

WWRP 2009 - 1

Recommendations for the Verification and
Intercomparison of QPFs and PQPFs from
Operational NWP Models

Revision 2
October 2008

For more information, please contact:

World Meteorological Organization

Research Department

Atmospheric Research and Environment Branch

7 bis, avenue de la Paix – P.O. Box 2300 – CH 1211 Geneva 2 – Switzerland

Tel.: +41 (0) 22 730 83 14 – Fax: +41 (0) 22 730 80 27

E-mail: cpa@wmo.int – Website: http://www.wmo.int/pages/prog/arep/index_en.html

WWRP 2009 - 1 Recommendations for the Verification and Intercomparison of QPFs and PQPFs from Operational NWP Models



World
Meteorological
Organization
Weather • Climate • Water

WMO/TD - No. 1485



© World Meteorological Organization, 2008

The right of publication in print, electronic and any other form and in any language is reserved by WMO. Short extracts from WMO publications may be reproduced without authorization provided that the complete source is clearly indicated. Editorial correspondence and requests to publish, reproduce or translate this publication (articles) in part or in whole should be addressed to:

Chairperson, Publications Board
World Meteorological Organization (WMO)
7 *bis* avenue de la Paix
P.O. Box No. 2300
CH-1211 Geneva 2, Switzerland

Tel.: +41 22 730 8403
Fax.: +41 22 730 8040
E-mail: publications@wmo.int

NOTE

The designations employed in WMO publications and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of WMO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Opinions expressed in WMO publications are those of the authors and do not necessarily reflect those of WMO. The mention of specific companies or products does not imply that they are endorsed or recommended by WMO in preference to others of a similar nature which are not mentioned or advertised.

This document (or report) is not an official publication of WMO and has not been subjected to its standard editorial procedures. The views expressed herein do not necessarily have the endorsement of the Organization.

WORLD METEOROLOGICAL ORGANIZATION

WORLD WEATHER RESEARCH PROGRAMME

WWRP 2009 - 1

RECOMMENDATIONS

for the VERIFICATION AND INTERCOMPARISON of
QPFS and PQPFS from OPERATIONAL NWP MODELS

Revision 2

October 2008

WWRP/WGNE Joint Working Group on Verification



Table of Contents

1.	INTRODUCTION.....	1
2.	VERIFICATION STRATEGY	2
3.	REFERENCE DATA.....	6
4.	VERIFICATION METHODS.....	7
5.	REPORTING GUIDELINES.....	11
6.	SUMMARY OF RECOMMENDATIONS.....	12
	References	14
Annex 1	Scores description	17
Annex 2	Guidelines for computing aggregate statistics.....	24
Annex 3	Confidence intervals for verification scores	26
Annex 4	Examples of graphical verification products	27
Annex 5	Membership of WWRP/WGNE Joint Working Group on Verification (JWGV).....	35

1. INTRODUCTION

The Working Group on Numerical Experimentation (WGNE) began verifying quantitative precipitation forecasts (QPFs) in the mid 1990s. The purpose was to assess the ability of operational numerical weather prediction (NWP) models to accurately predict rainfall, which is a quantity of great interest and importance to both the forecasting and user communities. Many countries have national rain gauge networks that provide observations that can be used to verify the model QPFs. Since rainfall depends strongly on atmospheric motion, moisture content, and physical processes, the quality of a model's rainfall prediction is often used as an indicator of overall model health.

In 1995 the US National Centres for Environmental Prediction (NCEP) and the German Deutscher Wetterdienst (DWD) began verifying QPFs from a number of global and regional operational NWP models against data from their national rain gauge networks. The Australian Bureau of Meteorology Research Centre (BMRC) joined in 1997, followed by the UK Meteorological Office in 2000, Meteo-France in 2001, and the Japan Meteorological Agency (JMA) in 2002. Many climatological regimes are represented, allowing QPFs to be evaluated for a large variety of weather types.

The results of the WGNE QPF verification from 1997 through 2000 were summarized in a paper by Ebert et al. (2003). Focusing on a small number of verification measures, they assessed the relative accuracy of model forecasts of 24h rainfall accumulation in summer versus winter, mid-latitudes versus tropics, and light versus heavy rainfall. Quantitative results were provided for each category. However, Ebert et al. noted that it was not possible to *directly* compare the verification results for the United States, Germany, and Australia due to differences in verification methodology and rainfall observation time.

In order to maximize the usefulness of the precipitation verification (e.g., in a model intercomparison), the approach used should be as similar as possible across all regions. It is probably not feasible to change the rainfall observation times, but it is certainly desirable for participating centres to use a common verification methodology. The WWRP/WGNE Joint Working Group on Verification (JWGV, see Annex 4) was asked in 2003 to provide recommendations on a standard verification methodology to be used for QPFs. In 2004 the request was extended to include the verification of probabilistic QPFs (PQPFs).

The purpose of this document is to recommend a standard methodology for verification and intercomparison of QPFs and PQPFs from NWP models. These recommendations apply to the verification of direct model output precipitation forecasts as well as forecasts which have been subjected to post-processing.

A number of recent reports have summarized the state of the art in precipitation verification and have made specific recommendations for the practice of forecast verification. Wilson (2001) surveyed operational centres in Europe as to their QPF verification practices. He found that most centres used observations of 6h, 12h, and/or 24h rainfall accumulations from synoptic stations to verify model QPFs, and that a wide variety of verification scores were in use.

In a report to WGNE, Bougeault (2002) reviewed "standard" and emerging verification techniques with an emphasis on their application to mesoscale model forecasts. He concluded with several recommendations including (a) user-oriented and model-oriented verification may require different methodologies; (b) care is needed in handling the scale differences between model output and observations; (c) probabilistic methods may be more appropriate than the usual deterministic methods for verifying severe weather elements; (d) standard methodologies should be specified for weather elements from NWP; and (e) verification results should be accompanied by uncertainty measures.

WMO's Standardised Verification System for Long-Range Forecasts (SVS for LRF) was developed to reduce confusion among users of long-range forecasts by adopting a coherent

approach to verification (WMO, 2002). This system spells out the procedures to be used by operational centres for producing and exchanging a defined set of verification scores for forecasts of surface air temperature, precipitation, and sea surface temperature anomalies on time scales of monthly to two years. Both deterministic and probabilistic forecasts are to be verified on three levels: (a) large scale aggregated overall measures of forecast performance in tropics, northern extra-tropics and southern extra-tropics; (b) verification at gridpoints (maps); and (c) gridpoint by gridpoint contingency tables for more extensive verification. Because long-range forecasts have many fewer realizations than weather forecasts, information on the uncertainty of the verification scores is also a requirement.

In 2003 the European Centre for Medium range Weather Forecast (ECMWF) commissioned a review of existing verification practices for local weather forecasts in Member States and Co-operating States, to be reported in the annual "Green Book". The resulting report by Nurmi (2003) recommends a verification methodology for deterministic and probabilistic forecasts of weather elements, giving examples to illustrate the various verification methods. Nurmi specifies a set of minimum requirements (i.e., mandatory verification products) that all Member States should satisfy, as well as an optimum set that includes additional verification information.

With the ever growing availability of PQPF forecasts derived from ensembles and MOS systems, verification of PQPFs has become increasingly important. Reporting on a workshop on ensemble forecasting in the short to medium range, Hamill et al. (2000) suggested that verification efforts should place greater emphasis on sensible weather elements and recommended a suite of metrics to be used.

In 2003 the Japanese Meteorological Agency offered to be a Lead-Centre for the Verification of Global Medium Range Ensemble Prediction Systems. The arrangement, described in WMO (2003), calls for JMA to host a web site to display verification results for selected EPS products, including PQPFs, with each contributing centre sending new verification results every month. The URL for this password-protected site is <http://epsv.kishou.go.jp/EPSProducer/>. The specification for the verification products to be exchanged was updated at a later meeting of the WMO CBS expert team on EPS. The revisions are described in WMO (2006).

Two years later the verification working group at the TIGGE (THORPEX Interactive Grand Global Ensemble) First Planning Meeting endorsed the probabilistic verification suite set out by WMO (2003) for evaluating the quality of ensemble forecasts from single and multi-model EPSs (Richardson et al., 2005).

The recommendations made herein are essentially based on the above reports. We emphasize the need to produce a suite of verification measures to evaluate forecasts, rather than rely on some sort of summary score. Similar to Nurmi (2003), we rate verification measures as *highly recommended* (***), *recommended* (**), or *worth a try* (*).

Section 2 presents the verification strategy. Section 3 discusses the reference data, while verification methods and scores can be found in Section 4. Guidelines for reporting verification results are given in Section 5. Finally, the recommendations given in this document are summarized in Section 6. The appendices provide greater detail on the methods and scores that are considered here.

2. VERIFICATION STRATEGY

(a) Scope

Some of the most important reasons to verify forecasts are (a) to monitor forecast quality over time, (b) to compare the quality of different forecast systems, and (c) to improve forecast quality through better understanding of forecast errors. Reasons (a) and (b) may be the main motivators of the WGNE QPF study, but individual NWP centres also have a strong interest in how

to improve their forecasts (c). Different verification methods may be appropriate for each motivation. For example, monitoring the quality of precipitation forecasts usually entails plotting the time series of a small number of well-known scores such as RMSE, frequency bias, and equitable threat score. To evaluate multiple aspects of a forecast system and to intercompare models, a more detailed verification analysis is needed that employs a comprehensive set of verification scores. To really understand the nature of forecast errors so that targeted improvements can be made to a model, diagnostic verification methods are often used. These methods may include distributions-oriented approaches such as scatter plots and histograms, and some of the newer methods such as scale separation and object-oriented methods (see Bougeault (2002) and JWGV (2008)).

The choice of verification methodology depends not only on the purpose of the verification but also on the nature of the forecast being verified. The choice of methods does not, however, depend on the source of the forecast information. For example many centres post-process model forecasts using MOS or other statistical methods to generate forecast products. Ensemble systems do not produce QPF forecasts directly; their generation is an interpretation of the ensemble output. The verification measures discussed below can be applied to continuous, categorical or probabilistic precipitation forecasts whatever their source.

We recommend the use of a prescribed set of verification scores to evaluate and intercompare QPFs and PQPFs from NWP models (details given in Section 4). The use of additional scores and diagnostic methods to clarify the nature of forecast errors is highly desirable.

(b) Spatial matching

The primary interest of the NWP modelling community is anticipated to be model-oriented verification. Model-oriented verification, as we define it, includes processing of the observation data to match the spatial and temporal scales of the observations to those scales resolvable by the model. It addresses the question of whether the models are producing the best possible forecasts given their constraints on spatial and temporal resolution. Cherubini et al. (2002) showed that gridded, "upscaled", observations representing rainfall averaged over a gridbox are more appropriate than synoptic "point" observations for verification of models which produce areal quantities as opposed to grid point values. The upscaling leads to a better match between the model and observation spatial scales. The gridding of observations is discussed in Section 3. Acadia et al. (2003) found that a simple nearest-neighbour averaging method was less likely than bilinear interpolation to artificially smooth the precipitation field when coarser model output is downscaled to a higher resolution grid. To intercompare model precipitation forecasts on a common spatial scale, it is necessary that the forecasts be mapped onto a standard grid. Recent research has demonstrated methods to keep track of scales represented by both forecasts and observations in the verification process. (e.g., Casati and Wilson 2007) or verify forecasts on a variety of spatial scales (e.g., Ebert 2008).

Users of forecasts typically wish to know their accuracy for particular locations. They are also likely to be interested in a more absolute form of verification, without limiting the assessment to those space and time scales resolvable by the model. This is especially relevant now that some models are run at very high resolution, and direct model output is becoming increasingly available to the public via the internet. For this user-oriented verification it is appropriate to use the station observations to verify model output from the nearest gridpoint (or spatially interpolated if the model resolution is very coarse compared to the observations). Verification against a set of station observations that have been quality-controlled using model-independent methods, is the best way of ensuring truly comparable results between models.

Both approaches have certain advantages and disadvantages with respect to the validity of the forecast verification for their respective targeted user groups. The use of gridded observations addresses the scale mismatch and also avoids some of the statistical bias that can occur when stations are distributed unevenly within a network. A disadvantage is that the gridded data are not

"true" observations; that is, they contain some error associated with smoothing and insufficient sampling. Station data are true observations, unadulterated by any post-processing, but they usually contain information on finer scales than can be reproduced by the model, and they under-sample the spatial distribution of precipitation. Members of the JWGV strongly agree that both approaches give important information on forecast accuracy for their respective user groups.¹

We recommend that verification be done both against

- (a) Gridded observations (model-oriented verification) on a common 0.5° latitude/longitude grid (1.0 degree for ensembles)**
- (b) Station observations (user-oriented verification).**
- (c) Time scales**

The WGNE QPF verification/intercomparison has thus far used forecasts of 24h rainfall accumulation as the basic quantity to be verified. This approach is based on the large number of 24h rainfall observations available from national rain gauge networks. 24h observations are less prone to observational errors than those for shorter periods. (Radar and satellite can provide rainfall information with much higher spatial and temporal resolution -- the use of these remotely sensed rain estimates as reference data will be discussed Section 3.)

