Recommendations for the verification and intercomparison of QPFs from operational NWP models

WWRP/WGNE Joint Working Group on Verification

December 2004

Contents

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 the community. Many countries have national rain gauge networks that provide observations that can be used to verify the model QPFs. Since rainfall depends strongly on the 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 Centers 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 QPF verification/intercomparison to WGNE participants, it should be as similar as possible across all regions. It is probably not be feasible to change the rainfall observation times, but it is certainly desirable for participating centers to use a common verification methodology. The WWRP/WGNE Joint Working Group on Verification (JWGV, see Appendix 4) was asked in 2003 to provide recommendations on a standard verification methodology to be used for QPFs.

The purpose of this document is to recommend a standard methodology for verification and intercomparison of QPFs from NWP models. Verification of deterministic QPFs (i.e., those giving a precise value for rain amount) is considered here; verification of probabilistic forecasts will be the subject of a future document .

A number of recent reports have summarized the state of the art in precipitation verification and made specific recommendations for the practice of forecast verification. Wilson (2001) surveyed operational centers in Europe as to their QPF verification practices. He found that most centers 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. He argues that observations that have been "upscaled" by averaging to the model's grid resolution give a fairer evaluation of model performance than direct comparison to gauge data, which ignores scale differences.

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 centers 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: (1) large scale aggregated

overall measures of forecast performance in tropics, northern extra-tropics and southern extra-tropics, (2) verification at gridpoints (maps), and (3) 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.

ECMWF recently 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.

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 are described 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.

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 most relevant to the WGNE QPF study, while individual NWP centers also have a strong interest in how to improve their forecasts (c). Different verification methods may be appropriate for each. 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 using a comprehensive set of verification scores is needed. 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 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 (2004)).

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

(b) Spatial matching

The primary interest of WGNE is anticipated to be model-oriented verification. By this, we mean the evaluation of model output on the space and time scales of the models. It addresses the question of whether the models are producing the best possible forecasts given their constraints on spatial and temporal resolution. For model-oriented verification Cherubini et al. (2002) showed that gridded, "upscaled", observations representing gridbox averaged rainfall are more appropriate than synoptic "point" observations because the spatial scales are better matched. The gridding of observations is discussed in Section 3. A fair intercomparison of QPFs from different models requires that they be mapped onto a common grid. This is because forecasts at coarse resolution tend to score better according to standard measures than those at high resolution. Acadia et al. (2003) found that a simple nearest-neighbor averaging method was less likely than bilinear interpolation to artificially smooth the precipitation field when downscaling coarser model output to a higher resolution grid.

On the other hand, users of forecasts typically wish to know their accuracy for particular locations. This is especially relevant now that 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). It should be pointed out that the verification against a set of station observations, as is required in other WMO intercomparison programs (e.g. WMO, 2002), 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. 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, i.e., 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 undersample the spatial distribution of precipitation. Members of the JWGV are divided as to which approach is preferable, but strongly agree that both give important information on forecast accuracy.[1]

We recommend that verification be done both against
(a) gridded observations (model-oriented verification) on a common 0.5° latitude/longitude grid
(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 is due to the large number of 24h rainfall observations available from national rain gauge networks. 24h observations are less prone to observational and representativeness errors than those for shorter periods. (Radar and satellite can provide rainfall information with high spatial and temporal resolution -- the use of these remotely sensed rain estimates as reference data will be discussed Section 3.)

For the model intercomparison is it 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 but 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 behavior 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. Some scores such as the correlation coefficient and the ETS 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. 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 Appendix 1.

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

[1]Tustison et al. (2001), advocate using an 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, 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]. 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!).

We recommend that the verification of persistence forecasts be reported along with the model verification. The verification of climatology forecasts is highly desirable. The use of model 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 samples 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 error. Confidence intervals contain more information about the uncertainty of a score than a simple significance test, and can be fairly easily computed using bootstrapping methods (see Appendix 2). The median and interquartile range (middle 50% of the sample distribution reported as the 25^{th} and 75^{th} percentiles) give the "typical" values for the score.

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 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.

The verification should strive to use all available observations. Some countries have cooperative networks of rainfall observers who report to their meteorological centers well outside of real time.

[2]The 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.

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 centers. 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. Experiments with several objective analysis schemes (including averaging) suggest that the accuracy of the analysis is determined primarily by the density of observations, rather than the chosen method of analysis (G. Weymouth, unpublished results).

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 in WGNE. Research on the best use of reference data uncertainty in the verification process is strongly encouraged.

(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, satellites 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.

We recommend that, where possible, combined radar-gauge rainfall analyses be used to verify model QPFs at high spatial and temporal resolution. Research on using satellite estimates for verifying the forecast rainfall location and timing in remote regions should be encouraged.

