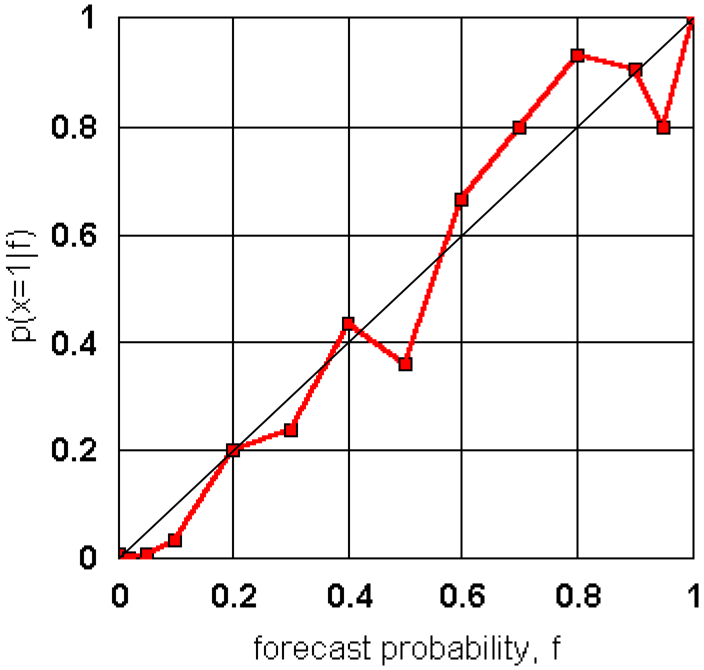
Table 2 shows the components of the CR factorisation for the JD of table 1.

Table 2: Calibration-refinement factorisation. Rain probabilities for Canberra issued 0530K for 0600-1200 1987-1995

forecast	**p(x=0|f)**	**p(x=1|f)**	**p(f)**
0	0.990	0.010	0.121
0.02	0.998	0.002	0.177
0.05	0.991	0.009	0.232
0.1	0.965	0.035	0.226
0.2	0.799	0.201	0.094
0.3	0.762	0.238	0.051
0.4	0.563	0.437	0.026
0.5	0.640	0.360	0.015
0.6	0.333	0.667	0.013
0.7	0.200	0.800	0.012
0.8	0.069	0.931	0.009
0.9	0.095	0.905	0.006
0.95	0.200	0.800	0.003
1	0.000	1.000	0.014

Note that p(x=0|f) = 1-p(x=1|f), since x=0 and x=1 are the only possibilities.

Figure 1: Reliability diagram corresponding to the data of table 2



A graph of p(x|f) against f is known as a reliability diagram or, with some extensions, an attributes diagram (Hsu and Murphy, 1986; Murphy and Winkler, 1992). Figure 1 shows the reliability diagram corresponding to table 2.

p(x|f) is the property of the forecasts that is perhaps of most immediate interest to users. It is the probability of rain given that the forecast probability is f. The forecasts are reliable or well-calibrated if p(x|f) = f for all f. On the reliability diagram this is indicated when all data points lie on the diagonal from 0,0 to 1,1.

The other graphical display used with the CR factorisation shows the marginal or predictive distribution p(f). Figure 2 shows this marginal distribution plotted as a bar graph.

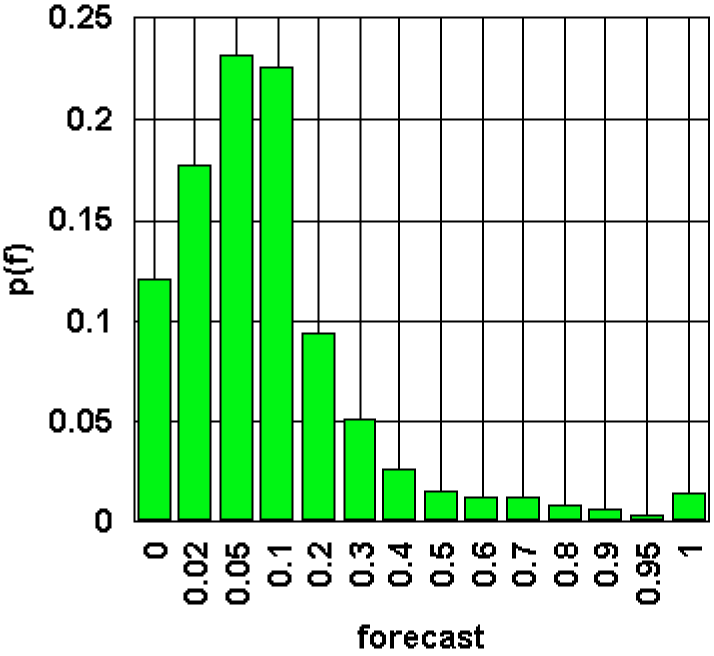


Figure 2: Marginal distribution p(f) for the data of table 2.

The likelihood-base rate (LBR) factorisation

In this case the JD is factored into distributions of the forecasts conditional on the observations and the marginal distribution of the observations:

$$p(f,x) = p(f|x).p(x) \qquad (3)$$

The p(f|x) are likelihoods, and p(x) is the climatological probability, sometimes referred to as the base rate. The term *likelihood* is used in statistics to refer to a probability for data conditional on a hypothesis (the reverse of the quantity usually of interest, the probability of a hypothesis conditional on data). In the present context the forecasts are regarded as data and the hypotheses are that rain will or will not occur.

For binary predictands there are two conditional likelihood distributions, p(f|x=0) and p(f|x=1), which show how often each forecast is given before non-occurrence and occurrence respectively. Comparison of these distributions indicates how successful the forecasting system has been in sorting meteorological situations into those that do and do not precede the event of interest, ie the intrinsic skill or discrimination capacity of the system. More skillful systems produce forecasts in which these distributions are more widely separated.

Table 3 shows the LBR factors of the JD and table 4 shows the corresponding inverse cumulative distributions used to plot the ROC.

Table 3: Likelihood-base rate factorisation. Rain probabilities for Canberra issued 0530K for 0600-1200 1987-1995

forecast f	p(f|x=0)	p(f|x=1)
0	0.134	0.011
0.02	0.199	0.003
0.05	0.258	0.020
0.1	0.244	0.073
0.2	0.084	0.174
0.3	0.044	0.112
0.4	0.017	0.106
0.5	0.011	0.050
0.6	0.005	0.078
0.7	0.003	0.090
0.8	0.001	0.076
0.9	0.001	0.053
0.95	0.001	0.022
1	0.000	0.132
p(x)→	0.891	0.109

These distributions are sometimes plotted as “likelihood diagrams” (Murphy and Winkler 1992). Ideally the distributions p(f|x=0) and p(f|x=1) do not overlap at all. Low probabilities would always be assigned to non-occurrences (eg of rain) and high probabilities to occurrences. Figure 3 shows the data of table 3 plotted in this way. Assessment of skill based on inspection of this type of graph is necessarily somewhat qualitative.

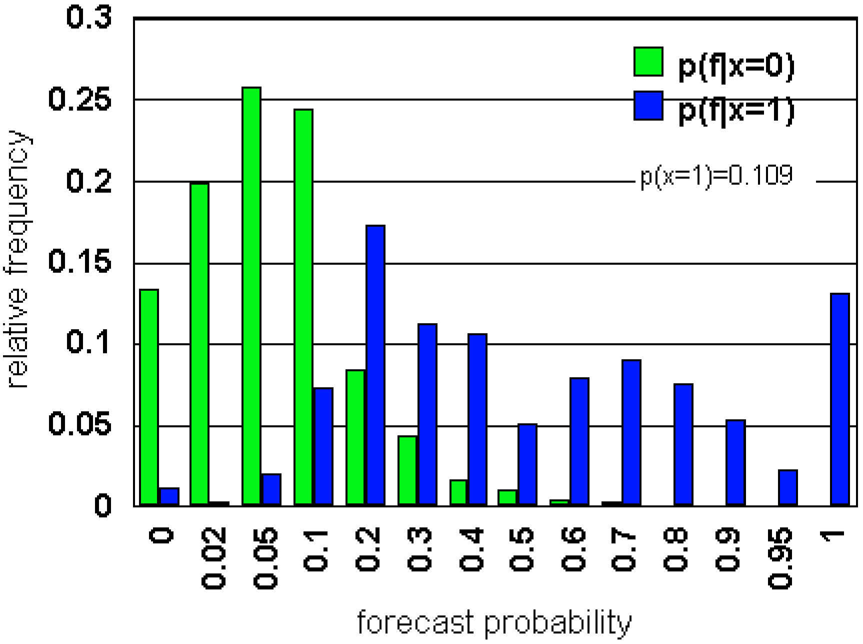


Figure 3: Likelihood diagram for the data of table 3.

forecast f	**p(F≥f|x=0)**	**p(F≥f|x=1)**
0	1.000	1.000
0.02	0.866	0.989
0.05	0.667	0.986
0.1	0.409	0.966
0.2	0.165	0.894
0.3	0.081	0.720
0.4	0.037	0.608
0.5	0.020	0.501
0.6	0.010	0.451
0.7	0.005	0.373
0.8	0.002	0.283
0.9	0.001	0.207
0.95	0.001	0.154
1	0.000	0.132

Table 4: Inverse cumulative distributions from the likelihood-base rate factorisation, used to plot the ROC. Rain probabilities for Canberra issued 0530K for 0600-1200 1987-1995

The ROC is a graph of p(F≥f|x=1) on the X-axis against p(F≥f|x=0) on the y-axis as f is stepped through its range (Mason 1982; see section 3.2.3 below).