Some users of precipitation forecasts are interested in knowing their quality on longer time scales, for example, synoptic storm total precipitation or total precipitation over several days. Verification of this nature is easily accomplished by summing observations over successive 24h periods and matching with the corresponding forecasts of multi-day total precipitation.

For model intercomparison it is important to include only those days or time periods that are common to all models being compared. Alternatively, those models with patchy or non-existent output can be excluded from the verification.

We recommend that WGNE continue to use 24h accumulation as the primary temporal scale for the rainfall verification. Additional verification at higher temporal resolution (6h or 12h) is highly desirable, especially for high resolution models, but optional. Verification of storm total or multi-day total precipitation is encouraged to support specific users, but is also optional. Only those days common to all models should be included in the intercomparison.

- (d) Stratification of data**

Stratifying the samples into quasi-homogeneous subsets helps to tease out forecast behaviour in particular regimes. For example, it is well known that forecast performance varies seasonally and regionally. Some pooling, or aggregation, of the data is necessary to get sample sizes large enough to provide robust statistics, but care must be taken to avoid masking variations in forecast performance when the data are not homogeneous. Many scores can be artificially inflated if they are actually reflecting the ability of the model to distinguish seasonal or regional trends instead of the ability to forecast day to day or local weather (Atger, 2003; Hamill and Juras, 2006). Pooling may bias the results toward the most commonly sampled regime (for example, regions with higher station density, or days with no severe weather). Care must be taken when computing aggregate verification scores. Some guidelines are given in Annex 1.

Many stratifications are possible. The most common stratification variables reported in the literature appear to be lead time, season, geographical region, and intensity of the observations.

¹Tustison et al. (2001) advocate using a composite scale matching approach that combines interpolation and upscaling to reduce representativeness error. However, this approach has not been tested for scores other than the RMSE, and is much less transparent than verification using either gridded or raw observations.

We recommend that verification data and results be stratified by:

- (a) Lead time (24h, 48h, etc.)
- (b) Season (winter, spring, summer, autumn, as defined by the 3-month periods, DJF, MAM, JJF, SON)
- (c) Region (tropics, northern extra-tropics, southern extra-tropics, where tropics are bounded by 20° latitude following WMO (2002, 2003), and appropriate mesoscale subregions)
- (d) Observed rainfall intensity threshold (1, 2, 5, 10, 20, 50 mm d⁻¹).

Use of other stratifications relevant to individual countries (altitude, coastal or inland, etc.) is strongly encouraged. Stratification of data and results by forecast rainfall intensity threshold (1, 2, 5, 10, 20, 50 mm d⁻¹) is strongly encouraged.

(e) Reference forecasts

To put the verification results into perspective and show the usefulness of the forecast system, "unskilled" forecasts such as persistence and climatology should be included in the comparison. *Persistence* refers to the most recently observed weather (i.e., the previous day's rainfall in the case of 24h accumulation), while *climatology* refers to the expected weather (for example, the median of the climatological daily rainfall distribution for the given month)², or the climatological frequency of the event being predicted. The verification results for unskilled forecasts hint at whether the weather forecast was "easy" or "difficult". Skill scores measure the relative improvement of the forecast compared to the unskilled forecast (see Section 4). Many of the commonly used verification scores give the skill with respect to *random chance*, which is an absolute and universal reference, but in reality random chance is not a commonly used forecast (despite what our critics say!).

Whether to use persistence or climatology as the standard of comparison depends on the application and on the availability of relevant data. Comparison with persistence gives a good idea of how well the model predicts changes in the weather, while comparison with climatology gives an indication of how well the forecast performs in unusual situations. Persistence forecasts may be hard to beat at shorter forecast ranges, while climatology is usually a stronger competitor at longer forecast projections. Persistence is not a meaningful standard of comparison for probability forecasts, since it is defined deterministically ("yesterday's rainfall").

We recommend that the verification of climatology and/or persistence forecasts be reported along with the forecast verification. The use of skill scores with respect to persistence, climatology, and random chance is highly desirable.

(f) Uncertainty of results

When aggregating and stratifying the data, the subsets should contain enough cases to give reliable verification results. This may not always be possible for rare events. In either case it is necessary to provide quantitative estimates of the uncertainty of the verification results themselves. This allows us to judge whether differences in model performance are likely to be real or just an artifact of sampling variability. Confidence intervals contain more information about the uncertainty of a score than a simple significance test, and can be fairly easily computed using parametric or resampling (e.g., bootstrapping) methods (see Annex 2). The median and interquartile range (middle 50% of the sample distribution reported as the 25th and 75th percentiles) give the "typical" values for the score.

²The long term climatology is preferred over the sample climatology because it is more stable. However, it requires information from outside the verification sample, which may not be easily available. Verification results for climatology forecasts should indicate whether they refer to the long term or the sample climatology.

We recommend that all aggregate verification scores be accompanied by 95% confidence intervals. Reporting of the median and interquartile range for each score is highly desirable.

3. REFERENCE DATA

(a) Observations from rain gauges

Most reference data for verifying model QPFs and PQPFs comes from national rain gauge networks. It is important that these data be as free as possible from error. We recognize that quality control of rainfall observations is extremely difficult due to their highly variable nature. Quality control measures should include screening for unphysical or unreasonable values using buddy checking and/or auxiliary information. Sometimes quality control checks are performed by comparing the observations with a short range forecast from a model. Model-based quality control methods act as selective filters, biasing the verification sample by removing those observations which are significantly different from the model estimate, even if they are accurate. This biasing will have the effect of inflating verification scores, especially for the model used in the quality control. Since precipitation is not usually an initialized variable in NWP, this practice may be less common for precipitation than for other variables, but in any case it is not recommended for verification purposes.

The verification should strive to use all available observations. Some countries have cooperative networks of rainfall observers who report to their meteorological centres well outside of real time. Municipal and other agencies may also make rain measurements. These additional sources can greatly increase the number of the verification data as compared to datasets from synoptic sites only, particularly in high impact areas such as urban centres. A disadvantage of including non-synoptic data is that the verification cannot be done until well after the forecasts are made.

Going from point observations to a gridded analysis is most easily done by averaging all of the observations within each grid box (Cherubini et al., 2002). This method matches observations with gridpoint values locally and independently. An alternate approach is to first use an objective analysis scheme such as kriging to analyze the data onto a fine grid, then upscale by averaging onto a coarser resolution grid. This approach has the advantage of regularizing the data distribution prior to averaging, but makes the assumption that the observations are spatially related on a restricted set of spatial scales.

Efforts should be made to estimate the error associated with the gridded rainfall values (using withdrawal methods, for example). If the magnitude of the analysis errors approaches that of the forecast errors, those grid boxes should be withdrawn from the verification sample. The issue of how to best make use of information on reference data error in the QPF verification is a topic of current research.

We recommend that quality-controlled point and gridded observations from rain gauge networks be the primary reference data for verifying model QPFs and PQPFs. Research on the best use of information about the uncertainty in reference data in the verification process is strongly encouraged. Quality control methods and analysis methods should not involve the model whose forecasts are being verified.

(b) Remotely sensed rainfall estimates

Reference data are also available from remotely sensed observations. Radar data provide rain estimates with very high spatial and temporal resolution (on the order of 1 km and 10 minutes). Many quality control procedures are needed to make radar data usable, and even then the rain estimates may not be accurate if fixed Z-R relationships are applied. Methods to correct local biases using coincident gauge observations have been developed, and high quality combined

radar-gauge rainfall analyses are now available in the US, UK, and Japan, and are being developed in many other countries. They are especially useful for verifying mesoscale rain forecasts where more precise information on timing and location are desired.

Rainfall estimates derived from satellite measurements are of greatest value where gauge and radar observations are not available, in remote regions and over the oceans. Because the measured passive infrared or microwave radiances are only indirectly related to rainfall, satellite rainfall estimates are less accurate than those from radar and gauges. Their use should be focused on giving the location and timing of rainfall, rather than the rainfall amount. High resolution satellite data can also be used as a check on station data in situations where “false alarm” observation errors can occur, on the assumption that no cloud means no precipitation (e.g., Ebert and Weymouth 1999).

We recommend that, where possible, combined radar-gauge rainfall analyses be used to verify model QPFs and PQPFs at high spatial and temporal resolution. Research on the use of satellite estimates for verifying the forecast rainfall location and timing in remote regions should be encouraged. Satellite or radar data that have been processed using a trial field from the model being verified should not be used in verification.

4. VERIFICATION METHODS

Verification begins with a matched set of forecasts and observations for identical locations and times, traditionally treated as samples of independent events. In reality there are often strong spatial and temporal correlations within subsets of the data (coherent spatial patterns, for example), which some of the advanced diagnostic methods explicitly take into account (JWGV, 2008).

Deterministic forecasts can be verified as *categorical events* or *continuous variables*, with different verification scores appropriate for each view. QPFs are usually viewed categorically according to whether or not the rain exceeds a given threshold. The continuous variable of interest is rain amount. Because rainfall amount is not normally distributed and can have very large values, the continuous verification scores (especially those involving squared errors) which are sensitive to large errors may give less meaningful information for precipitation verification than categorical verification scores.

Probabilistic QPFs are usually obtained either by statistical post-processing of deterministic model output (MOS or perfect prog methods), or by calculation from the precipitation probability distribution function (PDF) estimated from ensemble forecasts. Statistical post-processing methods lead to estimates of probabilities of specific events (>2mm/24h, for example) using a training sample of selected variables from previous runs of the model in question, and conditional on the values of the selected variables.

For ensemble output, verification methods are available both for the full estimated distribution and for probabilities calculated from the distribution by setting thresholds³. In the simplest approach, ensemble-based probabilities are defined by the proportion of ensemble members which predict the event to occur. This results in a limited set of probabilities, 0, 1/M, 2/M..., (M-1)/M, 1, where M is the number of members in the ensemble. More advanced forms of processing of ensembles include removal of the bias in the first moment of the ensemble distribution (e.g., Cui et al. 2008), and the fitting of distributions to the ensemble (e.g., Peel and Wilson 2008a). Such methods often result in probability forecasts that may take on any value in the (0,1) range.

³ Methods are available for verifying other aspects of ensemble forecasts, such as their spread, but these are beyond the scope of this document.

A large variety of verification scores are used operationally to verify QPFs and PQPFs (Wilson 2001; Nurmi 2003). Details of these scores can be found in the textbooks of Wilks (2005) and Jolliffe and Stephenson (2003), or on the JWGV (2008) web site and references therein. Most readers of this document will already be familiar with most or all of the scores given in this section. Here we list the scores and indicate whether we consider it to be *highly recommended* (***), *recommended* (**) or *worth a try* (*). This distinction attempts to balance the need to evaluate the important aspects of the forecast while recognizing that most users of verification output may not want to wade through a large array of scores. In Annex 1 we give a very short definition of each score.

Note that once the forecast and observed data have been extracted and the code written to compute the *highly recommended* scores, very little extra effort is required to compute the other scores.

(a) Forecasts of rain meeting or exceeding specified thresholds

For binary (yes/no) events, an event ("yes") is defined by rainfall greater than or equal to the specified threshold; otherwise it is a non-event ("no"). The joint distribution of observed and forecast events and non-events is shown by the categorical contingency table, as represented in Table 1.

Table 1. Categorical contingency table

		Observed		
		Yes	no	
Forecast	yes	<i>Hits</i>	<i>false alarms</i>	<i>forecast yes</i>
	no	<i>Misses</i>	<i>correct rejections</i>	<i>forecast no</i>
		<i>observed yes</i>	<i>observed no</i>	<i>N = total</i>

The elements of the table, *hits*, *false alarms*, *misses*, and *correct rejections*, count the number of times each forecast and observed yes/no combination occurred in the verification dataset. A large number of verification scores are computed from these four values. **Reporting the number of *hits*, *false alarms*, *misses*, and *correct rejections* for each of the rain thresholds specified in Section 2 is mandatory.**

The list of recommended scores includes:

- Frequency bias (***)
- Proportion correct (PC) (***)
- Probability of detection (POD) (***)
- False alarm ratio (FAR) (***)
- Probability of false detection (POFD) (**)
- Threat score (TS) (**)
- Equitable threat score (ETS) (***)
- Hanssen and Kuipers score (HK) (**)
- Heidke skill score (HSS) (**)
- Odds ratio (OR) (**)
- Odds ratio skill score (ORSS) (**).

(b) Forecasts of rain amount

Other statistics measure the quality of forecasts of a continuous variable such as rain amount. As discussed previously, some continuous verification scores are sensitive to outliers. One strategy for lessening their impact is to normalize the rain amounts using a square root transformation (Stephenson et al., 1999). The verification quantities are computed from the square root of the forecast and observed rain amounts, then inverse transformed by squaring, if necessary, to return to the appropriate units. As the resulting errors are smaller than those computed from unnormalized data it is necessary to indicate whether the errors or scores apply to normalized or unnormalized data.