4. Verification methods

Verification begins with a matched set of forecasts and observations for identical locations and times, traditionally treated as independent samples. In reality there are often strong among-sample correlations within subsets of the data (coherent spatial patterns, for example), which some of the advanced diagnostic methods explicitly take into account (JWGV, 2004).

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, many of the continuous verification scores (especially those involving squared errors) are sensitive to large errors.

For this reason categorical verification scores may give more meaningful information for precipitation verification.

A large variety of verification scores are used operationally to verify QPFs (Wilson, 2001; Nurmi, 2004). Details of these scores can be found in the textbooks of Wilks (1995) and Jolliffe and Stephenson (2003), or on the JWGV (2004) 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 give a very short definition of each score 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 WGNE participants may not want to wade through a large array of scores.

Note that once the forecast and observed data have been extracted and the code written to compute the *highly recommended* scores, it is very little extra effort 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 forecasts events and non-events is shown by the 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 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 *frequency bias* (***) gives the ratio of the forecast rain frequency to the observed rain frequency.

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

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

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

The *probability of detection* (*POD*) (***) measures the fraction of observed events that were correctly forecast.

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

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

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

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.

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

The *threat score* (*TS*) (**) 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*) (***) 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

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

The *Hanssen and Kuipers score* (*HK*) (***) 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.[3]

$$HSS = \frac{2\,(hits \times correct\ rejections - misses \times false\ alarms)}{observed\ yes \times forecast\ no + forecast\ yes \times observed\ no}$$

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.

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

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

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

(b) Forecasts of rain amount

[3]The *HSS* is related to the *ETS* according to *HSS* = 2 *ETS* / (1+*ETS*) (Schaefer, 1990) and therefore gives no new information. It is included here because it is familiar to many users.

Other statistics measure the quality of forecasts of a continuous variable such as rain amount. 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.

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

$$\overline{O} = \frac{1}{N}\sum_{i=1}^{N} O_i \qquad \overline{F} = \frac{1}{N}\sum_{i=1}^{N} F_i$$

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

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

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.

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

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.

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

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

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

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

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

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.

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

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 x/÷ *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.

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

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*.

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

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 is persistence for forecasts of 24h or less, and climatology for longer forecasts.

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

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

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.

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

(c) 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.**

(d) Advanced diagnostic methods (*)

Several advanced diagnostic methods have proven very useful in a research setting, and we encourage WGNE scientists to begin to experiment with them in research and operations. Some examples include multiscale spatial statistics (e.g., Harris et al., 2001), scale decomposition methods (e.g., Casati et al., 2004), object oriented methods (Ebert and McBride, 2000; Bullock et al., 2004), and spatial multi-event contingency tables (Atger, 2001). More information on these and other methods can be found on the JWGV (2004) web site.

5. Reporting guidelines

The QPF verification will be most useful to WGNE participants 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.

We recommend that the system description and a full selection of numerical and graphical verification results be accessible from a user-friendly web site that is updated on a regular basis. Password protection should be included to ensure confidentiality.

(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 then 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 Appendix 3.

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
(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. 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.

The verification of persistence forecasts should be reported along with the model verification. The verification of climatology forecasts is highly desirable. The use of model skill scores with respect to persistence, climatology, and random chance is highly desirable.

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

Quality-controlled point and gridded observations from rain gauge networks should be the primary reference data for verifying model QPFs in WGNE. Research on the best use of reference data uncertainty in the verification process is strongly encouraged.

Where possible, combined radar-gauge rainfall analyses should be used to verify model QPFs at high spatial and temporal resolution. Research on using satellite estimates for verifying the forecast rainfall location and timing in remote regions should be encouraged.

The verification results should include the following information:

Model: (ex: ECMWF)
Verification Country: (ex: Australia)
Region: (ex: Tropics)
Season and year: (ex: JJA 2003)
Initialization time: (ex: 00 UTC)
Forecast projection: (ex: 48h)
Accumulation period: (ex: 24h)
Spatial scale: (ex: 1° grid)

Forecast	**Highly recommended**	**Recommended**
Rain occurrence ≥ specific thresholds: 1, 2, 5, 10, 20, 50 mm d^{-1}	Elements of contingency table: *hits*, *misses*, *false alarms*, *correct rejections* *PC, BIAS, POD*, *FAR*, *ETS, HK* (including 95% confidence intervals on scores)	*POFD*, *TS*, *HSS*, *OR*, *ORSS* (including 95% confidence intervals on scores) maps, time series
Rain amount	$\overline{O}$, $\overline{F}$, s_O, s_F *median O, median F* *ME*, *RMSE, r* (including 95% confidence intervals on scores) maps, time series, scatter plots	s_O^2, s_F^2 IQR_O, IQR_F *MAE, MSE, RMSF,* r_s*, MAE_SS*, *MSE_SS, LEPS* (including 95% confidence intervals on scores)

		error scatter plots, histograms, quantile-quantile plots, exceedance probability

The system description and a full selection of numerical and graphical verification results should be accessible from a user-friendly web site that is updated on a regular basis. Password protection should be included to ensure confidentiality.