The ROC may be plotted in two ways, on linear axes or on axes transformed to the standard normal deviate of the probabilities. The rationale for this transformation is discussed in section 3.2.3 below. Figure 4 shows the ROC corresponding to the data of table 4 on linear axes, and figure 5 on transformed axes.

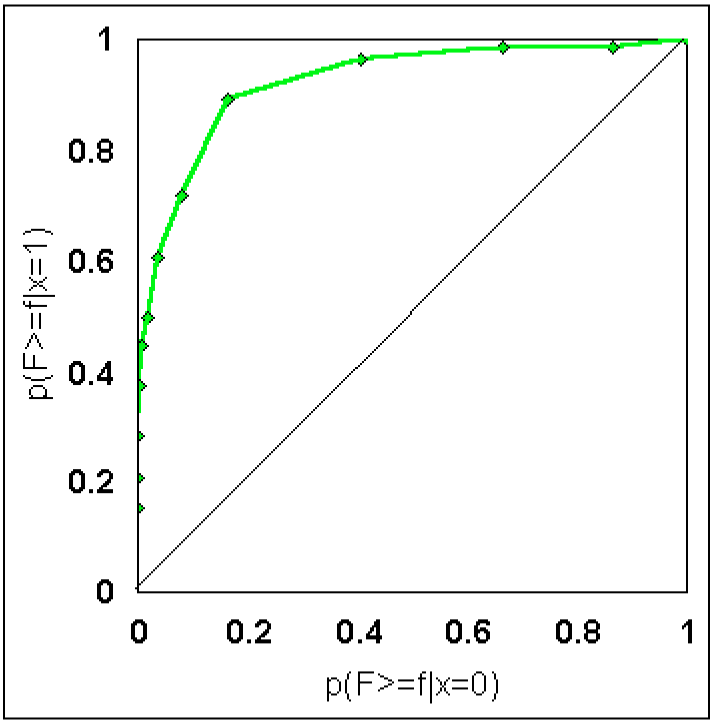


Figure 4: ROC on axes linear in probability, for the data of table 4

Figure 5 illustrates one of the few robust empirical results in the field of forecast verification, that ROCs plotted on axes transformed in this way are close to a straight line. Various measures of forecast skill are based on the ROC.

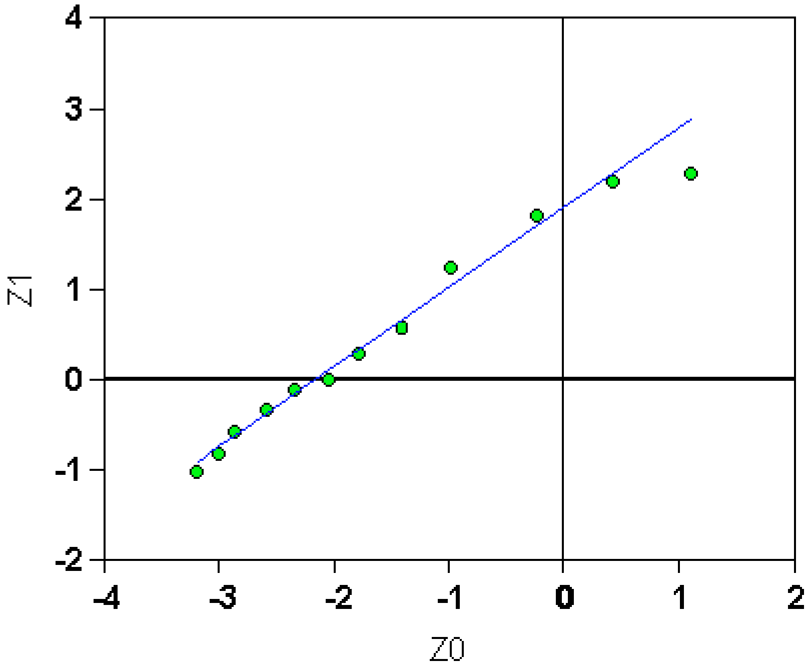


Figure 5: ROC on "bi-normal" axes for the data of table 4. Z0 and Z1 are the standard normal deviates corresponding to the axes of figure 4.

The marginal distribution p(x) is the other element of the LBR factorisation. It is the (sample) climatological distribution of the predictand. For rain/no rain forecasts p(x=1) is the sample relative frequency of rain, referred to as the base rate, and p(x=0) = 1-p(x=1).

p(x) is the only element of either factorisation that does not involve the forecasts in any way. It is a characteristic of the environment in which the forecasting system is operating, rather than of the system itself.

Verification measures for probability forecasts

It is useful to have succinct summary measures of the properties of the forecasts, the observations and their relationships revealed by the JD and factorisations.

Traditionally verification of probabilistic forecasts has relied heavily on the Brier score (Brier 1950), equal to mean square error, and the related skill score. In view of its widespread use, theoretical and mathematical properties of this measure are discussed in some detail below, drawing largely on work by Murphy and Winkler (1992).

Notation

μ_f, σ_f, μ_x, σ_x are the means and standard deviations of the forecasts (subscript f) and observations (subscript x).

$\mu_{f|x}$ is the mean forecast given x, and $\mu_{x|f}$ is the mean of x given f, equal to p(x=1|f) for a binary predictand.

When $x\mathrm{Y}\{0,1\}$, $\mu_x = p(x=1)$ and $\sigma_x^2 = p(x=1).p(x=0) = p(x=1).[1-p(x=1)]$.

ρ_{fx} is the correlation coefficient, defined by

$\rho_{fx} = \sigma_{fx}/\sigma_f\sigma_x$, where σ_{fx} is the covariance of forecasts and observations.

Brier score (mean square error)

The Brier score is just the mean square difference between the forecast probability and the observation indicator, 0 or 1. More rigorously, if the forecast on the ith occasion is considered as a two-component vector $\underline{f_i}=(f_{i0},f_{i1})$, where f_{i0} is the probability forecast for non-occurrence and f_{i1} for occurrence, and the corresponding observation is $\underline{x_i}=(x_{i0},x_{i1})$, where non-occurrence gives $\underline{x_i}=(1,0)$ and occurrence $\underline{x_i}=(0,1)$, then MSE can be defined by

$$\mathrm{MSE} = (1/n)\Sigma_i\Sigma_j(f_{ij}-x_{ij})^2 \qquad (4)$$

where i=1,..,n and j=0,1.

Since $f_{i0}=1-f_{i1}$, $x_{i0}=1-x_{i1}$ and $x_{ij}\mathrm{Y}\{0,1\}$,

$$\mathrm{MSE} = (2/n)\Sigma_i(f_{i1}-x_{i1})^2 \qquad (5)$$

The range of MSE for probabilistic forecasts is [0,2] and smaller values indicate greater accuracy (MSE has a negative orientation).

A linear transformation of MSE occasionally encountered is the quadratic score (QR), defined by

$$\mathrm{QR} = 1-\mathrm{MSE}/2 \qquad (6)$$

QR has some appeal because it has a positive orientation (higher scores indicate greater accuracy) and its range is [0,1].

The form of the Brier score usually encountered is one half MSE, so that it has a range of [0,1].

Partitions of the Brier score: basic partition

Interesting relationships are revealed when the Brier score is expressed as an additive partition based on the factorisations described above. (In this and following sections the notation MSE will be used rather than Brier score).

A basic partition of MSE originally noted by Yates (1982) is

$$\mathrm{MSE}(f,x) = (\mu_f-\mu_x)^2 + \sigma_f^2 + \sigma_x^2 - 2\sigma_f\sigma_x\rho_{fx} \qquad (7)$$

The first term is the square of the bias, that is the difference between the average forecast and the average observation. This expression shows that MSE can be improved (reduced) by identifying and allowing for overall biases in the forecasts.

The remaining terms together are equal to the variance of the forecast errors.

Partitions of the Brier score: CR factorisation

Using the CR factorisation, it can be shown that

$$\mathrm{MSE}(f,x) = \sigma_x^2 + E_f(\mu_{x|f}-f)^2 - E_f(\mu_{x|f}-\mu_x)^2 \qquad (8)$$

(Sanders, 1963; Murphy and Winkler, 1992).

σ_x^2 is the variance of the outcome index, equal to $\mu_x(1-\mu_x)$. This term does not depend on the forecasts in any way. It is equal to MSE for a constant forecast of the (sample) climatological probability of rain. Since the forecasting system has no control over the number of occurrences of rain in the sample it seems inappropriate that assessment of the system's skill should depend on this term.

The second term is the mean square difference between the relative frequencies of the event corresponding to each forecast, and that forecast probability, weighted by the relative frequency of use of the forecast. This term is sometimes denoted REL, for reliability (Murphy and Daan 1985).

The third term is the mean square difference between the relative frequencies for the event corresponding to each forecast and the overall relative frequency of the event, again weighted by the relative frequency of use of each forecast. This term is a measure of the tendency of the sub-sample relative frequencies (estimates of the "true" probabilities of the event given the forecasts) to be "resolved" into two groups, above and below the climatological probability. It is sometimes denoted RES, for resolution.