The suggested scores are listed below, while their brief descriptions can be found in Annex 1:

- Mean observed (***)
- Sample standard deviation (s) (***)
- Conditional median (***)
- Interquartile range (IQR) (**)
- Mean error (ME) (***)
- Mean absolute error (**)
- Mean square error (MSE) (**)
- Root mean square error (RMSE) (***)
- Root mean square factor (RMSF) (**)
- (Product moment) correlation coefficient (r) (***)
- Spearman rank correlation coefficient (r_s) (**)
- MAE skill score (**)
- MSE skill score (**)
- Linear error in probability space (LEPS) (**).

(c) Probability forecasts of rain meeting or exceeding specified thresholds

The vast majority of PQPFs are expressed as the probability of occurrence of a specific pre-defined event. For the forecast to be verifiable, the event must be completely specified in terms of valid time interval (e.g., 24h precipitation accumulation > 1.0 mm) and location. Probabilities valid over a grid box are valid for every point in the gridbox, with all points assigned the same probability. Probabilities generally may take any value from 0 to 1 inclusive.

The suggested scores are listed below and are briefly described in Annex 1:

- Brier score (BS) (**)
- Brier skill score (BSS) (***)
- Reliability diagram (***)
- Relative operating characteristic (ROC) (***)
- ROC area (ROCA) (***)

(d) Verification of the ensemble PDF

There is interest, especially in the ensemble modelling community, to verify the full distribution as interpreted from the ensemble members, either as a PDF, or in its integrated cumulative (CDF) form. Several verification measures have been developed for this purpose.

The following scores are suggested. A brief description can be found in Annex 1.

- Ranked probability score (RPS) (**)
- Ranked probability skill score (RPSS) (**)

- Continuous ranked probability score (CRPS) (***)
- Continuous ranked probability skill score (CRPSS) (***) whenever comparisons are made)
- Ignorance score (**).

(e) Simple diagnostic methods

Diagnostic methods give more in-depth information about the performance of a forecast system. Some methods examine the joint distribution of independent forecast and observed values, while others verify the spatial distribution or intensity distribution. Most diagnostic methods produce graphical results.

Maps of observed and forecast rainfall show whether the overall spatial patterns are well represented by the forecast system. **Maps of seasonal mean rainfall are highly recommended. Maps of the frequency of rainfall exceeding certain thresholds (for example, 1 mm d⁻¹ and 10 mm d⁻¹) are recommended.**

Time series of observed and forecast domain mean rainfall allow us to see how well the temporal patterns are simulated by the model. **Time series of seasonal mean rainfall are highly recommended. Time series of mean rainfall for shorter time series are recommended. Time series of the seasonal frequency of rainfall exceeding certain thresholds (for example, 1 mm d⁻¹ and 10 mm d⁻¹) are recommended.**

A *scatter plot* simply plots the forecast values against the observed values to show their correspondence. The results can be plotted as individual points, or if there are a very large number, as a contour plot. **Scatter plots of forecast versus observed rain are highly recommended. Scatter plots of forecast error versus observed rainfall are recommended.**

The distribution of forecast and observed rain amounts can be compared using *histograms* for discrete rainfall bins, *quantile-quantile* plots for discrete percentiles of the forecast and observed (wet) distributions, or by plotting the *exceedance probability* (1-CDF) as a function of rain amount. **These plots are recommended.**

(f) Advanced diagnostic methods (*)

Several advanced diagnostic methods have proven very useful for assessment of deterministic models in a research setting, and we encourage scientists to begin to experiment with them in research and operations. Some examples include multi-scale spatial statistics (e.g., Harris et al., 2001), scale decomposition methods (e.g., Casati et al., 2004), object oriented methods (Ebert and McBride, 2000; Davis et al., 2006), neighbourhood (fuzzy) verification methods (Ebert 2008), and field verification methods (e.g., Keil and Craig 2007). More information on these and other methods can be found on the JWG (2008) web site.

The *rank histogram* is a commonly used diagnostic tool for assessment of the spread of an ensemble PDF, averaged over a verification sample. Although it is widely used for continuous unbounded variables, it is more difficult to apply and interpret for bounded variables such as QPF, and is less frequently used for such variables. It is also not often used to quantitatively compare ensembles, even though quantitative methods have been reported in the literature (Candille and Talagrand 2005). Hamill and Colucci (1997) describe a method to compute the rank histogram for variables such as precipitation where several ensemble members may forecast the same value. We recommend that the rank histogram be explored further for precipitation ensemble diagnostic purposes, with the following caveats. First, it is a diagnostic tool only and not a true verification method since it is possible to obtain a perfect result by randomly selecting the verification category from climatology without reference to the forecast. Unlike most of the above verification measures, the rank histogram does not involve a systematic match of each forecast with observation, but rather has meaning only when computed over a sufficiently large verification sample. Finally, precipitation amounts over specified periods tend to follow a gamma-type distribution, which

means that the spread is related to the mean. Thus, the rank histogram results must be interpreted in combination with the bias of the ensemble mean, since it is possible to obtain higher average spread simply by overforecasting the precipitation.

5. REPORTING GUIDELINES

The QPF verification will be most useful to users if the results are available in a timely fashion via the internet. A QPF report will also normally be made to WGNE on an annual basis.

(a) Information about verification system

Information must be provided on the data and methodology used in the verification. This should include:

Reference data:

- Description of reference data type(s)
- Locations of rain gauges and other reference data
- Domain for verification
- Reporting time(s) for reference data
- Description of quality control measures used
- Description of method used to put point observations onto a grid
- Information on observation and analysis errors, if available
- Description of climatology reference forecasts.

Quantitative precipitation forecasts:

- Models included in the verification
- For each model:
 - Name and origin
 - Initialization time(s)
 - Spatial resolution of grid
 - Other information (spectral vs. gridpoint, vertical levels, cloud/precipitation physics, etc.) is useful but not required
- Method used to remap to the common grid.

Verification methods:

- List of scores, plots, and other type of verification products used in the system
- Method used to estimate confidence intervals on scores.

(b) Information about verification products

Every verification product must be accompanied by information on exactly what is being verified:

- Score being computed / diagnostic technique being used
- Transformation applied to rain amounts, if any
- Model(s) being verified
- Country of verification
- Region within country (if applicable)
- Season(s) and year(s)
- Initialization time(s)
- Lead time(s)
- Accumulation period(s)
- Spatial scale(s)
- Rain thresholds(s).

(c) Display of verification results

Graphical products are generally easier to digest than tables full of numbers. There are many ways to "slice and dice" the results to show different aspects of the QPF evaluation. Some suggestions for graphical products are:

- Plot of scores for multiple models/seasons/rain thresholds as a function of lead time
- Plot of scores for multiple models/lead times/rain thresholds as a function of season (time series)
- Plot of scores for multiple models/seasons/lead times as a function of rain threshold
- *POD* versus *FAR* for multiple models as a function of rain threshold/lead times
- Bar chart of scores as a function of model, lead time, season, rain threshold, etc.
- Box and whiskers plot of daily scores as a function of model, lead time, season, rain threshold, etc.
- Taylor diagram (Taylor, 2001) including multiple models, lead times, seasons, thresholds, etc.

Some examples of graphical verification products are shown in Annex 4.

In addition to the graphical products, the numerical verification results should be made available to other WGNE scientists. All scores should be tabulated in downloadable files for ease of use.

Confidence intervals may be shown as error bars on the diagrams. At the very least they should be reported in the tables.

6. SUMMARY OF RECOMMENDATIONS

The mandatory set of verification scores should be used to evaluate and intercompare QPFs from NWP models (details given in Section 4). The additional use of optional measures and diagnostic methods to clarify the nature of model errors is highly desirable.

The verification should be done both against

- (a) Gridded observations (model-oriented verification) on a common 0.5° latitude/longitude grid (1.0° for ensemble forecasts)
- (b) Station observations (user-oriented verification).

24h accumulation should continue to be used as the primary temporal scale for the rainfall verification. Additional verification at higher temporal resolution (6h or 12h) is highly desirable especially for high resolution models. Verification of storm total or multi-day total precipitation is encouraged to support specific users. Only those days common to all models should be included in the intercomparison.

The verification data and results should be stratified by:

- (a) Lead time (24h, 48h, etc.)
- (b) Season (winter, spring, summer, autumn, as defined by 3-month periods, DJF, MAM, JJF, SON)
- (c) Region (tropics, northern extra-tropics, southern extra-tropics, where tropics are bounded by 20° latitude, and appropriate mesoscale subregions)
- (d) Observed rainfall intensity threshold (1, 2, 5, 10, 20, 50 mm d⁻¹).

Use of other stratifications relevant to individual countries (altitude, coastal or inland, etc.) is strongly encouraged. Stratification of data and results by forecast rainfall intensity threshold (1, 2, 5, 10, 20, 50 mm d⁻¹) is strongly encouraged.

References

- Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbour average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932.
- Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlin. Proc. Geophys.*, **8**, 401-417.
- Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.*, **131**, 1509-1523.
- Bougeault, P., 2002: WGNE survey of verification methods for numerical prediction of weather elements and severe weather events. *CAS/JSC WGNE Report No. 18, Annex C*. Available at <http://www.wmo.ch/web/wcrp/documents/wgne18rpt.pdf>.
- Candille, G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Met. Soc.*, **131**, 231-250.
- Casati, B., G. Ross and D.B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Met. Applications*, **11**, 141-154.
- Casati, B. and L. J. Wilson, 2007: A new spatial-scale decomposition of the Brier Score: Application to the verification of lightning probability forecasts. *Mon. Wea. Rev.* **135**, 3052–3069.
- Cherubini, T., A. Ghelli, and F. Lalaurette, 2002: Verification of precipitation forecasts over the Alpine region using a high-density observing network. *Wea. Forecasting*, **17**, 238-249.
- Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2008: Statistical downscaling approach and its application. Preprints, *19th Conf. Probability and Statistics in Atmospheric Sciences, Amer. Met. Soc., New Orleans, Jan, 2008*
- Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methods and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772-1784.
- Ebert, E.E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteorol. Appls.* **15**, 51-64.
- Ebert, E.E., U. Damrath, W. Wergen and M.E. Baldwin, 2003: The WGNE assessment of short-term quantitative precipitation forecasts. *Bull. Amer. Met. Soc.*, **84**, 481-492.
- Ebert, E.E. and J.L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrology*, **239**, 179-202.
- Ebert, E.E. and G.T. Weymouth, 1999: Incorporating satellite observations of “no rain” in an Australian daily rainfall analysis. *J. Appl. Climatol.*, **38**, 44-56.
- Golding, B.W., 1998: Nimrod: A system for generating automated very short range forecasts. *Meteorol. Appls.*, **5**, 1-16.
- Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327.
- Hamill, T.M., S.L. Mullen, C. Snyder, Z. Toth and D.P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Met. Soc.*, **81**, 2653-2664.

- Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Q J Roy. Met Soc.*, **132**, 2905-2923.
- Harris, D., E. Foufoula-Georgiou, K.K. Droegemeier and J.J. Levit, 2001: Multiscale statistical properties of a high-resolution precipitation forecast. *J. Hydrometeorology*, **2**, 406-418.
- Jolliffe, I.T., 2007: Uncertainty and inference for verification measures. *Wea. Forecasting*, **22**, 637-650.
- Jolliffe, I.T., and D.B. Stephenson, 2003: *Forecast Verification. A Practitioner's Guide in Atmospheric Science*. Wiley and Sons Ltd, 240 pp.
- JWGV, 2008: Forecast verification – Issues, methods, and FAQ.
http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html
- Keil, C. and G.C. Craig, 2007: A Displacement-Based Error Measure Applied in a Regional Ensemble Forecasting System. *Mon. Wea. Rev.*, **135**, 3248-3259.
- Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291-303.
- Mason, S. J., 2007: Do high skill scores mean good forecasts? Presentation at the *Third International Verification Methods Workshop*, ECMWF, Reading, UK, Jan 2007. Available at http://ecmwf.int/newsevents/meetings/workshops/2007/jwqv/workshop_presentations/S_Mason.pdf
- Mason, S. J. and N. E. Graham, 2002: Areas beneath the relative operating characteristic (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Q J Roy. Met. Soc.*, **128**, 2145-2166.
- Murphy, A.H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595-600.
- Nurmi, P., 2003: Recommendations on the verification of local weather forecasts. *ECMWF Tech. Memo. 430*, 18 pp. Available at <http://www.ecmwf.int/publications/library/ecpublications/pdf/tm430.pdf>
- Peel, S., and L. J. Wilson, 2008a: Modelling the distribution of precipitation forecasts from the Canadian ensemble prediction system using kernel density estimation. *Wea. Forecasting*, **23**, 596-616.
- Peel, S., and L. J. Wilson, 2008b: A diagnostic verification of the precipitation forecasts produced by the Canadian ensemble prediction system. *Wea. Forecasting*, **23**, 575-595.
- Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Royal Met. Soc.*, **126**, 649-667.
- Richardson, D., R. Buizza, and R. Hagedorn, 2005: Report of the 1st Workshop on the THORPEX Interactive Grand Global Ensemble (TIGGE), 1-3 March 2005, ECMWF. 34 pp.
- Roulston, M. S. & L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory, *Monthly Weather Review*, **130**, 1653-1660.
- Schaefer, J.T., 1990: The critical success index as an indicator of forecasting skill. *Wea. Forecasting*, **5**, 570-575.
- Stanski, H.R., L.J. Wilson, and W.R. Burrows, 1989: *Survey of common verification methods in meteorology*. World Weather Watch Tech. Rept. No.8, WMO/TD No.358, WMO, Geneva, 114 pp.