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-neighbor 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.

Bougeault, P., 2002: WGNE survey of verification methods for numerical prediction of weather elements and severe weather events. *CAS/JSC WGNE Report No. 18, Appendix C*. Available on the internet at http://www.wmo.ch/web/wcrp/documents/wgne18rpt.pdf.

Bullock, R., B.G. Brown, C.A. Davis, M. Chapman, K.W. Manning, and R. Morss, 2004: An object-oriented approach for the verification of quantitative precipitation forecasts: Part I – Methodology. *20th Conf. Weather and Forecasting, Amer. Met. Soc., January 2004, Seattle.*

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.

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.

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.

Golding, B.W., 1998: Nimrod: A system for generating automated very short range forecasts. *Meteorol. Appl.*, **5**, 1-16.

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., and D.B. Stephenson, 2003: *Forecast Verification. A Practitioner's Guide in Atmospheric Science*. Wiley and Sons Ltd, 240 pp.

JWGV, 2004: Forecast verification – Issues, methods, and FAQ.
http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

Kane, T.L. and B.G. Brown, 2000: Confidence intervals for some verification measures - a survey of several methods. *15th Conference on Probability and Statistics in the Atmospheric Sciences, Amer. Met. Soc., 8-11 May 2000, Asheville, North Carolina*.

Nurmi, P., 2003: Recommendations on the verification of local weather forecasts. *ECMWF Tech. Memo. 430*, 18 pp. Available on the internet at
http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm430.pdf

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Royal Met. Soc.*, **126**, 649-667.

Schaefer, J.T., 1990: The critical success index as an indicator or forecasting skill. *Wea. Forecasting*, **5**, 570-575.

Stephenson, D.B., R. Rupa Kumar, F.J. Doblas-Reyes, J.-F. Royer, F. Chauvin, and S. Pezzulli, 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., 1995: *Statistical Methods in the Atmospheric Sciences. An Introduction*. 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.

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

Appendix 1. 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 correction 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 $s_F^2 = \frac{1}{N-1}\sum_{i=1}^{N} F_i^2 - \frac{N}{N-1}\overline{F}^2$. To compute the pooled variance from the daily variances, subtract the second term (computed from the daily $\overline{F}$) 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 $\overline{F}$ 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 x s_O to get the covariance, s_{FO}. The covariance can be expressed as $s_{FO} = \frac{1}{N-1}\sum_{i=1}^{N} F_i O_i - \frac{N}{N-1}\overline{F}\,\overline{O}$. 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.

Appendix 2. 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. 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.

Kane and Brown (2000) give 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 or independence.

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.

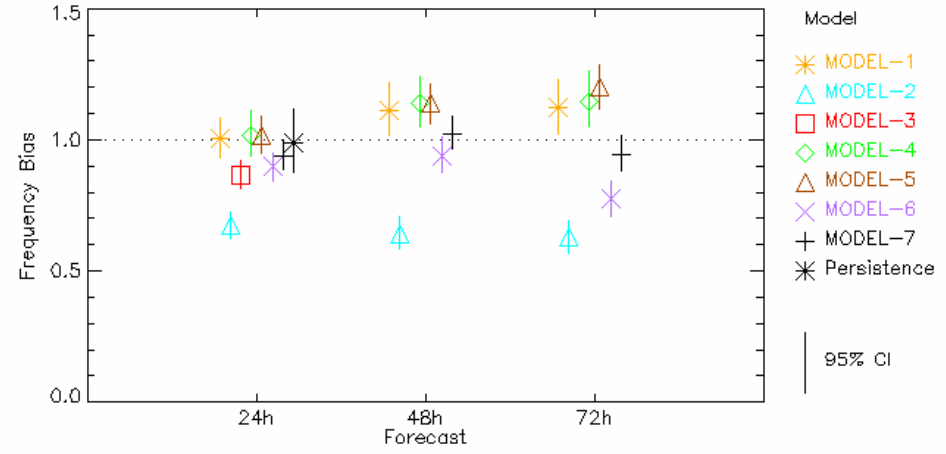
3. Repeat steps 1 and 2 a large number of times, say 1000, to generate 1000 estimates for that verification statistic.

4. Order the estimates from smallest to largest. The (1-α) confidence interval is easily obtained by finding the values for which the fraction α /2 of estimates are lower and higher, respectively.