"Raw" values of MSE are unreliable as indicators of forecast accuracy unless the dependence on σ_x^2 is allowed for. The usual way of allowing for this dependence is through skill scores standardised against climatology (section 3.2.2.6 below).

Partitions of MSE: LBR factorisation

MSE can also be partitioned using the LBR factorisation. In this case it is expressed as

$$\mathrm{MSE}(f,x) = \sigma_f^2 + E_x(\mu_{f|x}-x)^2 - E_x(\mu_{f|x}-\mu_f)^2 \qquad (9)$$

(Murphy & Winkler 1992).

The first term, σ_f^2, is the variance of the forecasts. It does not depend explicitly on the event being forecast.

It is possible to reduce MSE to some degree by the unskilled strategy of "hedging" the forecast probabilities towards their mean value to reduce σ_f^2, although this leads to concurrent changes in the second and third terms which counteract and eventually outweigh the improvement in MSE. The optimal amount of hedging depends on the correlation between forecasts and observations, ρ_{fx} (Barnston 1992; Potts et al 1996).

σ_f^2 is a measure of the sharpness or refinement of the forecasts, the degree to which the forecasts tend to be near to zero or one.

The second term in eqn 16 depends on the differences between the mean forecasts before occurrences and non-occurrences and the outcome index on those occasions. It is optimised (equal to zero) for non-probabilistic forecasts (ie always 0 or 1) which are always correct. Otherwise it acts to increase MSE. This term is a measure of "type 2 conditional bias" (Murphy and Winkler 1992).

The third term involves the differences between the mean forecasts before occurrence and non-occurrence and the overall mean forecast. It vanishes for any kind of random or constant forecasting strategy, for which $\mu_{f|x=0} = \mu_{f|x=1} = \mu_f$. Otherwise it improves (reduces) MSE. This term is a measure of discrimination capacity, but is deficient in that it only involves the means of the conditional distributions. The ROC provides an indication of discrimination that uses the whole distributions and is thus more informative.

Skill scores based on MSE

The "skill" of a forecasting system is often defined as the accuracy of the forecasts relative to the accuracy of forecasts produced by a "no-skill" method. A measure of skill in this sense is the skill score based on MSE, defined as

$$SS(f,g,x) = 1 - MSE(f,x)/MSE(g,x) \qquad (10)$$

where g represents the no-skill forecasts.

The most frequently used no-skill forecasts are climatology and persistence, and occasionally a combination of the two (Murphy 1992; Williams 1997). As a general rule it seems sensible to use the most accurate method that might be available to users in the absence of the official forecasts.

The no-skill standard forecast should be known to forecasters at the time of issuing the forecasts

When climatology is used as a standard for rain probabilities the no-skill baseline is a constant forecast equal to the climatological probability of the event. Murphy usually took as this climatological probability the sample estimate from the VDS itself, μ_x, as this facilitates an interesting expression for the skill score in terms of sample means, variances and the correlation coefficient (Murphy 1996).

In practice, and particularly when SS is used as a scoring rule for daily feedback to forecasters, the sample value would not be known. In the case of rain sample values may be quite variable, affecting σ_x^2 and hence MSE (eqn 8 above). It is therefore preferable to use the long-term climatological probability as a standard. This would be available to users in the absence of the actual forecasts, can be known to the forecaster, and the resulting skill score gives forecasters credit for recognising deviations from long-term climate.

Several different long-term values for the climatological probability have been used to assess skill at one time or another. These include the annual, monthly and daily values, and for daily values alternatives are the actual daily means, calculated for each day separately, and smoothed daily values as found for example by Williams (1997) by fitting harmonics to daily values using least squares.

A constant forecast of the average annual rain probability is a weak standard for skill, particularly where the variation through the year is large. Values of MSE(g,x) in the denominator of the expression for SS would be inflated at times when the annual mean is a poor estimate, making skill appear unrealistically high. Monthly values are better, but there are still unrealistic discontinuities from the end of one month to the beginning of the next, simply due to the change in μ_x.

Actual daily means can also fluctuate unrealistically due to the high natural variability of rainfall.

It appears that the fairest value for climatological probability for use in skill scores is a smoothed daily value (Williams 1997).

There are likely to be difficulties in estimating daily climatological values for rain probability at some locations if the forecast validity period does not coincide with the standard 9 am – 9 am period for rainfall observations. This should not be a problem for major cities at which 3-hourly observations are available. For other locations it will be necessary to estimate the appropriate value.

The other common no-skill standard is persistence. Persistence in the case of rain forecasts is in general an unequivocal forecast of occurrence or non-occurrence, depending on what happened in the preceding forecast period. A more sophisticated persistence forecast might be a probability produced using a Markov model, eg Raible et al 1999.

Murphy (1996) shows that the MSE for persistence forecasts, MSE_p, is equal to

$$MSE_p = 2(1-r)\sigma_x^2 \qquad (11)$$

where r is the sample first order autocorrelation coefficient and σ_x^2 is the sample variance of the observations.

It can be shown (Murphy 1992) that an optimal linear combination of climatology and persistence may be a more stringent standard for skill. This combination is a forecast g given by

$$g = rx^o + (1-r)\mu_x \qquad (12)$$

where x^o is the persistence forecast, μ_x is the climatological forecast and r is the first order autocorrelation coefficient.

The 1992 paper by Murphy, and Williams' 1997 workshop paper should be consulted for further details.

Partitions of the climatological sample skill score

When the sample values are used for relevant means and variances and sample climatology is the no-skill standard, SS can be partitioned as follows (Murphy and Winkler 1992):

$$SS\ (f,x) = \rho_{fx}{}^2 - (\rho_{fx} - \sigma_f/\sigma_x)^2 - [(\mu_f\text{-}\mu_x)/\sigma_x]^2 \qquad (13)$$

The first term is the square of the correlation coefficient, a measure of the degree of linear association between f and x.

In the second term, Murphy & Winkler (1992) show that perfectly calibrated forecasts have $\sigma_f = \rho_{fx}\sigma_x$, so the second term vanishes. This term is thus a measure of the miscalibration of the forecasts, and reduces SS if it is not zero.

The third term involves the difference between the mean forecast and the mean outcome index, and is thus a measure of unconditional bias. Non-zero values reduce SS.

For many forecast sets the second and third terms are quite small compared with the first, so SS is usually close to the square of the correlation coefficient.

Proper scores

Scores used as feedback to forecasters on a daily basis should be *strictly proper*. A scoring rule is *proper* if its expected value is optimised when the forecast probability issued is the same as the forecaster’s judgement of the true probability of the event. A *strictly* proper scoring rule is optimised *only* when the forecast and the forecaster’s judgement coincide. Proper and strictly proper scoring rules are thus defined in the context of “subjective” probability forecasts on single occasions (Winkler and Murphy 1968, Murphy and Daan 1985, Murphy 1997).

If on some occasion the forecaster’s best judgement of the probability of rain is p, his/her issued forecast is f, and the forecasts are being evaluated using a scoring rule S(f,x) so that a score S(f,1) is awarded if rain occurs and S(f,0) if not, then the forecaster’s expected E(p,f) score is

$$E(p,f) = p.S(f,1) + (1\text{-}p).S(f,0) \qquad (14)$$

Assuming positive orientation, S is a *proper* score if

$$E(p,p) \geq E(p,f) \text{ for all } f \neq p \qquad (15)$$

and S is *strictly proper* if

$$E(p,p) > E(p,f) \text{ for all } f \neq p \qquad (16)$$

with equality for f=p.

ie, the maximum expected value of the score is attained only when the issued forecast is equal to the forecaster’s judgement of the true probability.

To show that the Brier score is a strictly proper score, note that for a single forecast the score if the event occurs is

$$S(f,1) = (f\text{-}1)^2 + [(1\text{-}f)\text{-}0]^2 = 2(1\text{-}f)^2 \qquad (17)$$

and similarly if the event does not occur

$$S(f,0) = 2f^2 \qquad (18)$$

so the forecaster’s expected value of BS, written E for brevity, is

$$E = p.2(1\text{-}f)^2 + (1\text{-}p).2f^2 \qquad (19)$$

Differentiating w.r. to f and equating to zero to find the extremum gives

$$\partial E/\partial f = f\text{-}p = 0 \qquad (20)$$

or p=f, ie the value of f at which S has an extremum is equal to p, and this is the only such value of f, so S is strictly proper.

The extremum is a minimum because

$$\partial^2 E/\partial f^2 = 4 \qquad (21)$$

Thus the expected value of MSE is optimised (minimised) if and only if the forecaster’s issued forecast f is equal to his/her best judgement of the probability, p.

There are other strictly proper scores, for example the log score and the spherical score (Winkler and Murphy, 1968). They have been little used in forecast verification. It has been suggested that the log score used as daily feedback to forecasters may be more effective than the Brier score in rectifying a general tendency to overconfidence in subjective probability assessment (eg Fischer 1982), but Mason (1997) found that used over a period of months it actually induced substantial underconfidence. For daily feedback the skill score based on MSE, using a smoothed long-term daily climatology, is probably satisfactory.