- Stephenson, D.B., R. Rupa Kumar, F.J. Doblas-Reyes, J.-F. Royer, F. Chauvin, and S. Pezulli, 1999: Extreme daily rainfall events and their impact on ensemble forecasts of the Indian monsoon. *Mon. Wea. Rev.*, **127**, 1954-1966.
- Taylor, K.E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183-7192.
- Tustison, B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res.*, **106** (D11), 11,775-11,784.
- Wilks, D.S., 2005: *Statistical Methods in the Atmospheric Sciences. An Introduction*. 2nd ed., Academic Press, San Diego, 467 pp.
- Wilson, C., 2001: Review of current methods and tools for verification of numerical forecasts of precipitation. COST717 Working Group Report on Approaches to Verification. Available on the internet at http://pub.smhi.se/cost717/doc/WDF_02_200109_1.pdf.
- Wilson, L., 2000: Comments on "Probabilistic predictions of precipitation using the ECMWF ensemble prediction system". *Wea. Forecasting*, **15**, 361-364.
- WMO, 2002: Standardised Verification System (SVS) for Long-Range Forecasts (LRF). New attachment II-9 to the *Manual on the GPDS* (WMO-No.485), Volume 1. Available on the internet at <http://www.wmo.ch/web/www/DPS/LRF-standardised-verif-sys-2002.doc>
- WMO, 2003: Commission for Basic Systems, OPAG on Data Processing and Forecasting Systems, Meeting of the Expert Team on Ensemble Prediction Systems, Geneva, 27-31 October 2003, Final Report. 33 pp. Available at http://www.wmo.ch/web/www/DPS/Reports/ET-EPS_Geneva2003.pdf.
- WMO, 2006: Commission for Basic Systems, OPAG on Data Processing and Forecasting Systems, Meeting of the Expert Team on Ensemble Prediction Systems, Exeter, 6-10 February, 2006. 38 pp. Available at http://www.wmo.int/pages/prog/www/DPFS/Reports/ET-EPS_Exeter2006.pdf.

SCORES DESCRIPTION

a) Forecast of rain meeting or exceeding specified thresholds

The following scores are based on a categorical contingency table whereby an event ("yes") is defined by rainfall greater than or equal to the specified threshold; otherwise it is a non-event ("no"). The joint distribution of observed and forecast events and non-events is shown by the categorical contingency table, as represented in Table 2.

Table 2. Categorical contingency table

		Observed		
		Yes	no	
Forecast	yes	Hits	false alarms	forecast yes
	no	misses	correct rejections	forecast no
		observed yes	observed no	N = total

The *frequency bias* gives the ratio of the forecast rain frequency to the observed rain frequency.

$$BIAS = \frac{\text{hits} + \text{false alarms}}{\text{hits} + \text{misses}}$$

The *proportion correct (PC)* gives the fraction of all forecasts that were correct.

$$PC = \frac{\text{hits} + \text{correct rejections}}{N}$$

The *probability of detection (POD)*, also known as the hit rate, measures the fraction of observed events that were correctly forecast.

$$POD = \frac{\text{hits}}{\text{hits} + \text{misses}}$$

The *false alarm ratio (FAR)* gives the fraction of forecast events that were observed to be non-events.

$$FAR = \frac{\text{false alarms}}{\text{hits} + \text{false alarms}}$$

The *probability of false detection (POFD)*, also known as the false alarm rate, measures the fraction of observed non-events that were forecast to be events.

$$POFD = \frac{\text{false alarms}}{\text{correct rejections} + \text{false alarms}}$$

The *threat score (TS)*, also known as the critical success index and hit rate, gives the fraction of all events forecast and/or observed that were correctly diagnosed.

$$TS = \frac{hits}{hits + misses + false\ alarms}$$

The *equitable threat score (ETS)*, also known as the Gilbert skill score, measures the fraction of all events forecast and/or observed that were correctly diagnosed, accounting for the hits that would occur purely due to random chance.

$$ETS = \frac{hits - hits_{random}}{hits + misses + false\ alarms - hits_{random}}$$

where

$$hits_{random} = \frac{1}{N}(\text{observed yes} \times \text{forecast yes})$$

The *Hanssen and Kuipers score (HK)*, also known as the Pierce skill score and the true skill statistic, measures the ability of the forecast system to separate the observed "yes" cases from the "no" cases. It also measures the maximum possible relative economic value attainable by the forecast system, based on a cost-loss model (Richardson, 2000).

$$HK = POD - POFD$$

The *Heidke skill score (HSS)* measures the increase in proportion correct for the forecast system, relative to that of random chance.⁴

$$HSS = \frac{hits + correct\ rejections - correct_{random}}{N - correct_{random}}$$

where

$$correct_{random} = \frac{1}{N}(\text{observed yes} \times \text{forecast yes} + \text{observed no} \times \text{forecast no})$$

The *odds ratio (OR)* gives the ratio of the odds of making a hit to the odds of making a false alarm, and takes prior probability into account.

$$OR = \frac{hits \times correct\ rejections}{misses \times false\ alarms}$$

The odds ratio skill score (ORSS) is a transformation of the odds ratio to have the range [-1,+1].

$$ORSS = \frac{hits \times correct\ rejections - misses \times false\ alarms}{hits \times correct\ rejections + misses \times false\ alarms}$$

b) Forecasts of rain amounts

In the expressions to follow F_i indicates the forecast value for point or grid box i , O_i indicates the observed value, and N is the number of samples

The *mean value* is useful for putting the forecast errors into perspective.

⁴For the two-category case the *HSS* is related to the *ETS* according to $HSS = 2 ETS / (1+ETS)$ (Schaefer 1990).

$$\bar{O} = \frac{1}{N} \sum_{i=1}^N O_i \quad \bar{F} = \frac{1}{N} \sum_{i=1}^N F_i$$

Another descriptive statistic, the *sample variance* (s^2) describes the rainfall variability.

$$s_O^2 = \frac{1}{N-1} \sum_{i=1}^N (O_i - \bar{O})^2 \quad s_F^2 = \frac{1}{N-1} \sum_{i=1}^N (F_i - \bar{F})^2$$

The *sample standard deviation* (s) is equal to the square root of the sample variance, and provides a variability measure in the same units as the quantity being characterized.

$$s_O = \sqrt{s_O^2} \quad s_F = \sqrt{s_F^2}$$

The *conditional median* gives the "typical" rain amount, and is more resistant to outliers than the mean. Since the most common rain amount will normally be zero, the conditional median should be drawn from the wet samples in the distribution.

The *interquartile range (IQR)* is equal to [25th percentile, 75th percentile] of the distribution of rain amounts, and reflects the sample variability. It is more resistant to outliers than the standard deviation. As with this conditional median, the *IQR* should be drawn from the wet samples.

The *mean error (ME)* measures the average difference between the forecast and observed values.

$$ME = \frac{1}{N} \sum_{i=1}^N (F_i - O_i) = \bar{F} - \bar{O}$$

The *mean absolute error (MAE)* measures the average magnitude of the error.

$$MAE = \frac{1}{N} \sum_{i=1}^N |F_i - O_i|$$

The *mean square error (MSE)* measures the average squared error magnitude, and is often used in the construction of skill scores. Larger errors carry more weight.

$$MSE = \frac{1}{N} \sum_{i=1}^N (F_i - O_i)^2$$

The *root mean square error (RMSE)* measures the average error magnitude but gives greater weight to the larger errors.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (F_i - O_i)^2}$$

It is useful to decompose the *RMSE* into components representing differences in the mean and differences in the pattern or variability.

$$RMSE = \sqrt{(\bar{F} - \bar{O})^2 + \frac{1}{N} \sum_{i=1}^N [(F_i - \bar{F}) - (O_i - \bar{O})]^2}$$

The *root mean square factor (RMSF)* is the exponent of the root mean square error of the logarithm of the data, and gives a scale to the multiplicative error, i.e., $F = O \times / \div RMSF$ (Golding

1998). Statistics are only accumulated where the forecast and observations both exceed 0.2 mm, or where either exceeds 1.0 mm; lower values are set to 0.1 mm.

$$RMSF = \exp \left(\sqrt{\frac{1}{N} \sum_{i=1}^N \left[\log \left(\frac{F_i}{O_i} \right) \right]^2} \right)$$

The (*product moment*) *correlation coefficient* (r) measures the degree of linear association between the forecast and observed values, independent of absolute or conditional bias. As this score is highly sensitive to large errors it benefits from the square root transformation of the rain amounts.

$$r = \frac{\sum_{i=1}^N (F_i - \bar{F})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^N (F_i - \bar{F})^2} \sqrt{\sum_{i=1}^N (O_i - \bar{O})^2}} = \frac{s_{FO}}{s_F s_O}$$

The (*Spearman*) *rank correlation coefficient* (r_s) measures the linear monotonic association between the forecast and observations, based on their ranks, R_F and R_O (i.e., the position of the values when arranged in ascending order). r_s is more resistant to outliers than r .

$$r_s = 1 - \frac{6}{N(N^2 - 1)} \sum_{i=1}^N (R_{F_i} - R_{O_i})^2$$

Any of the accuracy measures can be used to construct a *skill score* that measures the fractional improvement of the forecast system over a reference forecast. The most frequently used scores are the *MAE* and the *MSE*. The reference estimate could be either climatology or persistence for 24 h accumulations, but persistence is suggested as a standard for short range forecasts and shorter accumulation periods.

$$MAE_SS = \frac{MAE_{forecast} - MAE_{reference}}{MAE_{perfect} - MAE_{reference}} = 1 - \frac{MAE_{forecast}}{MAE_{reference}}$$

$$MSE_SS = \frac{MSE_{forecast} - MSE_{reference}}{MSE_{perfect} - MSE_{reference}} = 1 - \frac{MSE_{forecast}}{MSE_{reference}}$$

The *linear error in probability space* (LEPS) measures the error in probability space as opposed to measurement space, where $CDF_o()$ is the cumulative probability density function of the observations, determined from an appropriate climatology.

$$LEPS = 3 \left[\frac{1}{N} \sum_{i=1}^N (1 - |CDF_o(F_i) - CDF_o(O_i)| + CDF_o^2(F_i) - CDF_o(F_i) + CDF_o^2(O_i) - CDF_o(O_i)) \right] - 1$$

c) Probability forecasts of rain meeting or exceeding specified thresholds

In this section P_i represents the forecast rain probability for point or grid box i , and O_i is the observed occurrence equal to 0 or 1. These scores are explained in some detail here and in Annex 4, since they may be less familiar to many scientists.

The *Brier score* (BS) (**) measures the mean squared error in probability space.

$$BS = \frac{\sum_{i=1}^N (P_i - O_i)^2}{N}$$

The Brier score can be partitioned into three components, following Murphy (1973)

$$BS = \underbrace{\frac{1}{N} \sum_{j=1}^J n_j \sum_{i=1}^{n_j} (P_{ij} - \bar{O}_j)^2}_{\text{Reliability}} - \underbrace{\frac{1}{N} \sum_{j=1}^J n_j (\bar{O}_j - \bar{O})^2}_{\text{Resolution}} + \underbrace{\bar{O}(1 - \bar{O})}_{\text{Uncertainty}}$$

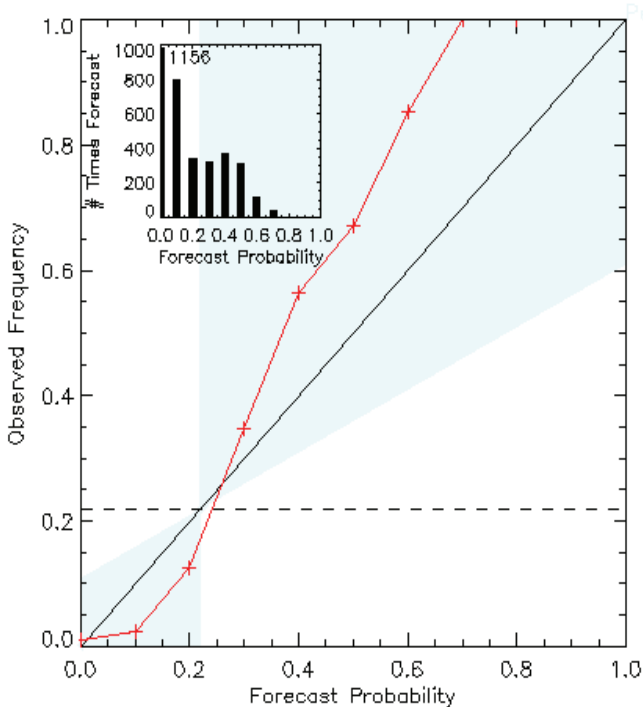
Verification samples of probability forecasts are frequently partitioned into ranges of probability, for example deciles, 0 to 0.1, 0.1 to 0.2 etc. The above form of the partitioned Brier score reflects this binning, where J is the number of bins into which the forecasts have been partitioned, and n_j is the number of cases in the j^{th} bin. The bins should be chosen such that there are sufficient number of cases in each to provide a stable estimate of the frequency of occurrence of the event in each bin, \bar{O}_j . Reliability and resolution can be evaluated quantitatively from this equation, or can be evaluated graphically from the reliability table (see below). The uncertainty term does not depend on the forecast at all; it is a function of the observations only, which means that the Brier scores are not comparable when computed on different samples. The Brier Skill Score with respect to sample climatology avoids this problem.