Comments on MSE and MSE-based skill scores

MSE is widely used as a performance measure for probabilistic forecasts, but has characteristics which need to be appreciated if weak or misleading conclusions are to be avoided.

The dependence of MSE on sample climate makes unpartitioned values unreliable as performance measures. Comparative verification using MSE should be based on the components of the partitions of MSE or on the skill score. The procedures in Murphy and Winkler (1992) provide a model.

MSE as a performance measure implies a squared error penalty function, ie the “seriousness” of an error is proportional to the square of that error. There appears to be little discussion in the meteorological literature of the appropriateness of this assumption. Many others are possible, for example linear, logarithmic or exponential penalty functions, or penalty functions based on the positions of forecast and observation in the distribution of the observations (the LEPS score; Ward and Folland, 1991), and it is known that scores using different penalty functions may rank forecasts differently (Winkler and Murphy 1968). The apparently privileged position of squared error may ultimately be a result of the ubiquity of the normal distribution in statistical practice, the appearance of a squared difference in the normal probability density function, and the central limit theorem. However, the particular validity of the square of the error as a penalty function in forecast verification is not always obvious and rarely considered.

The skill score based on MSE, SS, is dependent on the no-skill baseline forecasts (eqn 10). On the principle that the most accurate standard of comparison should be used, the optimal combination of climatology and persistence is appropriate. The parameters of this optimal combination for rain forecasts may be dependent on external factors, for example ENSO, and some investigation is desirable before implementation. Comparisons of different forecast sets based on SS need to take the no-skill baseline into account.

SS using sample climatology is widely used, even though this is a weaker standard than long-term climate or climatology/persistence. Nevertheless, it would be useful for purposes of comparison to provide values of SS based on sample climate, and the components of the partition in eqn 13 above.

Recommended output format: summary performance measures

Each of the partitions of MSE and the skill score contains useful information, and should be provided in a full DO analysis of probabilistic forecasts. The following form suggested for the output of summary measures based on MSE is based on Murphy and Winkler (1992). The specific values are those for the Canberra rain probabilities (table 1 above).

Table 5: Summary measures: joint distribution

means		variances		correlation coefficient	sample size
μ_f	μ_x	σ_f^2	σ_x^2	ρ_{fx}	N
0.136	0.097	0.039	0.097	0.669	3287

Table 6: Summary measures: LBR factorisation

means	variances	sample sizes

$\mu_{f\|x=0}$	$\mu_{f\|x=1}$	$\sigma^2_{f\|x=0}$	$\sigma^2_{f\|x=1}$	N(x=0)	N(x=1)
0.090	0.516	0.015	0.241	2930	357

Note that the means for the CR factorisation, $\mu_{x|f}$ are available directly from the table (table 2), and since xY{0,1}, the variances are $\sigma^2_{x|f} = \mu_{x|f}(1- \mu_{x|f})$.

Table 7: Components of the basic partition of MSE

MSE =	$(\mu_f – \mu_x)^2$	$+ \sigma_f^2$	$+ \sigma_x^2$	$- 2\sigma_f\sigma_x\rho_{fx}$
(accuracy)	(bias)	(variance of forecast errors)		
0.054	.00076	0.039	0.097	0.082

Table 8: Components of MSE associated with the CR factorisation

MSE =	σ_x^2	$+ E_f(\mu_{x\|f} – f)^2$	$- E_f(\mu_{x\|f} – \mu_x)^2$
(accuracy)	(uncertainty)	(reliability)	(resolution)
0.054	0.097	0.002	0.045

Table 9: Components of MSE associated with the LBR factorisation

MSE =	σ_f^2	$+ E_x(\mu_{f\|x} – x)^2$	$- E_x(\mu_{f\|x} – \mu_f)^2$
(accuracy)	(refinement)	(“discrimination-1”)	(“discrimination-2”)
0.054	0.039	0.033	0.018

Table 10: Components of the MSE sample skill score

SS =	ρ_{fx}^2	$- [\rho_{fx} – (\sigma_f/\sigma_x)]^2$	$- [(\mu_f – \mu_x)/\sigma_x]^2$
(skill)	(association)	(calibration)	(bias)
0.439	0.448	0.001	0.008

Verification measures from signal detection theory

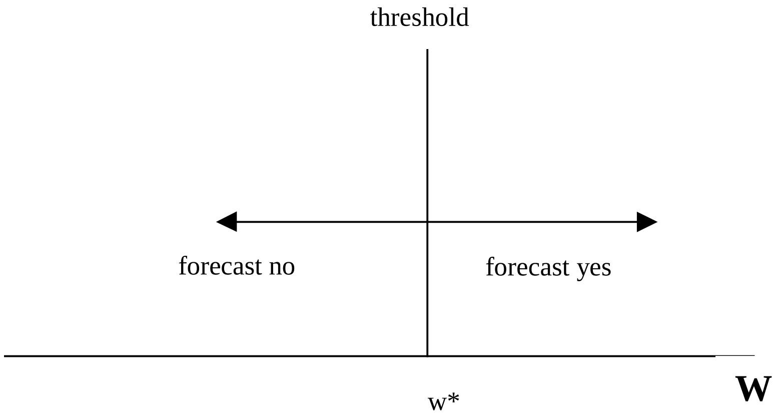
This section is a much abbreviated outline of the basis for certain verification measures for probabilistic forecasts based on the signal detection model.

Presentations or applications of this model in a meteorological context can be found in Mason (1980, 1982a,b, 1989. 1997b), Levi (1985), McCoy (1986), Harvey (1992), Harvey et al (1992), Buizza et al (1998), Palmer et al (1999), and Hamill et al (1999), and there are some comments in Stanski et al (1989) and in Murphy (1997). A good review of the historical development is in Swets (1973). John Swets was one of the pioneers in the field and a collection of some of his papers is available (Swets 1996). A standard text on the methodology, possibly more accessible than Swets, is Macmillan and Creelman (1991).

Forecasting under uncertainty: the signal detection model

In the simplest case of non-probabilistic forecasts for a two-state weather event it is assumed that the forecast (“yes” or “no”) is selected on the basis of a threshold on a scalar quantity related to the weight of evidence for the event. If the weight of evidence is greater than the threshold then the event is forecast, and if less then not forecast. Figure 5 illustrates this situation

Figure 5: W represents the weight of evidence for a weather event and w is the threshold for forecasting the event.*



The model postulates a certain fixed and known probability density for W when the event does not occur (the "noise alone" distribution), and a different distribution when the event does occur (the "signal plus noise" distribution). This is illustrated in figure 6

The horizontally hatched area in figure 6 represents the probability of a forecast of occurrence given that the event occurs, which is estimated by POD in a set of verified yes/no forecasts. The diagonally hatched area represents the probability of a forecast of occurrence given that the event does not occur, which is estimated by POFD ("Probability of False Detection", not to be confused with FAR).

If the distributions are assumed to be Gaussian with equal variance, then POD and POFD from a set of verified forecasts can be used to estimate the separation of the means (labelled Δm in the figure but often denoted d' when the variances are equal) and the location of the threshold, w*, in units of the common standard deviation of the distributions. Computational details can be found in the first Report (section 8.1.5.1).

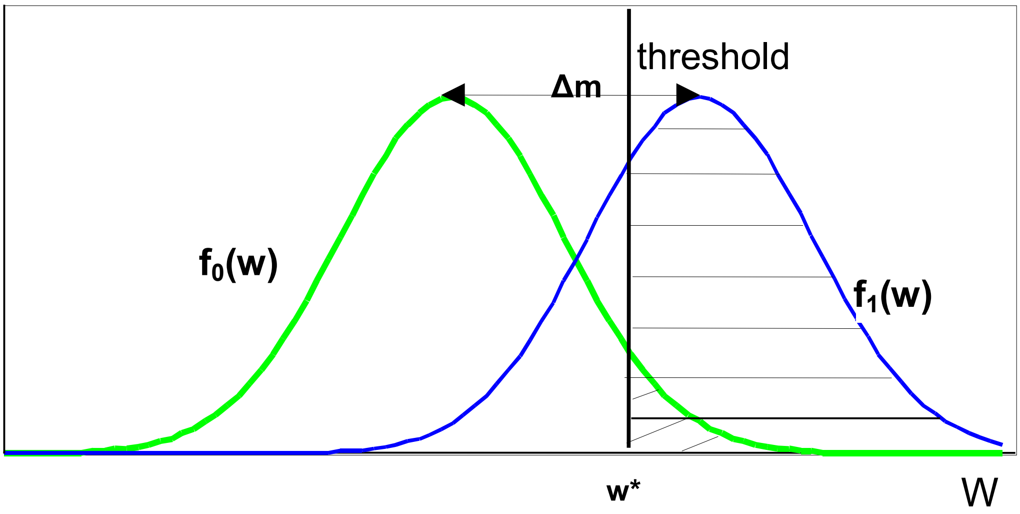


Figure 6: Assumed distributions of weight of evidence before non-occurrence of a weather event, f0(w), and before occurrence, f1(w). The separation of the means is Δm and the ratio of the standard deviations (not shown) is s. The horizontally hatched area represents the probability of a forecast of occurrence given that the event occurs and the diagonally hatched area the probability of a forecast of occurrence given that the event does not occur.

If the forecasts are given as probabilities then there is a sequence of thresholds on W. If there are k allowed forecast probabilities, so that the forecast f is selected from a set $\{p_i\}$; I=1,..k then there are k-l thresholds on W, and probability pj is forecast if $w^*_{j-1} \leq w < w^*_j$. Figure 7 illustrates this.

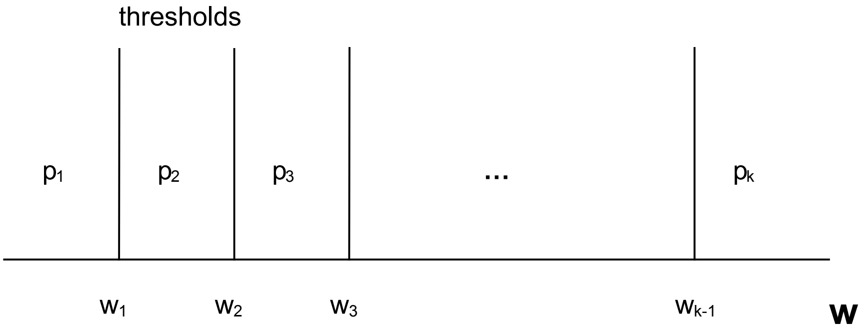


Figure 7: Thresholds for selection of probabilistic forecasts.

Fitting the model

The parameters of the model are Δm (the separation of the means of the f0 and f1 distributions in figure 6 above), s (the ratio of the standard deviations of the distributions) and $\{w^*_j\}$; j=l,..,k-1 (the set of thresholds corresponding to the allowed probabilities).

The most convenient means of estimating the values of these parameters from a set of verified probability forecasts is to use standard software. A number of packages are available. Possibly the most highly developed is ROCKIT developed by C.E. Metz at the University of Chicago and available free on

http://www-radiology.uchicago.edu/krl/toppage11.htm .

A FORTRAN listing of an early version of this program (“DORFALF”) is in Swets and Pickett (1982). The ROCKIT package provides ML estimates of the SDT model parameters and their variances, together with goodness-of-fit statistics. ROCKIT was developed primarily to analyse designed experiments in the medical field, but is readily adaptable to weather forecasts. The FORTRAN code is available.

Quick “first look” estimates can be obtained by plotting the Z-transforms of POD and POFD obtained when a threshold probability is stepped through the allowed set of forecasts, producing the so-called “bi-normal” ROC illustrated in figure 5 above. Δm is then the X-intercept of the line of best fit, $s = \sigma_0/\sigma_1$ is the slope of the line, and the w^*_j are the x values (z-transforms) of the POFDs. Figure 8 illustrates the geometry.

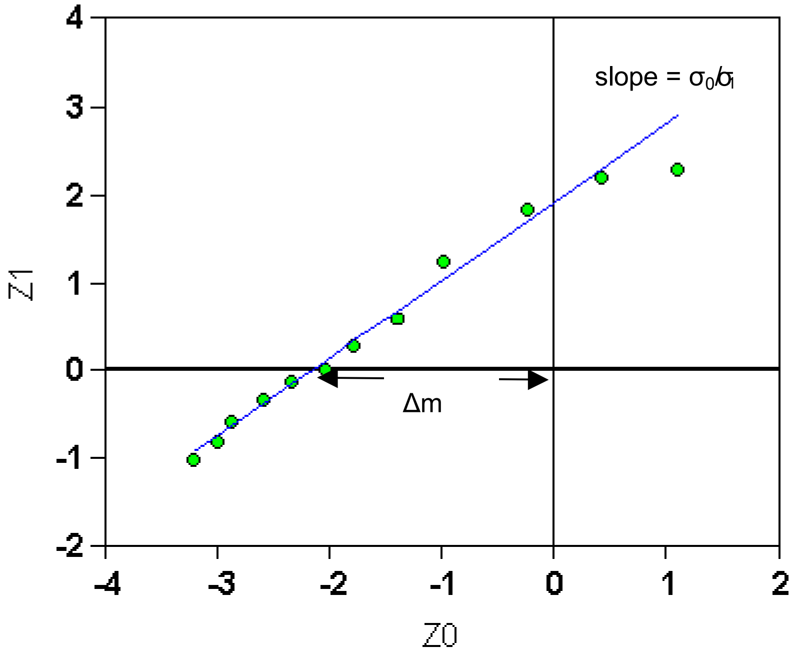


Figure 8: Estimation of the parameters of the SDT model from a fitted straight line. Z0 is the standard normal deviate corresponding to POFD and Z1 the standard normal deviate corresponding to POD. The w_{*j} would be found as the values of Z0 corresponding to the data points.

Measures based on the SDT model

(Δm,s)

Δm and s together are sometimes used as verification measures for diagnostic systems. The whole ROC curve can be reconstructed given values for these parameters.

Δm is the quantity most directly related to forecasting skill. Its lowest value is zero when the forecasts show no capacity to discriminate. Δm=3.0 indicates a high level of skill and would correspond to percent correct above 95% (the actual value depending on other factors including sample climate).

The slope parameter s is required for an adequate description of the data, but its meaning in practical terms is unclear. Typically s varies between about 0.5 and 1.5.

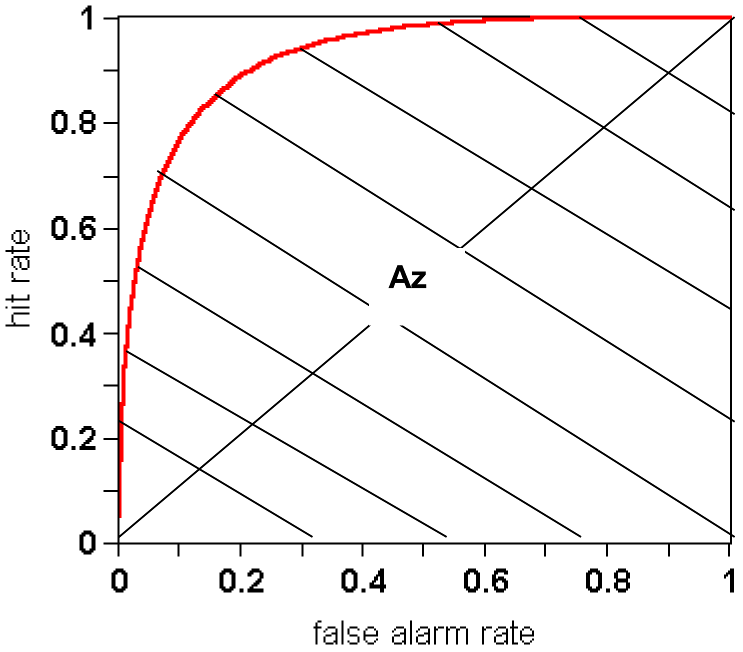
Swets (1986) regards (Δm,s) more as a data summary than an accuracy index.

Az

The most satisfactory single-number measure of discrimination capacity is Az, the area under the fitted ROC on bi-normal axes, transferred back to probability axes.

Figure 9 shows the fitted straight line of figure 8 transformed in this way. ROCKIT provides estimates of Az together with the variance of the estimate.

Figure 9. Fitted ROC on axes linear in probability. The hatched area is the SDT verification measure for discrimination capacity, Az.



The lowest value of Az is 0.5 for forecasts that show no discrimination for the event, when the ROC lies along the diagonal from 0,0 to 1,1. Perfect discrimination is shown by Az=1.0, when the ROC goes from 0,0 to 0,1 to 1,1.

It can be shown (Green and Swets 1966) that Az is equal to expected percent correct for an experimental design known as "two-alternative, forced-choice". In a weather forecasting context this would be a test situation in which paired sets of data were presented to a forecaster or forecasting system, one of each pair being followed by (say) rain and the other by fine weather. The task is to choose (unequivocally) the data set followed by rain. This design eliminates variability associated with sample climate and decision threshold, both of which must be 0.5.

For explanation to non-specialists Az expressed as a percentage might be loosely described a standardised form of percent correct.

There is a number of other performance measures based on SDT, for which see Swets (1988).

Communication with users

Output format type A: Diagnostic output for individual forecasters, forecasting teams, numerical modelers and the developers of Model Output Forecasts (MOF);

Output Format Type B: Simpler output for weather services users, the media, Bureau management and Government.

Type A output

Two levels of output should be provided for category A users. These are

1. Comprehensive output to provide a complete description of all aspects of the forecasts. This should comprise

- listings of the basic verification data set and any derived data sets,
- the joint distribution with CR and LBR factorisations,
- graphical output of p(f), the reliability diagram, likelihood diagrams and the ROC in linear probability and bi-normal form,
- the Brier score and skill score together with components of the various partitions,
- (Δm,s) for the SDT model with estimated variances, and

- Az and estimated variance.

2. Output suitable for daily feedback on performance to forecasters and forecasting teams. The skill score based on squared error is recommended for this purpose, ie

   $SS = 1 – (f-x)^2/(c-x)^2$

   where f is the forecast probability for rain, xY{0,1} is the observation and c is the corresponding long-run climatological probability of rain. SS can be expressed as percent improvement over climatology. It is strictly proper when c is long-run climatology, but only approximately so when sample climate is used.