The Brier skill score (BSS) references the value of the BS for the forecast to that of a reference forecast, usually climatology.

$$BSS = 1 - \frac{BS_f}{BS_c}$$

where the subscripts f and c refer to the forecast and climatological forecast respectively. When sample climatology is used as the reference the decomposition of the BSS takes the simple form,

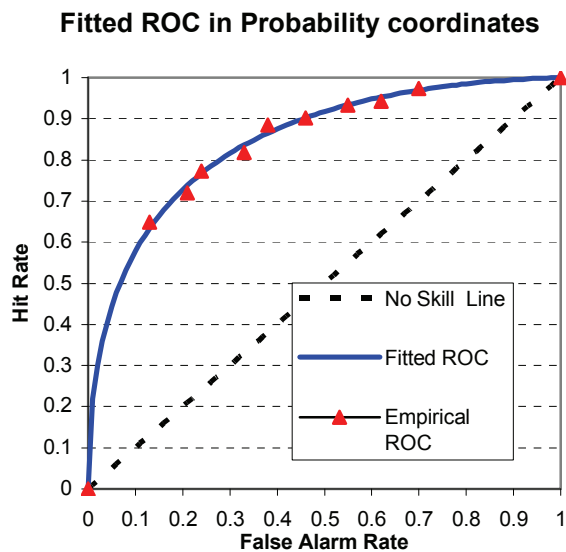
$$BSS = \frac{\text{resolution} - \text{reliability}}{\text{uncertainty}}$$



The *reliability diagram* is used to evaluate bias in the forecast. For each forecast probability category along the x-axis the observed frequency is plotted on the y-axis. The number of times each forecast probability category is used indicates its sharpness. This can be represented on the diagram either by plotting the bin sample size next to the points, or by inserting a histogram. In the example shown, the sharpness is represented by a histogram, the shaded area represents the positive skill region and the horizontal dashed line shows the sample climatological frequency of the event. Annex 4 shows a reliability diagram for probability forecasts of 10 day dry periods.

The *relative operating characteristic (ROC)* is a plot of the hit rate (H , same as POD) against the false alarm rate (F , same as

POFD) for categorical forecasts based on probability thresholds varying between 0 and 1. It measures the ability of the forecast to distinguish (discriminate) between situations followed by the occurrence of the event in question, and those followed by the non-occurrence of the event. The main score associated with the ROC is the *area under the curve* (ROCA)^{***}.

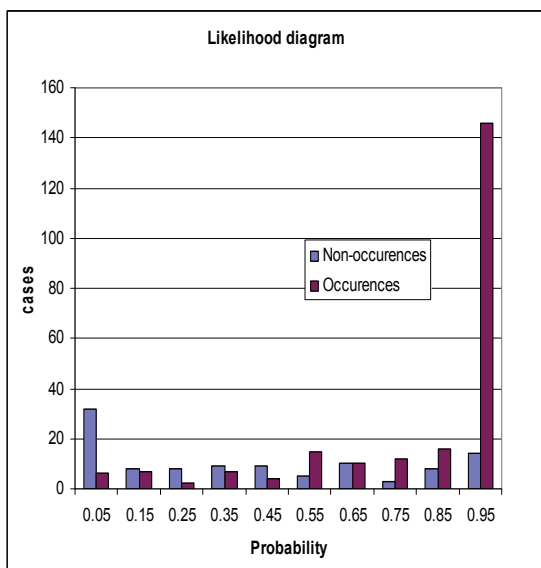


There are two recommended methods to plot the ROC curve and to calculate the ROCA. The method often used in climate forecasting, where samples are usually small, is to list the forecasts in ascending order of the predicted variable (in this case, rain amount), tally the hit rate and false alarm rate for each value considered as a threshold and plot the result (e.g. Mason and Graham 2002). This gives a zigzag line for the ROC, from which the area under the curve can be easily calculated directly as the total area contained within the rectangular boxes beneath the curve. This method has the advantage that no assumptions are made in the computation of the area – the ability of the forecast to discriminate between occurrences and non-occurrences, or

between high and low values of rain amount, can be computed directly. This method could and perhaps should be used to assess the discriminant information in deterministic precipitation forecasts, as suggested by Mason (2007).

The other recommended method, which works best for larger samples typical of short- or medium-range ensemble forecasts. is to bin the dataset according to forecast probability (deciles is most common), tabulate H and F for the thresholds between the bins, and plot the empirical ROC. Then, a smooth curve is fit according to the methodology described in Jolliffe and Stephenson (2003), Mason (1982), and Stanski et al (1989) and the area is obtained using the fitted curve. This method makes the assumption that H and F are transformable to standard normal deviates by means of a monotonic transformation, for example the square root transformation sometimes used to normalize precipitation data. There is considerable empirical evidence that this assumption holds for meteorological data (Mason, 1982). The commonly used method of binning the data into

deciles, then calculating the area using the trapezoidal rule isn't recommended, especially when ROCs are to be compared, for reasons discussed in Wilson (2000). In the example above, the binormal fitting method has been used, and both the fitted curve and the empirical points are shown, which helps evaluate the quality of the fit.

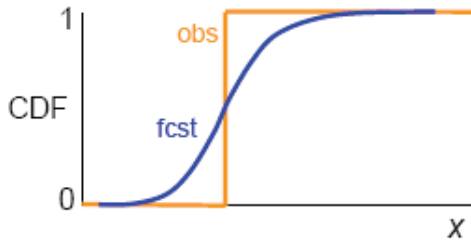


Visual depiction of discriminant ability of the forecast is enhanced by the inclusion of a likelihood diagram with the ROC. This is a plot of the two conditional distributions of forecast probability, given occurrence and non-occurrence of the predicted category. The diagram above corresponds to the forecasts shown in the ROC. These two distributions should be as separate as possible; no overlap at all indicates perfect discrimination.

d) Verification of the ensemble PDF

A continuous ranked probability score (CRPS) is given by

$$CRPS(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 dx$$



The CRPS is probably the best measure for comparing the overall correspondence of the ensemble-based CDF to observations. The score is the integrated difference between the forecast CDF and the observation, represented as a CDF, as illustrated at left. For a deterministic forecast, the score converges to the MAE, thus allowing ensemble forecasts and deterministic forecasts to be compared using this score.

The CRPSS is the skill score based on the CRPS,

$$CRPSS = 1 - \frac{CRPS_{fcst}}{CRPS_{ref}}$$

The ranked probability score (RPS) is an extension of the BS to multiple probability categories, and is a discrete form of the CRPS. It is usually applied to K categories defined by (K-1) fixed physical thresholds,

$$RPS = \frac{1}{K-1} \sum_{k=1}^K (CDF_{fcst,k} - CDF_{obs,k})^2$$

where CDF_k refers to the cumulative distribution evaluated at the k^{th} threshold. It should be noted that in practice the CRPS is evaluated using a set of discrete thresholds as well, but these are usually determined by the values forecast by the ensemble system, and change for each case of the verification sample.

The ranked probability skill score (RPSS) references the RPS to the unskilled reference forecast.

$$RPSS = 1 - \frac{RPS_{fcst}}{RPS_{ref}}$$

The ignorance score (Roulston and Smith, 2002) evaluates the forecast PDF in the vicinity of the observation. It is given by:

$$IGN = -\log_2 P(T = T_o)$$

It can be interpreted as the level of ignorance about the observation inherent in the forecast, and ranges from 0 when a probability of 1.0 is assigned to the verifying value to infinity when a probability of 0 is assigned to the verifying observation. In practice, ensemble systems quite frequently assign 0 probability to observed values which are within the range of possible occurrences. It is thus necessary to take steps to prevent the score from blowing up when the observation is far from the ensemble distribution, for example by setting a suitable small non-zero probability limit.

GUIDELINES FOR COMPUTING AGGREGATE STATISTICS

Real-time verification systems often produce daily verification statistics from the spatial comparisons of forecasts and observations, and store these statistics in files. To get aggregate statistics for a period of many days it is tempting to simply average all of the daily verification statistics. Note that in general this does not give the same statistics as those that would be obtained by pooling the samples over many days. For the linear scores such as mean error, the same result is obtained, but for non-linear scores (for example, anything involving a ratio) the results can be quite different.

For example, imagine a 30-day time series of the frequency bias score, and suppose one day had an extremely high bias of 10 because the forecast predicted an area with rain but almost none was observed. If the forecast rain area was 20% every day and this forecast was exactly correct on all of the other 29 days (i.e., bias=1), the daily mean frequency bias would be 1.30, while the frequency bias computed by pooling all of the days is only 1.03. These two values would lead to quite different conclusions regarding the quality of the forecast.

The verification statistics for pooled samples are preferable to averaged statistics because they are more robust. In most cases they can be computed from the daily statistics if care is taken. The guidelines below describe how to correctly use the daily statistics to obtain aggregate multi-day statistics. An assumption is made that each day contains the same number of samples, N (number of gridpoints or stations).

For pooled categorical scores computed from the 2x2 contingency table (Section 4a):

First create an aggregate contingency table of hits, misses, false alarms, and correct rejections by summing their daily values, then compute the categorical scores as usual.

For linear scores (mean, mean error, MAE, MSE, LEPS):

The average of the daily statistics is the same as the statistics computed from the pooled values.

For non-linear scores:

The key is to transform the score into one for which it is valid to average the daily values. The mean value is then transformed back into the original form of the score.

RMSE: First square the daily values to obtain the *MSE*. Average the squared values, then take the square root of the mean value.

RMSF: Take the logarithm of the daily values and square the result, then average these values. Transform back to *RMSF* by taking the square root and then the exponential.

s^2 : The variance can also be expressed as $s_F^2 = \frac{1}{N-1} \sum_{i=1}^N F_i^2 - \frac{N}{N-1} \bar{F}^2$. To compute the pooled

variance from the daily variances, subtract the second term (computed from the daily \bar{F}) from s_F^2 to get the daily value of the first term. Average the daily values of the first term, and use the average of the daily \bar{F} values to compute the second term. Recombine to get the pooled variance.

s : Square the daily values of s to get daily variances. Compute the pooled variance as above, then take the square root to get the pooled standard deviation.

r : Multiply the daily correlations by the daily $s_F \times s_O$ to get the covariance, s_{FO} . The covariance can be expressed as $s_{FO} = \frac{1}{N-1} \sum_{i=1}^N F_i O_i - \frac{N}{N-1} \bar{F} \bar{O}$. Follow the steps given for s^2 above to get a

pooled covariance. Divide by the product of the pooled standard deviations to get the pooled correlation.

MAE_SS, MSE_SS: Use the pooled values of *MAE* or *MSE* to compute the skill scores.

In addition to aggregating scores, it is often useful to show their distribution. This can be done using box-whisker plots, where the inter-quartile (25th to 75th percentile) of values is shown as a box, and the whiskers show the full range of values, or sometimes the 5th and 95th percentiles. The median is drawn as a horizontal line through the box, with a "notch" often shown to indicate the 95% confidence interval on the median.

CONFIDENCE INTERVALS FOR VERIFICATION SCORES

Any verification score must be regarded as a sample estimate of the "true" value for an infinitely large verification dataset. There is therefore some uncertainty associated with the score's value, especially when the sample size is small or the data are not independent, or both. It is a good idea to estimate some confidence intervals (CIs) to set some bounds on the expected value of the verification score. This also helps to assess whether differences between competing forecast systems are real.

Jolliffe (2007) gives a nice discussion of several methods for deriving CIs for verification measures. Mathematical formulae are available for computing CIs for distributions which are binomial or normal, assumptions that are reasonable for scores that represent proportions (*PC*, *POD*, *FAR*, *TS*). In general, most verification scores cannot be expected to satisfy these assumptions. Moreover, the verification samples are often non-independent in space and/or time. A non-parametric method such as the *bootstrap method* is ideally suited for handling these data because it does not require assumptions about distributions. The bootstrap is, however sensitive to dependence of the events in the verification sample. A strategy such as "block bootstrapping", where the data is resampled in blocks which can be considered independent of each other, is recommended for datasets with high spatial or temporal correlation. This point is discussed further below.

The non-parametric bootstrap is quite simple to do:

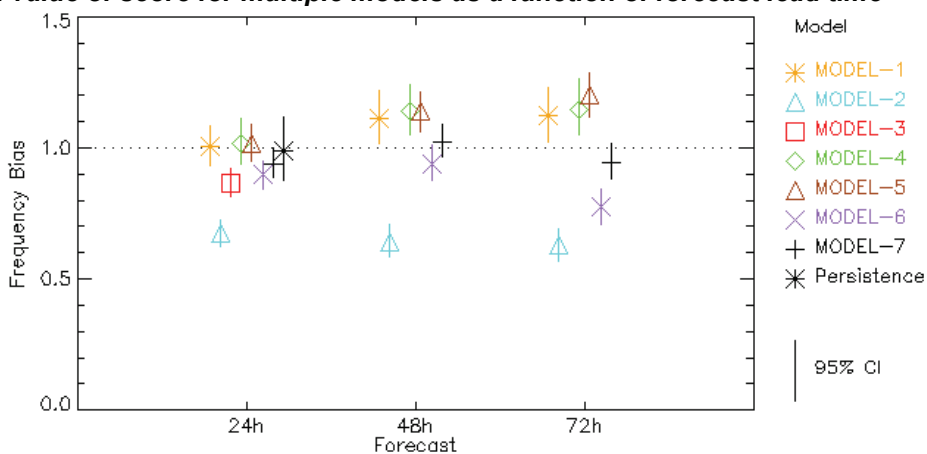
1. Generate a bootstrap sample by randomly drawing N forecast/observation pairs from the full set of N samples, *with replacement* (i.e., pick a sample, put it back, N times).
2. Compute the verification statistic for that bootstrap sample.
3. Repeat steps 1 and 2 a large number of times, say 1000, to generate 1000 estimates for that verification statistic.
4. Order the estimates from smallest to largest. The $(1-\alpha)$ confidence interval is easily obtained by finding the values for which the fraction $\alpha/2$ of estimates are lower and higher, respectively.

When comparing the scores for two or more forecasts one can use the degree of overlap between their confidence intervals to judge whether the differences between the forecasts are likely to be significant. A more precise method is to calculate confidence intervals for the mean *difference* between the scores. If the $(1-\alpha)$ confidence interval does not include 0, then the performance of the forecasts can be considered significantly different.

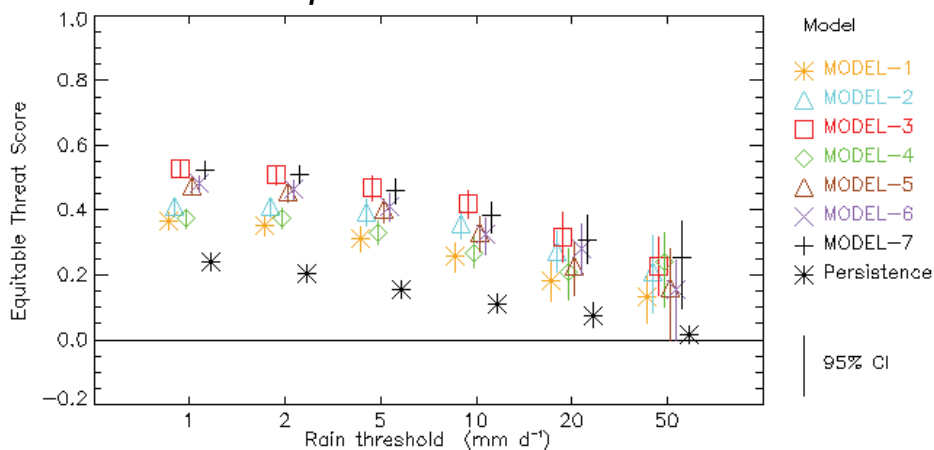
As mentioned above, bootstrap sampling also implies independent events in the sample. Thus it is often necessary to sample in "blocks" to obtain more reasonable estimates of the confidence limits. For example, if there is high spatial correlation in the dataset, which is often the case in gridded forecasts, then each full grid might be sampled as a single block. Or, if there is also temporal correlation in the forecasts, then it could be necessary to form blocks of two or three successive forecasts.

EXAMPLES OF GRAPHICAL VERIFICATION PRODUCTS

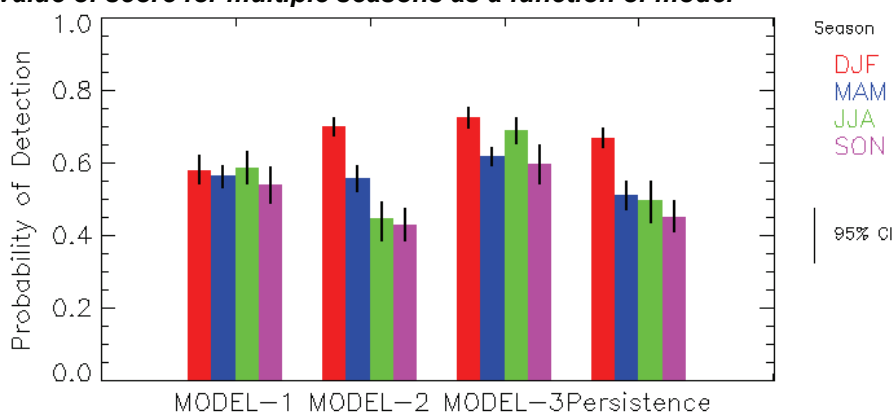
1. Pooled value of score for multiple models as a function of forecast lead time



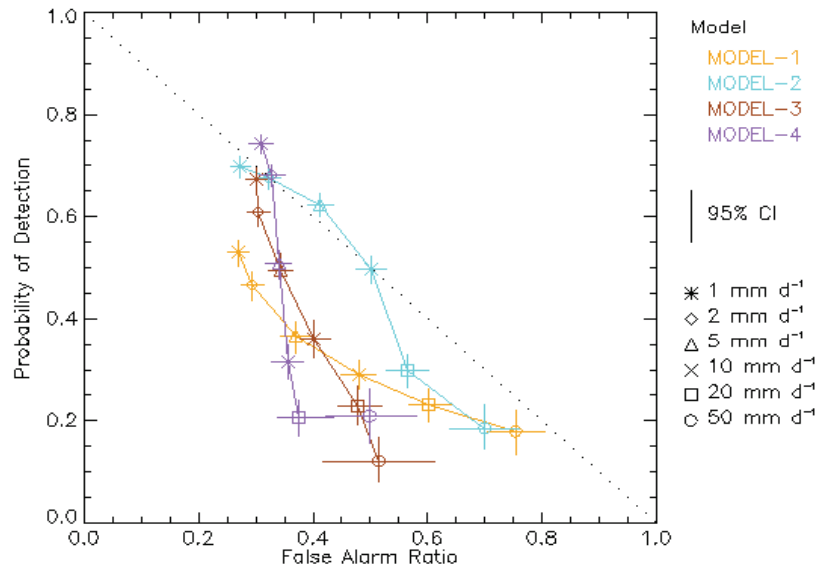
2. Pooled value of score for multiple models as a function of rain threshold



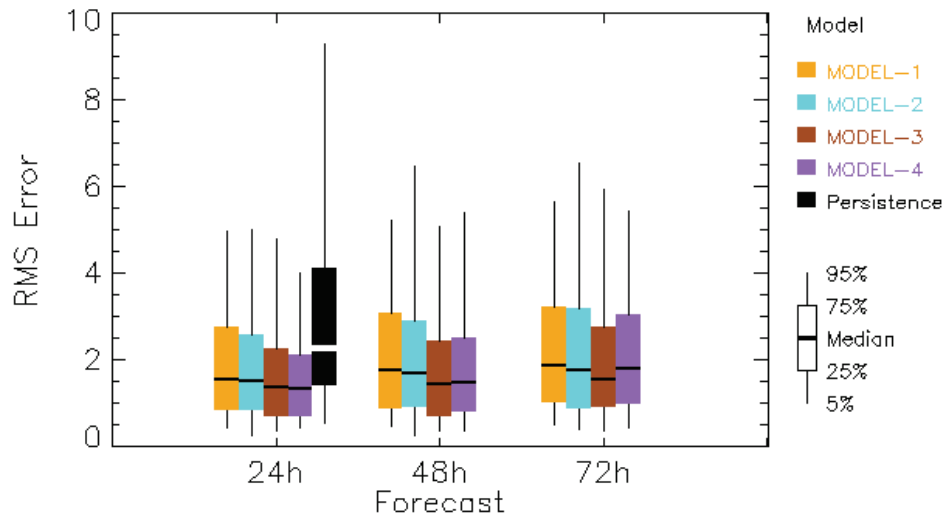
3. Pooled value of score for multiple seasons as a function of model



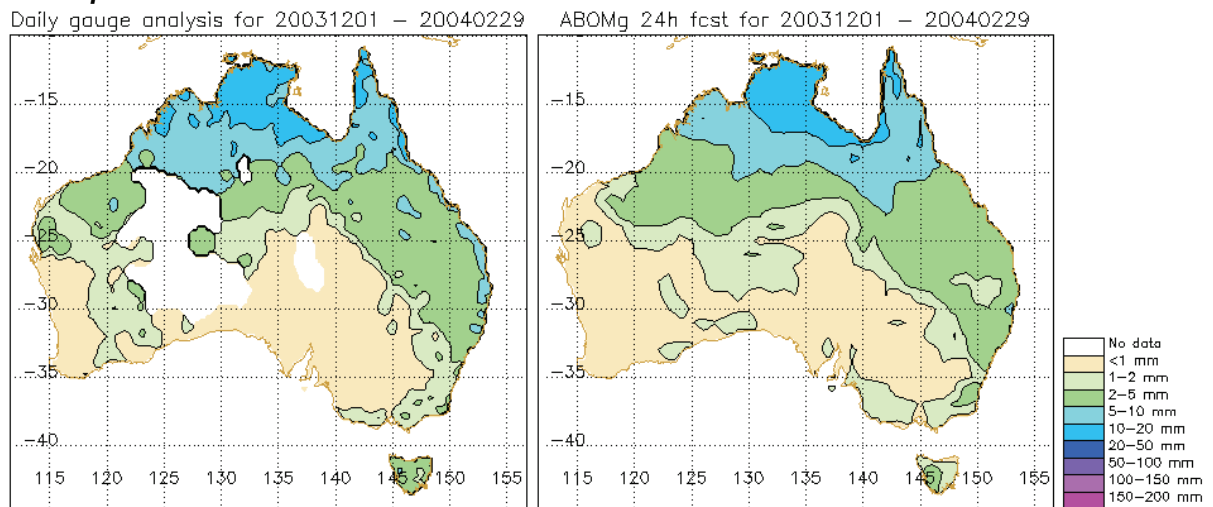
4. POD vs FAR for multiple models as a function of rain threshold



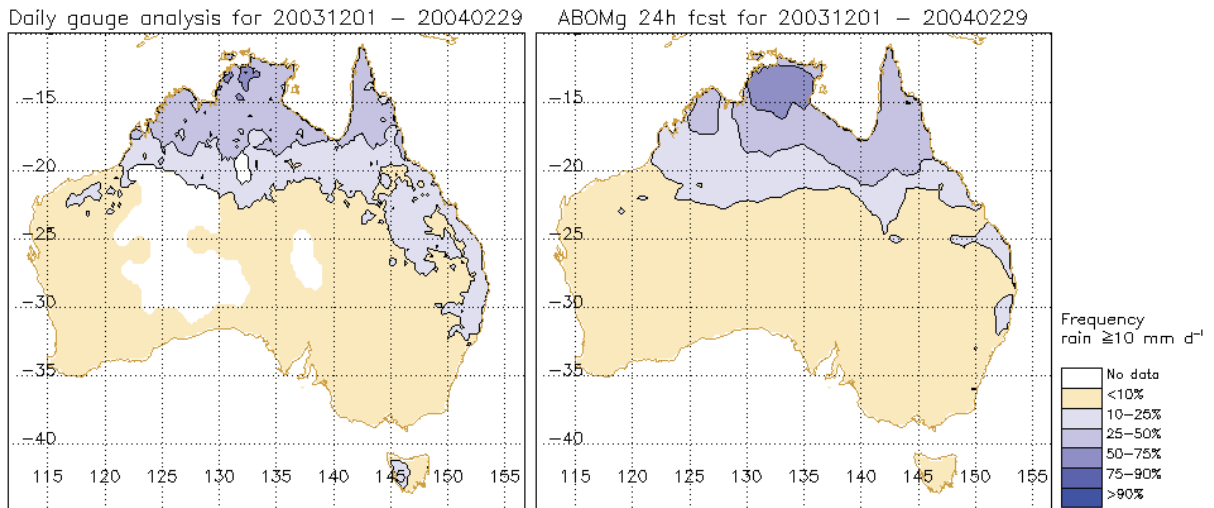
5. Box plot of daily values of score for multiple models as a function of lead time