   Investigation of the use of an optimal combination of climatology and persistence is recommended.

   At regular intervals, as sufficient data accumulates, reliability diagrams should be available for each forecaster. How much data is sufficient is difficult to define, but about 100 forecasts is probably the minimum.

Type B output

There are two main features of probabilistic forecasts that require description for non-specialists. These are calibration and skill or discrimination.

The REL component of the CR partition of MSE (section 3.2.2.4 above) is most often used to describe calibration, although it does not provide an indication of whether the forecasts are over- or under-confident. In fact there is no completely satisfactory summary measure of calibration, and for Category B users it is desirable to show the calibration diagram (figure 1) with a brief description in lay terms.

The summary measure of skill most often encountered is the skill score SS based on MSE (section 3.2.2.6), expressed as % improvement over climatology. It can be simply explained as percent improvement in the accuracy of the forecasts over unchanging forecasts of the average probability of rain. Az is more theoretically satisfying as a measure of discrimination, but is more difficult to explain to non-specialists.

Output recommended for Category B users (weather services users, the media, Bureau management and Government) is

- The REL component of the CR partition of MSE.
- Reliability diagrams should also be provided to management, and possibly to other B users at the discretion of management. The reason for this caution is that the sample relative frequencies plotted in reliability diagrams are unstable for small samples, and can be misleading for non-experts.
- The MSE-based skill score SS using long-run climatology as the baseline.

Wind forecasts and warnings

Verification of wind forecasts and warnings is complex, due to the variability in both forecasts and observations.

The Report suggests a general framework for implementation in AIFS. It should provide a basis for further development

In order to produce some valid results in a reasonable time, this Report will be restricted to situations in which forecasts and observations can be regarded as referring to a single location, and in which the forecasts, warnings and observations are available in AIFS.

There may be significant problems in parsing free-form forecasts and warnings to extract wind direction and speed, but these are beyond the scope of this Report

The procedure recommended follows the 3-stage verification process outlined in section 2. Most of the difficulties arise in extraction of the basic verification data set, that is in extracting a representative set of forecast, observed data pairs.

Upper wind forecasts for aviation are not considered in this section.

Some issues in verification of wind forecasts and warnings

This section outlines a number of issues in verification of wind forecasts and warnings. The list is not exhaustive but does demonstrate the complexity of the task.

Variable times of issue and validity periods

Warnings may be issued at any time, for any period of time. Verification thus has three periods to consider for each warning situation; from the start of the warning validity period or onset of wind above the warning threshold (whichever is earlier), the period in which warning and above threshold winds coincide, then the period to cessation of the validity period of the warning or cessation of above threshold winds (whichever is later). This is further complicated by the existence of multiple thresholds (strong wind, gale, storm, hurricane).

Discontinuities: variable times of forecast and observed changes

Similar considerations to the previous paragraph apply to forecasts of discontinuities in wind speed or direction. If the forecast and observed time of the change do not coincide then the periods before and after both observed and forecast time of change must be considered, and these can be at any time.

Availability of verifying observations

Observations may be sparse or absent for long stretches of coastline and for large ocean areas. There is no simple solution to this problem.

The basic verification data set should be flexible enough to enable “pseudo-observations” to be entered manually, as estimated by meteorologists or from objective analyses. Pseudo-observations should be clearly identified as such in the VDS and their presence in the data set flagged in subsequent analyses.

Difference between formats of forecasts and observations

Wind forecasts are normally given as ranges of both speed and direction, eg NW-W 12-18 kts, whereas the observations are in general in point values of degrees and knots, eg 350° 13kts.

Variable warning areas

The geographical area requiring a warning varies according to the meteorological situation. An adequate verification system should be able to cope with variable areas in both forecasts and observations.

The no/no problem

Warning situations are rare, and at the higher winds speeds very rare. When a joint distribution of warnings issued and warning required is produced in the form of table 11, the frequency of no warning/no requirement, p(f=No,x=No) in table 11, is typically much larger than the other elements, sometimes by several orders of magnitude.

Warning required→ Warning issued ↓	No	Yes
No	p(f=No,x=No) large	p(f=No,x=Yes) small
Yes	p(f=Yes,x=No) small	p(f=Yes,x=Yes) small

Figure 11: General 2x2 joint distribution for warning verification.

Difficulties in assessing the skill of forecasts in this situation are well-known (eg Mason 1989). Most performance measures are dominated by the high no/no frequency, so that the usually more interesting variations in the other frequencies have little impact. Many (but not all) of the correct no/no forecasts are trivially easy, and if they are all included they give the forecasts an unrealistic impression of skill. The usual solution to this problem has been to calculate performance measures that do not involve the no/no frequency, for example CSI, POD, FAR and bias. This is unsatisfactory as some at least of the no/no forecasts do require skill, and a valid assessment of skill requires knowledge of how well the system copes with forecasting non-occurrences in marginal situations. It will not do to just throw them all away.

A possible solution is to use some external criterion to select a VDS which contains, in addition to all the issued warnings and all the observed occurrences, those non-occurrences which could or should have required “a difficult conscious choice” not to issue a warning (Alford 1997). Just what this criterion should be in the case of wind warnings is a matter for further discussion. One possible starting point could be to include all occasions on which the observed wind was just below the threshold for issuing a warning. How far below is “just below” is again something that needs further investigation, as it should be low enough to include all the false alarms.

Verification of vector forecasts

Wind is a vector, so calculation of errors in forecasts suggests use of vector differences between forecasts and observation. This may perhaps be appropriate in the case of upper wind forecasts for aviation, where the operationally significant quantity is the component of the wind along the flight path. In the case of surface winds however wind speed and direction may have different effects on users of forecasts, with speed usually of more interest than direction. It therefore seems appropriate to verify these as separate forecasts.

Other issues

Alford (1997) gives a list of issues in severe thunderstorm verification many of which also apply to wind warnings, or indeed to any rare and severe event. Garske (1997) also gives a very useful account of warning verification in the Queensland region, highlighting problems and solutions which could have wider application.

This Report

This Report suggests a framework for verification of wind forecasts and warnings that should provide a flexible basis for further development. The basis is a comprehensive and flexible verification data set, from which a variety of analyses can be generated as required.

What AIFS can do

The strength of AIFS in forecast verification is the access it provides to data bases of forecasts and observations. This Report is therefore confined mainly to situations in which the relevant forecasts, warnings and observations are available in an AIFS archive.

It is recognised that adequate verification of warning situations, in particular, may involve non-standard observations (eg from members of the public or newspaper reports). The system needs the facility to archive such reports and retrieve them readily.

The VDS

The basic VDS should be as comprehensive as possible, and flexible in that subsets can be extracted using user-supplied criteria. Transformations of the variables should be possible.

In addition, it should be possible to attach text notes to specific occasions in the VDS so that non-standard observations or other information can be kept in readily accessible form associated with the relevant time.

It should be possible to download the VDS to standard spreadsheets to take advantage of non-AIFS data analysis systems (eg statistical packages).

The basic VDS should include all forecasts and observations (at least) at one hour intervals, with data at intermediate times when appropriate, for example for issue and finalisation time and start and end of the validity period of warnings, and for observations times at which warning thresholds are crossed in either direction.

The variables to be part of the basic VDS may need some further discussion. As a reasonable minimum the following are suggested.

- date/time of issue of forecast or warning
- date/time of forecast validity
- forecast direction and speed
- observed direction and speed
- date/time of finalisation of warning
- date/time when observed wind crosses thresholds for warnings and corresponding direction and speed
- comments (text)

Wind forecasts should be saved exactly as issued, eg direction NW-W speed 15-20 kts. Observations similarly should be saved as received, eg direction 310°, speed 21 kts, rather than categorised.

Textual comments could be entered by forecasters in real time or later by the officer responsible for warning verification, possibly with prompts based on Queensland (Garske 1997) or other forms.

Where forecasts are for a fixed area containing several wind observing stations, all stations should be included in the VDS.

Some comments on verification of warnings

Some important aspects of warnings do not fit readily into a formal DO verification framework, (although some may be derivable from a comprehensive VDS). Important aspects include

- lead time for warnings (from time of issue of first warning to development of warned event),
- duration of warning
- time of issue of last warning,
- duration of above-threshold winds,
- warned area vs. observed area,
- comments (can be anything, including damage reports, public criticism, the forecast situation etc)

The verification system should be able to include this information linked to any formal VDS, so that it is not lost.

The joint distribution

Following Murphy and Winkler's general framework, the joint distribution of forecasts and observations provides the basis for verification.

Direction

A similar layout to that proposed for verification of wind in the AIFS Fire Weather module (first Report) is recommended. Table 12 illustrates the proposed categorisation of direction. The system should provide flexibility to change these categories if required.

Observed (deg)→ Forecast ↓	350-070	080-160	170-250	260-340	Calm or L&V	**p(f)**
N, N-NE, NE, NE-E						
E, E-SE, SE, SE-S						
S, S-SW, SW, SW-W			p(f,x)			
W, W-NW, NW, NW-N						
Calm or L&V						
p(x)						

Table 12: Proposed format for joint distribution of forecasts and observations of wind direction.

Speed

The joint distribution for wind speed could use the categories in table 13 below. The category boundaries correspond approximately to the Beaufort scale, but should be variable by the user.