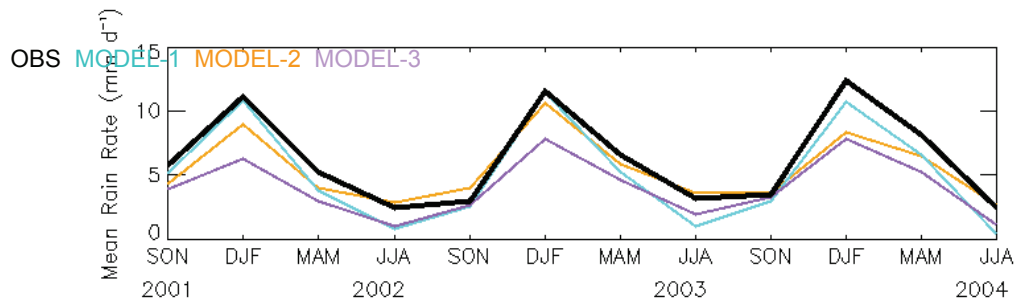
6. Maps of forecast and observed mean seasonal rainfall



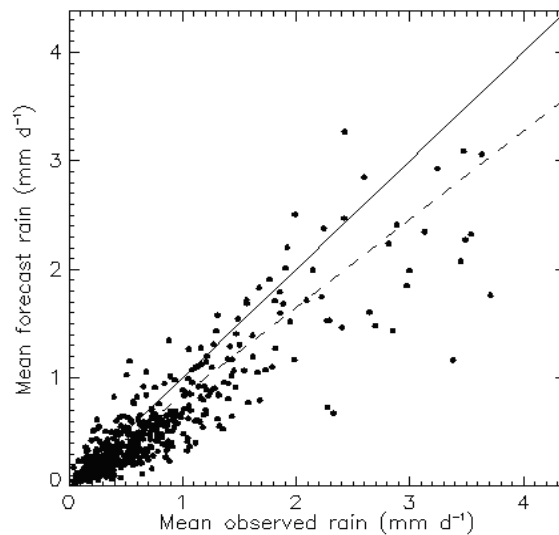
7. Maps of forecast and observed frequency of rain exceeding 10 mm d⁻¹



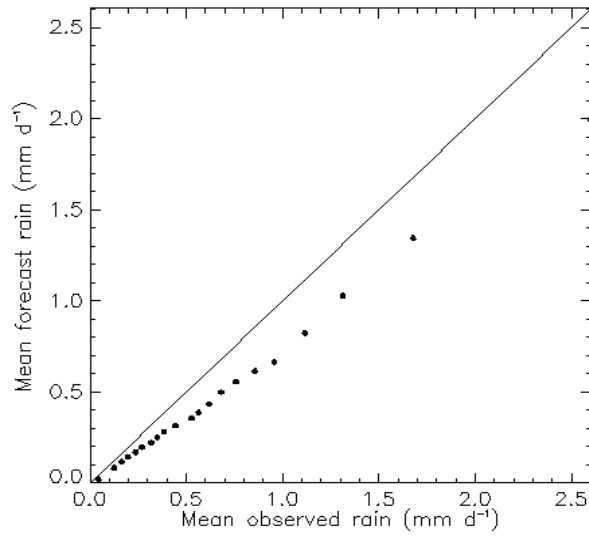
8. Seasonal time series of forecast and observed mean rainfall



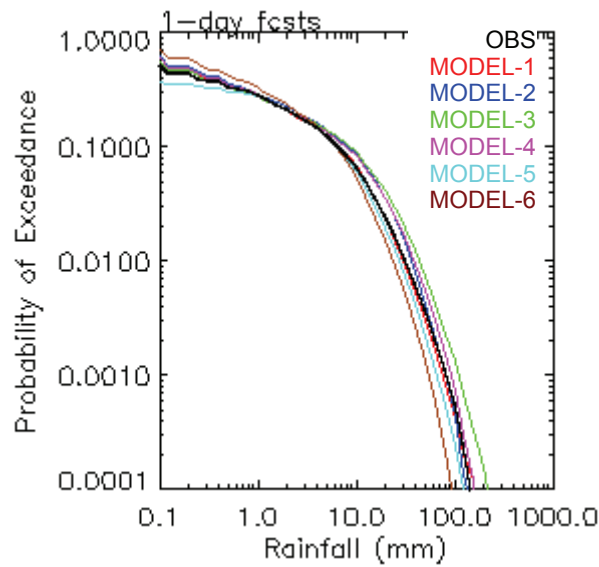
9. Scatter plot of forecast versus observed rainfall. The dashed line shows the best fit to the data when normalized using a square root transformation



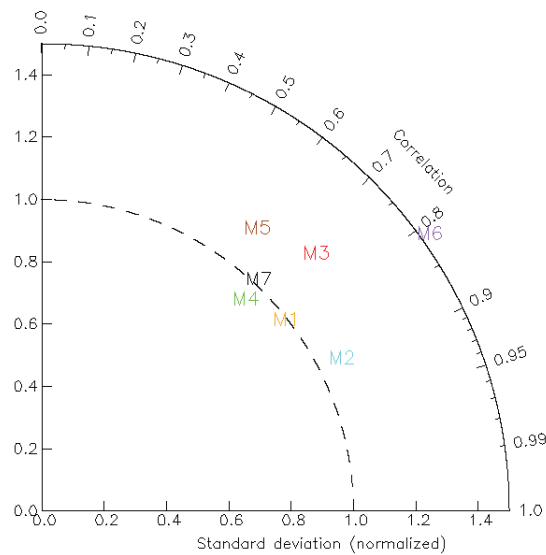
10. **Quantile-quantile plot of forecast versus observed rainfall. Quantiles are given in 5% increments**



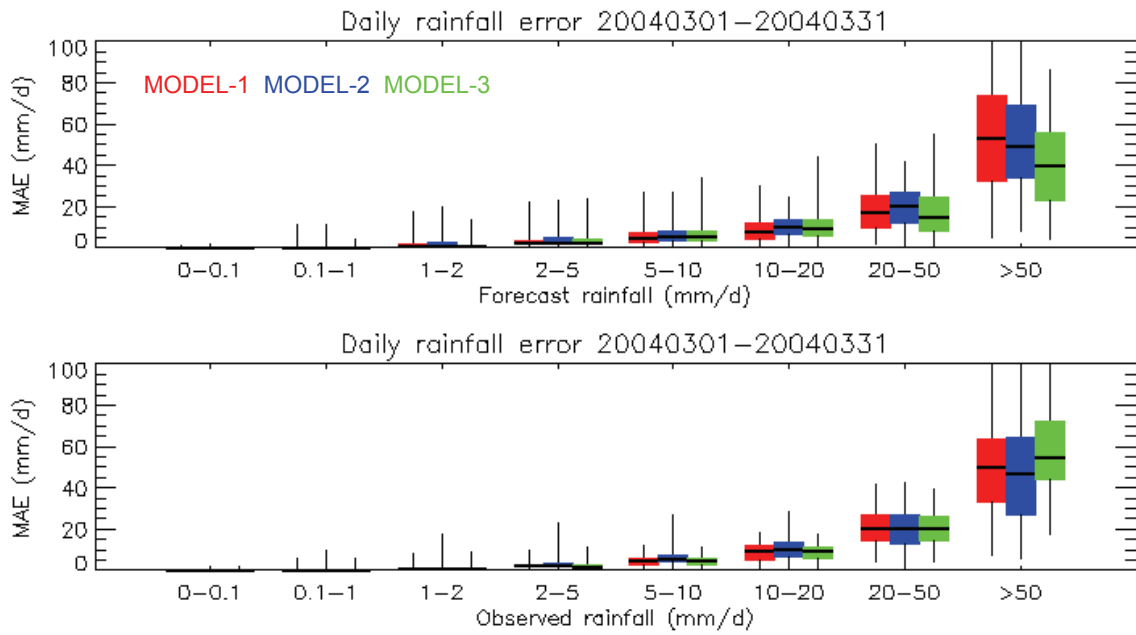
11. **Exceedance probability for forecast and observed rainfall**



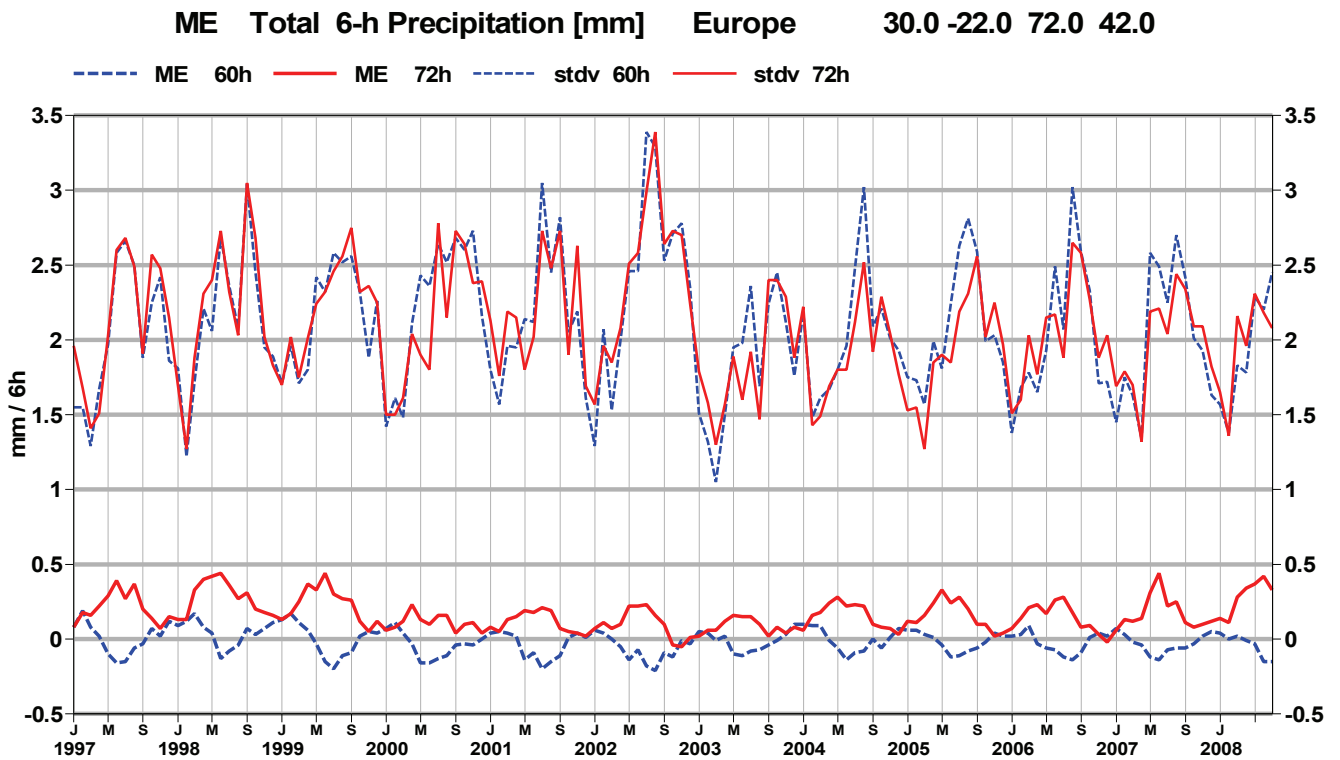
12. **Non-dimensional Taylor diagram**



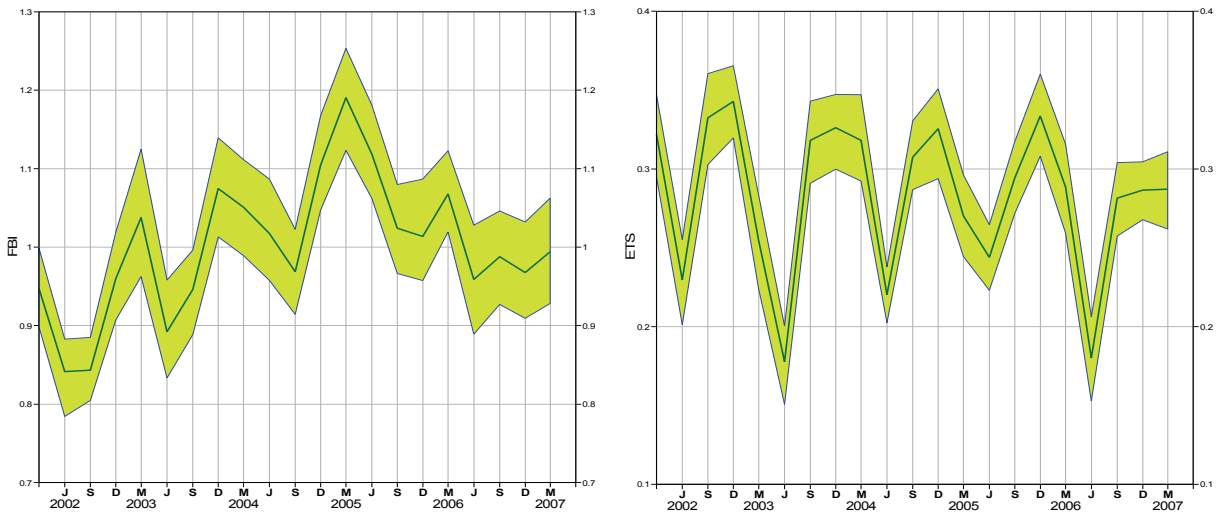
13. **Box plot of daily values of score for multiple models as a function of rain range for forecast or observed rain**



14. **Time series of monthly ME for 6h accumulated precipitation over Europe**

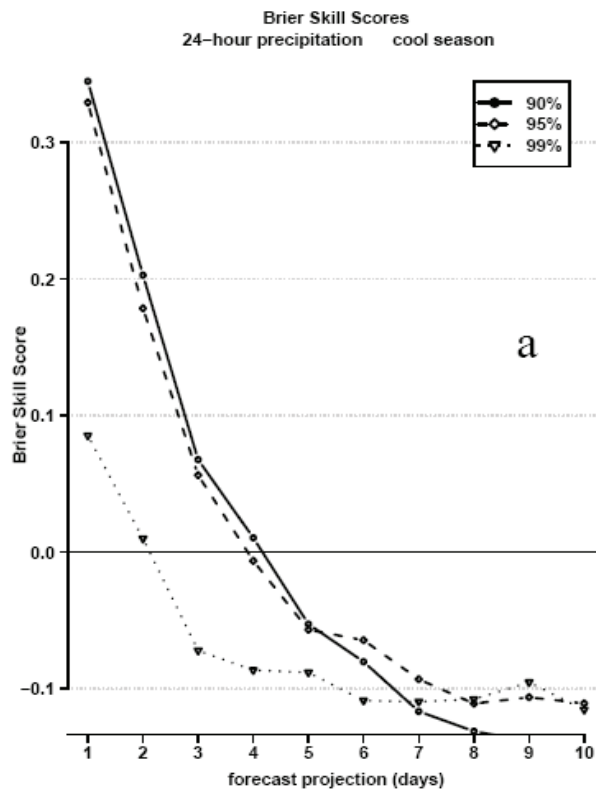


15. **Time series of Frequency Bias Index (BIAS) and ETS for 42h forecasts of accumulated precipitation > 10mm/24h, verified against gridded observations according to Cherubini et al. (2002). The score is calculated for each season (DJF, MAM, JJA and SON). The green shaded area represents the 5 to 95% confidence range on the bias values**

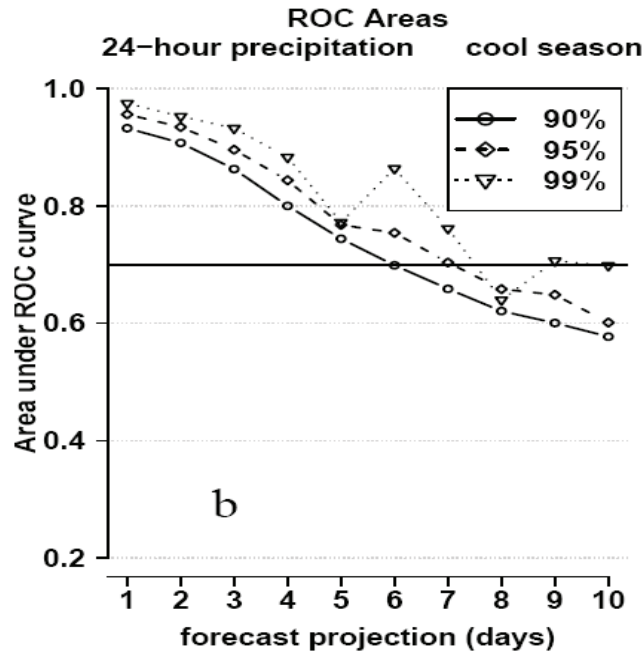


16 to 18: These are examples of verification of ensemble forecasts. The main intent of the verification was to determine the ability of the ensemble to forecast extremes of precipitation, defined as the 90th, 95th and 99th percentiles of the long term climatological distribution at each station. Since verification samples of such extremes are inherently small, it was necessary to pool the data over different stations. The underlying climatology has been “standardized” by using distribution centiles rather than physical thresholds for the pooled verification data. This avoids the problems described by Hamill and Juras (2006), while increasing the sample size to more acceptable levels.

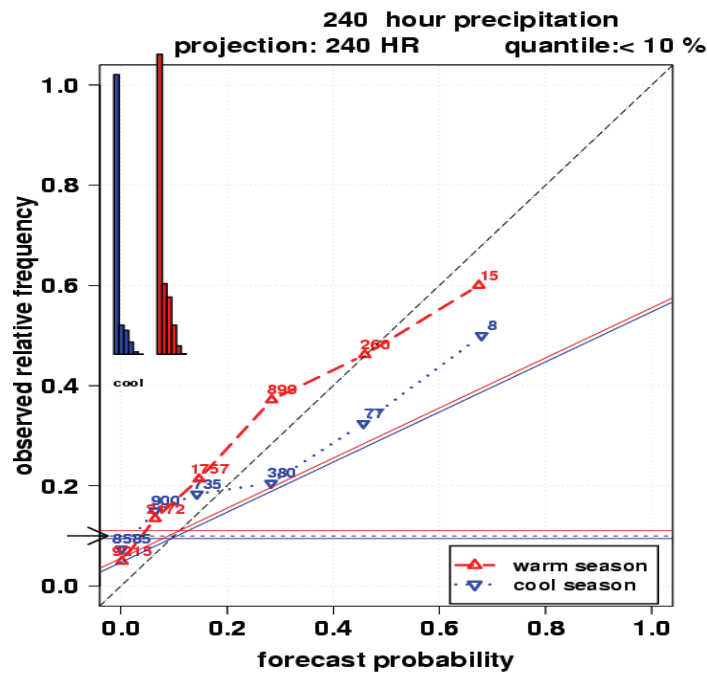
16. **Brier Skill Score as a function of forecast projection for 3.5 years of 24h winter season precipitation accumulation forecasts for 36 stations. Skill is positive to 4 days or so for the 90th and 95th percentiles, but only for day 1 and 2 for the 99th percentile. (After Peel and Wilson 2008b)**



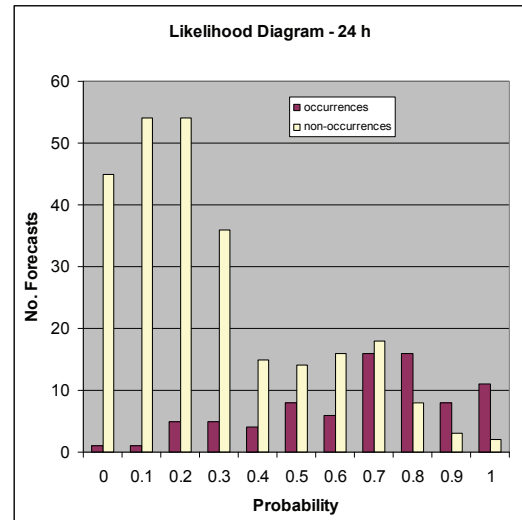
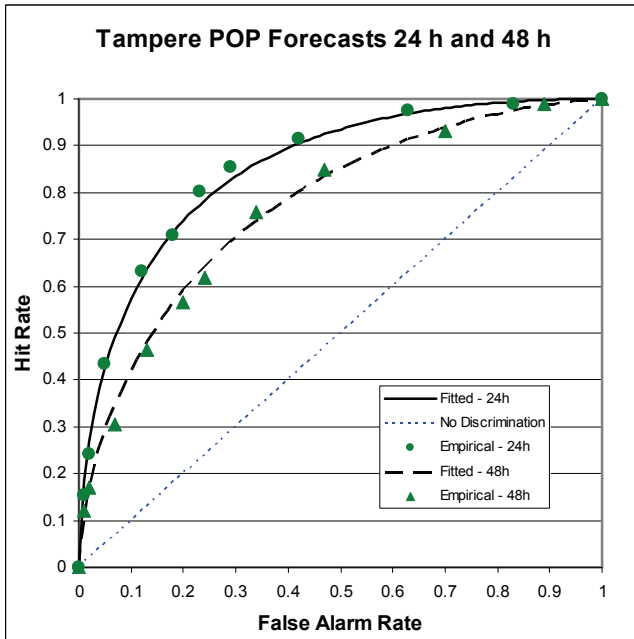
17. **ROCA for 3.5 years of 24h precipitation forecasts exceeding 90th, 95th and 99th climatological thresholds for 36 stations, as a function of forecast projection time. Although the BSS is lower for the higher thresholds, the ensemble shows greater discrimination ability for the higher threshold than the lower one. A ROCA of 0.7 is sometimes considered to represent the lowest useful discrimination ability in practice. (After Peel and Wilson, 2008b)**



18. **Reliability table for winter season (blue) and summer season (red), for probability of 10 day precipitation less than the 10th percentile of the long term climatology, over 36 stations. The histograms show the sharpness of the forecasts; the number of cases in each probability bin is also shown next to the plotted points of the reliability graph. The horizontal lines show the sample climatological frequency of the event, for each of the two samples. The dashed horizontal line is the long term climatological frequency, 0.1 in this case. The lines which lie halfway between the diagonal and the climatology are the “no skill” lines, where the reliability component of the Brier Score exactly balances the resolution component, giving a skill score of 0.**



19. ROCs, empirical and fitted with the binormal model, for 24h and 48h forecasts of probability of precipitation. The ROCA is 0.86 and 0.77 for the 24- and 48-h forecasts respectively. On the graph, the diagonal is the “no discrimination” line, for which the forecast probability distribution for cases where the event occurred lies on top of the forecast probability distribution for the cases where the event did not occur, and the ROCA is 0.5. Perfect discrimination is represented by a curve which travels up the left hand side of the diagram, then across the top, giving an area of 1.0. The ROC is not sensitive to bias, and should be used in combination with the reliability table scores for a more complete diagnostic verification. The likelihood diagram for the 24h forecasts is shown at right.



MEMBERSHIP OF WWRP/WGNE JOINT WORKING GROUP ON VERIFICATION (JWGV) (2008)

Barb Brown (chair), National Center for Atmospheric Research, USA

Joël Stein, Météo-France, France

Harold Brooks, National Severe Storms Laboratory, USA

Barbara Casati, Ouranos, Canada

Martin Göber, Deutscher Wetterdienst, Germany

Beth Ebert, Bureau of Meteorology, Australia

Anna Ghelli, ECMWF

Marion Mittermaier, The Met Office, UK

Pertti Nurmi, Finnish Meteorological Institute, Finland

David Stephenson, University of Reading, UK

Clive Wilson, The Met Office, UK

Laurie Wilson, Meteorological Research Division, Canada

World Weather Research Programme (WWRP) Report Series

Sixth WMO International Workshop on Tropical Cyclones (IWTC-VI), San Jose, Costa Rica, 21-30 November 2006 (WMO TD No. 1383) (**WWRP 2007 - 1**).

Third WMO International Verification Workshop Emphasizing Training Aspects, ECMWF, Reading, UK, 29 January - 2 February 2007) (WMO TD No. 1391) (**WWRP 2007 - 2**).

WMO International Training Workshop on Tropical Cyclone Disaster Reduction (Guangzhou, China, 26 - 31 March 2007) (WMO TD No. 1392) (**WWRP 2007 - 3**).

Report of the WMO/CAS Working Group on Tropical Meteorology Research (Guangzhou, China, 22-24 March 2007) (WMO TD No. 1393) (**WWRP 2007 - 4**).

Report of the First Session of the Joint Scientific Committee (JSC) for the World Weather Research Programme (WWRP), (Geneva, Switzerland, 23-25 April 2007) (WMO TD No. 1412) (**WWRP 2007 - 5**).

Report of the CAS Working Group on Tropical Meteorology Research (Shenzhen, China, 12-16 December 2005) (WMO TD No. 1414) (**WWRP 2007 - 6**).

Preprints of Abstracts of Papers for the Fourth WMO International Workshop on Monsoons (IWM-IV) (Beijing, China, 20-25 October 2008) (WMO TD No. 1446) (**WWRP 2008 - 1**).

Proceedings of the Fourth WMO International Workshop on Monsoons (IWM-IV) (Beijing, China, 20-25 October 2008) (WMO TD No. 1447) (**WWRP 2008 - 2**).

WMO Training Workshop on Operational Monsoon Research and Forecast Issues – Lecture Notes, Beijing, China, 24-25 October 2008 (WMO TD No. 1453) (**WWRP 2008 - 3**).

Expert Meeting to Evaluate Skill of Tropical Cyclone Seasonal Forecasts (Boulder, Colorado, USA, 24-25 April 2008) (WMO TD No. 1455) (**WWRP 2008 - 4**).