Observed → Forecast ↓ km/h	≤19	20-29	30-39	40-62	63-75	76-87	≥88	**p(f)**
≤19								
20-29								
30-39			p(f,x)					
40-62								
63-75								
76-87								
≥88								
p(x)								N

Table 13: Proposed categories for a joint distribution for wind speed

Warnings

At the simplest level warnings can be summarised in a 2x2 JD (table 14)

Warning required→ Warning issued ↓	No	Yes
No	p(f=No,x=No)	p(f=No,x=Yes)
Yes	p(f=Yes,x=No)	p(f=Yes,x=Yes)

Table 14: Simple JD for warnings

Tables like this should be produced for each warning type.

Analysis

The DO analysis proceeds in two stages, extraction of joint and marginal distributions in the CR and LBR forms, and calculation of summary verification measures.

The general form of the CR and LBR factorisations has been covered in the first Report, particularly section 8.

CR and LBR factorisations should be available for JDs for wind direction and speed, and also for the 2x2 warning JD.

Verification measures and graphical displays

Forecasts

Joint distribution

The JD for direction suggested above is 5x5, and for speed, 7x7. An appropriate graphical display is the bivariate histogram, as used by Murphy et al (1989) to display the joint distribution of forecast and observed temperatures. Box plots should also be available.

Histograms of the marginal distributions p(f) and p(x) should also be available, accompanied by the mean values for forecast and observed direction, μ_f and μ_x and the corresponding standard deviations σ_f and σ_x.

There is no satisfactory summary measure of association for multicategory contingency tables (Murphy 1997). A correlation coefficient ρ_{fx} can be calculated, and multi-category forms of Hansen and Kuipers and the Heidke score are available (Wilks 1995). However, any single-number index must lose detail, particularly when it is considered that the dimensionality of a 5x5 JD is 24; 24 joint probabilities (or joint and marginal probabilities) are required to completely describe the original JD. The dimensionality of the JD for speed is 48.

In the first report a procedure was recommended to cope with high dimensionality to some extent by collapsing a NxN JD into N-1 2x2 JDs by thresholding at successively higher category boundaries, and summarising each of these by the signal detection measures d’ and β. d’ is a measure of discrimination capacity and β indexes the decision threshold (first Report section 8.1.5.1). This is recommended for wind speed, but is questionable in the case of a JD for circular variables like wind direction, because the meaning of the boundaries is unclear.

For want of a satisfactory alternative, it is recommended that the correlation coefficient be provided with the JD for wind direction.

MSE is widely used as a summary measure of performance, so values of MSE should be provided for both speed and direction, including components of the basic partition as described by Murphy (1997),

$$\mathrm{MSE}(f,x) = (\mu_f - \mu_x)^2 + \sigma_f^2 + \sigma_x^2 - 2\sigma_f\sigma_x\rho_{fx} \qquad (22)$$

$(\mu_f-\mu_x)^2$ is a measure of unconditional bias. σ_f^2 and σ_x^2 are measures of the variability in the forecasts and observations respectively and $\sigma_f\sigma_x\rho_{fx}$ is a measure of the association between forecasts and observations. Each of these terms can very independently of the others.

The skill score based on MSE using long-run climate as the standard should also be provided, and also components of Murphy’s (1997) partition of the score using sample climate.

$$SS(f,c,x) = 1 - MSE(f,x)/MSE(c,x) \quad (23)$$

where c represents a no-skill baseline, in this case forecasts of the long-run climatological mean.

Murphy's partition is

$$SS(f,\mu_x,x) = \rho_{fx}^2 - [\rho_{fx} - (\sigma_f/\sigma_x)]^2 - [(\mu_f - \mu_x)/\sigma_x]^2 \quad (24)$$

where ρ_{fx}^2 is a measure of the linear association between forecasts and observations, $[\rho_{fx} - (\sigma_f/\sigma_x)]^2$ is a measure of reliability or "conditional bias", and $[(\mu_f - \mu_x)/\sigma_x]^2$ is a measure of overall bias.

CR factorisation

The CR factorisation is described in section 8.1.2 of the first Report.

This factorisation provides (sample estimates of) the probabilities of observing wind direction or speed within the specified ranges given that it was forecast to be in a certain range. This is one of the verification statistics of most interest to both forecasters and users.

Summary statistics that should be provided with this factorisation are $\mu_{x|f}$, that is the mean observation corresponding to each forecast range f, and also a measure of the dispersion of the observations about this range. The conditional standard deviations $\sigma_{x|f}$ may be appropriate, although care is necessary in dealing with circular quantities.

Appropriate graphical displays would be histograms and box plots of observed directions and speeds corresponding to each forecast category, similar to those presented by Brown and Murphy (1987).

The components of the CR partition of MSE should also be provided,

$$MSE_{CR} = \sigma_x^2 + E_f(\mu_{x|f} - f)^2 - E_f(\mu_{x|f} - \mu_x)^2 \quad (25)$$

E_f denotes expectation with regard to the sample distribution of forecasts. The three terms in this partition are measures of the climatological uncertainty, reliability and resolution respectively. Further discussion of the meaning of these terms for non-probabilistic forecasts of continuous quantities is in Murphy (1997) and Murphy et al, (1989).

LBR factorisation

The LBR factorisation is discussed in section 8.1.2 of the first Report.

The distributions p(f|x) show discrimination capacity, broadly, the propensity of the forecasting system to issue different forecasts before different observations. This is the property of the forecasts most directly related to pure meteorological expertise; it shows the ability of the system to sort meteorological situations into groups in which the observed value of the predictand has different values.

Appropriate graphical displays would be histograms and box plots of forecast directions and speeds against observed categories.

Summary statistics that should be provided with this factorisation include the conditional means and standard deviations, $\mu_{f|x}$ and $\sigma_{f|x}$.

The LBR partition of MSE should also be provided,

$$MSE_{LBR}(f,x) = \sigma_f^2 + E_x(\mu_{f|x} - x)^2 - E_x(\mu_{f|x} - \mu_f)^2 \quad (26)$$

E_x denotes expectation with regard to the sample distribution of observations. The three terms on the RHS measure the spread of the forecasts, a form of conditional bias, and discrimination, respectively. Further discussion of the meaning of these terms is in Murphy (1997) and Murphy et al (1989).

Warnings

The basic JD for warnings is a 2x2 array of the form of table 15. Marginal probabilities should be provided.

Warning required→ Warning issued ↓	No	Yes	p(f)
No	p(f=No,x=No)	p(f=No,x=Yes)	p(f=No)
Yes	p(f=Yes,x=No)	p(f=Yes,x=Yes)	p(f=Yes)
p(x)	p(x=No)	p(x=Yes)	(N)

Table 15: General 2x2 joint distribution for warnings

As discussed in section 8 of the first Report, the most satisfactory measures of performance are the SDT measures d' and β, or if a single number is required, Az.

There are many common measures of accuracy for 2x2 tables, and for comparison with other forecasts it may be desirable to provide values for the following:

POD, FAR, bias, POFD, CSI, Hansen & Kuipers' score, the Heidke score, proportion correct.

Computational formulas are in the first Report.

Adequate verification of warnings requires considerably more information than is contained in the 2x2 JD. Ideally the basic VDS is comprehensive enough to provide at least the following information.

- Distribution of warning lead times (time from issue of warning to first observation of warning conditions)
- Distributions of waning finalisations (time from last observation of warning conditions to time of finalisation of warning)
- Distribution of durations of warnings
- Distribution of durations of warning conditions
- Duration of warning validity period not coincident with observations of warning conditions ("false alarm" periods)
- Duration of warning validity period coincident with warning conditions
- Duration of no warning/no observation period
- Duration of observations of warning conditions not coincident with a warning validity period ("miss" periods).

Anecdotal or non-AIFS data

The basic VDS should be able to link to text files containing comments or non-standard observation that can be entered by RFC staff in real time. These kinds of data are sometimes lost or forgotten in the period between the warning situation and verification.

Communication with users

Category A

Output recommended for Category A users has two levels.

As a "first look" option, the joint distributions for direction, speed and warnings should be provided, together with descriptive statistics (μ_f, μ_x, σ_f, σ_x, ρ_{fx}) and values for the skill score standardised against long-run climate.

The second level includes the full DO verification and all statistics, graphs and verification measures.

Category B

As a single summary statistic for wind forecasts, the MSE skill score using long-run climate should be provided, for both direction and speed, referred to as “improvement over climatology”. It may be necessary to provide some further explanation for non-specialists.

For warnings, presentation of the 2x2 JD alone may be satisfactory. If a single number is required, d’ is recommended. It can be referred to simply as a measure of forecasting skill, with the information that zero represents forecasts with no skill and values above about 4.0 indicate almost perfect performance

Recommendations

Qualitative precipitation forecasts

A verification system for probabilistic precipitation forecasts should have the following components:

Data

A comprehensive basic verification data set should be available, containing in addition to the sequence of matched pairs of forecasts and observations enough additional information to fully characterise the forecasting situation.

The basic data set should be flexible, so that subsidiary data sets can be derived from it by selection of cases on user-defined criteria or by transforming variables

Analysis

The joint distribution of forecasts and observations should be available, together with the CR and LBR factorisations, since these are the basic data structures for DO verification. The means and variances corresponding to each of these factorisations should be available, and the correlation coefficient.

Analyses and summary measures should include

- MSE in the raw form and components of the partitions described in the text
- The skill score SS based on MSE using both long-run and sample climate as baselines for skill. Use of the optimal combination of climatology and persistence as a no-skill baseline should be investigated. Components of the partition of SS should be available.
- STD-based measures (Δm,s) and Az, together with variances.
- The sharpness diagram (p(f) histogram), reliability diagram and likelihood diagram.
- The ROC on both linear probability and “binormal” axes.

Communication

Category A

Two levels of output should be provided for category A users. These are

1. Comprehensive output to provide a complete description of all aspects of the forecasts. This should comprise
   - listings of the basic verification data set and any derived data sets,
   - the joint distribution with CR and LBR factorisations,
   - graphical output of p(f), the reliability diagram, likelihood diagrams and the ROC in linear probability and bi-normal form
   - the Brier score and skill score together with components of the various partitions
   - (Δm,s) for the SDT model with estimated variances, and
   - Az and estimated variance.

2. Output suitable for daily feedback on performance to forecasters and forecasting teams. The skill score based on squared error is recommended for this purpose, ie

$$SS = 1 – (f-x)^2/(c-x)^2$$

where f is the forecast probability for rain, xY{0,1} is the observation and c is the corresponding long-run climatological probability of rain. SS can be expressed as percent improvement over climatology.

Investigation of the use of an optimal combination of climatology and persistence is recommended.

At regular intervals, as sufficient data accumulates, reliability diagrams should be available. How much data is sufficient is difficult to define, but about 100 forecasts is probably the minimum.

Category B

For category B users there are two main features of probabilistic forecasts that require description. These are calibration and skill or discrimination.

The REL component of the CR partition of MSE is most often used to describe calibration, although it does not provide an indication of whether the forecasts are over- or under-confident. In fact there is currently no completely satisfactory summary measure of calibration, and at least for management it is desirable also to show the calibration diagram with a brief description in lay terms.

The summary measure of skill most often encountered is the skill score SS based on MSE, expressed as % improvement over climatology. It can be simply explained as percent improvement in the accuracy of the forecasts over unchanging forecasts of the average probability of rain. Az is more theoretically satisfying as a measure of discrimination, but is more difficult to explain to non-specialists.

Output recommended for Category B users (weather services users, the media, Bureau management and Government) is

- The REL component of the CR the skill score based on MSE.
- Reliability diagrams should also be provided to management, and possibly to other B users at the discretion of management. The reason for this caution is that the sample relative frequencies plotted in reliability diagrams are unstable for small samples, and can be misleading for non-experts.
- The MSE-based skill score SS using long-run climatology as the baseline.

Wind forecasts and warnings

Data

The main recommendation relating to verification of wind forecasts and warnings is that the basic verification data set should include hourly values of forecast and observed wind speed and direction, plus values at any intermediate times at which either forecasts were issued or wind speed passed through one of the warning thresholds.

The basic VDS should also include whatever additional information is necessary to adequately describe the situation, including for warnings the facility to link to text files containing comments or non-standard observation that can be entered by RFC staff in real time. These kinds of data are sometimes lost or forgotten in the period between the warning situation and verification (section 4.4.1).

It should be possible to extract subsidiary VDS by selection on the basis of user-supplied criteria, or by transformation of variables.

Separate joint distributions should be available for direction, speed and warnings (section 4.4.2) together with the corresponding CR and LBR factorisations.

Analysis

As a summary performance measure for wind direction the correlation coefficient is recommended (4.5.1.1.1).

For wind speed it is recommended that the 7x7 JD be collapsed into a sequence of six 2x2 JDs by thresholding on successive category boundaries, and the SDT measures (d',β) be calculated for each threshold. The correlation coefficient should also be provided.

For the 2x2 warning JD d' and β are recommended as summary performance measures.

Values for the more familiar "scores", POD, FAR, bias, POFD, CSI, Hansen & Kuipers' score, the Heidke score, and proportion correct should also be provided.

Values for MSE with components of the basic, CR and LBR factorisations should also be available.

The skill score based on MSE should be available (4.5.1.1.1), using both long-run climatology and sample climate as the no-skill baseline, and use of the optimal combination of climatology and persistence should be investigated.

Graphical displays should include histograms for marginal distributions of both forecasts and observations with descriptive statistics (μ_f, μ_x, σ_f, σ_x, ρ_{fx}), bivariate histograms (eg Murphy 1997) and box plots for wind direction and speed. The box plots should be based on the raw observations, rather than categorised.

Communication

Category A

Output recommended for Category A users has two levels.

As a "first look" option, the joint distributions for direction, speed and warnings should be provided, together with descriptive statistics (μ_f, μ_x, σ_f, σ_x, ρ_{fx}) and values for the skill score standardised against long-run climate.

The second level includes the full DO verification and all statistics, graphs and verification measures.

Category B

As a single summary statistic for wind forecasts for B users, the MSE skill score using long-run climate should be provided, for both direction and speed, referred to as “improvement over climatology”, with climatology represented by an unchanging forecast of average conditions.

For warnings, presentation of the 2x2 JD alone may be satisfactory, with cells indicating numbers of occasions rather than relative frequencies. If a single number is required, d’ is recommended. It could be referred to simply as a measure of forecasting skill, with the information that zero represents forecasts with no skill and values above about 4.0 indicate almost perfect performance.

References

Alford, Phil 1997. Developing an improved approach to severe thunderstorm advice and warning verification in Australia. *Forecast and Warning Verification Workshop, Bureau of Meteorology, Melbourne*, 15-17 September 1997.

Brier, G.W., 1950 Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*, 1-3.

Brown, B.G, and A.H. Murphy 1987. Quantification of uncertainty in fire weather forecasts: some results of operational and experimental programs. *Weather and Forecasting, vol*, pages.

Garske, Ray 1997. Notes on warning verification in Queensland: current practice… future possibilities? *Forecast and Warning Verification Workshop, Bureau of Meteorology, Melbourne*, 15-17 September 1997.

Green, D.M. and J.A. Swets 1966. *Signal Detection Theory and Psychophysics*. Reprinted 1974 Robert E. Kreiger New York. 479pp.

Harvey, Lewis O. Jr, Kenneth R. Hammond, Cynthia M. Lusk, and Ernest F. Mross 1992. The application of signal detection theory to weather forecasting behavior. *Monthly Weather Review, 120*, 863-883.

Levi, Keith 1985. A signal detection framework for the evaluation of probabilistic forecasts. *Organizational behaviour and human decision processes, 36*, 143-166.

Mason, I.B. 1979. On reducing probability forecasts to yes/no forecasts. *Monthly Weather Review, 107*, 207-211.

Mason, I.B. 1989. Dependence of the Critical Success Index on sample climate and threshold probability. *Australian Meteorological Magazine, 37*, 75-81.

Mason, I.B. 1997. Verification of some rain probabilities for Canberra. *Forecast and Warning Verification Workshop, Bureau of Meteorology, Melbourne*, 15-17 September 1997.

Mason, I.B. 1997b. The weather forecast as a statistical decision: an outline of signal detection theory and ROC analysis in assessment of forecast quality. *Forecast and Warning Verification Workshop, Bureau of Meteorology, Melbourne*, 15-17 September 1997.

McCoy, Mary Cairns 1986. Severe-storm-forecast results from the PROFS 1983 forecast experiment. *Bulletin American Meteorological Society*, *67*, 155-164.

Murphy, A. H. 1997. Forecast verification. In Katz, Richard W. and Allan H. Murphy (eds), The Economic Value of Weather and Climate Forecasts. *Cambridge University Press, Cambridge.*

Murphy, A. H. and Winkler, R.L., 1987. A general framework for forecast verification. *Monthly Weather Review, 115*, 1330-1338.

Murphy, A. H. and Winkler, R.L., 1987. A general framework for forecast verification. *Monthly Weather Review, 115*, 1330-1338.

Murphy, A. H. and Winkler, R.L., 1992. Diagnostic verification of probability forecasts. *International Journal of Forecasting, 7*, 435-455.

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

Murphy, A.H. and H. Daan 1985. Forecast evaluation. In *Probability, Statistics and Decision Making in the Atmospheric Sciences*. Allan H. Murphy and Richard W. Kats, eds. Westview Press, Boulder and London.

Swets, John A. 1986. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin, 99*, 100-117.

Thomson, J.C., and G.W. Brier, 1955. The economic utility of weather forecasts. *Monthly Weather Review, 83*, 249-254.

Wilks, Daniel S. 1995. Statistical methods in the atmospheric sciences : an introduction. *Academic Press*, pp 467.

Winkler, R.L. and A.H. Murphy 1968. “Good” probability assessors. *Journal of Applied Meteorology, 7*, 751-758.

Winkler, R.L., and A.H. Murphy, 1985. Decision analysis. In *Probability, Statistics and Decision Making in the Atmospheric Sciences* (Allan H. Murphy and Richard W. Katz, Eds), Westview Press, Boulder, Colorado. 493-524.

Yates, J.F. 1982. External correspondence: decompositions of the mean probability score. *Organizational behaviour and human performance, 30*, 132-156.