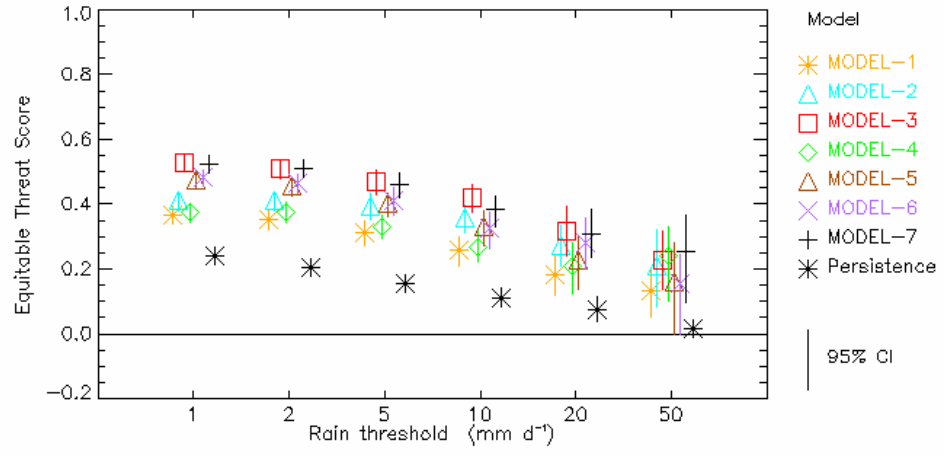
When comparing the scores for two or more forecasts one can use the degree of overlap between their confidence intervals to judge whether the differences between the forecasts are likely to be significant. A more precise method is to calculate confidence intervals for the mean *difference* between the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

Appendix 3. Examples of graphical verification products

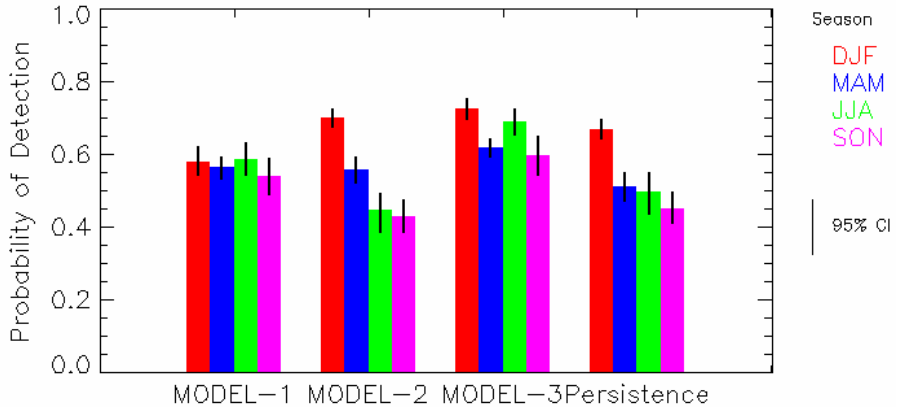
1. Pooled value of score for multiple models as a function of forecast lead time



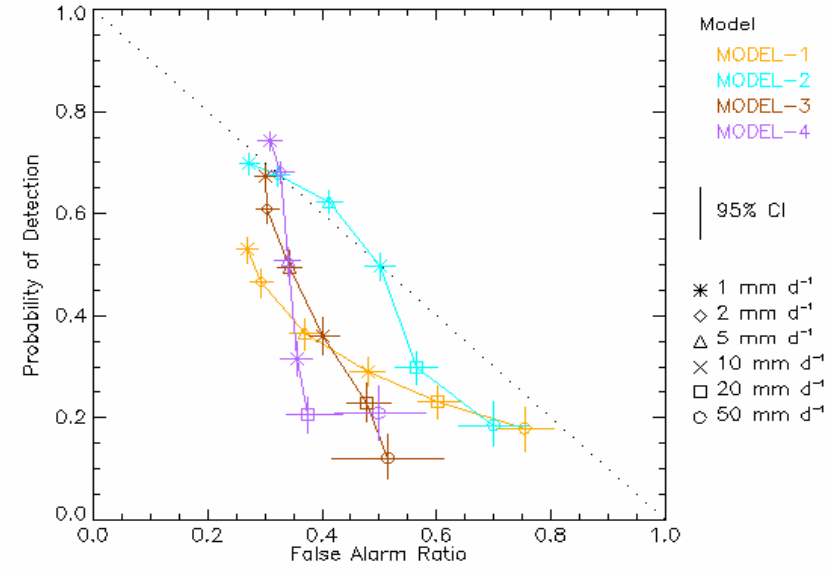
2. Pooled value of score for multiple models as a function of rain threshold



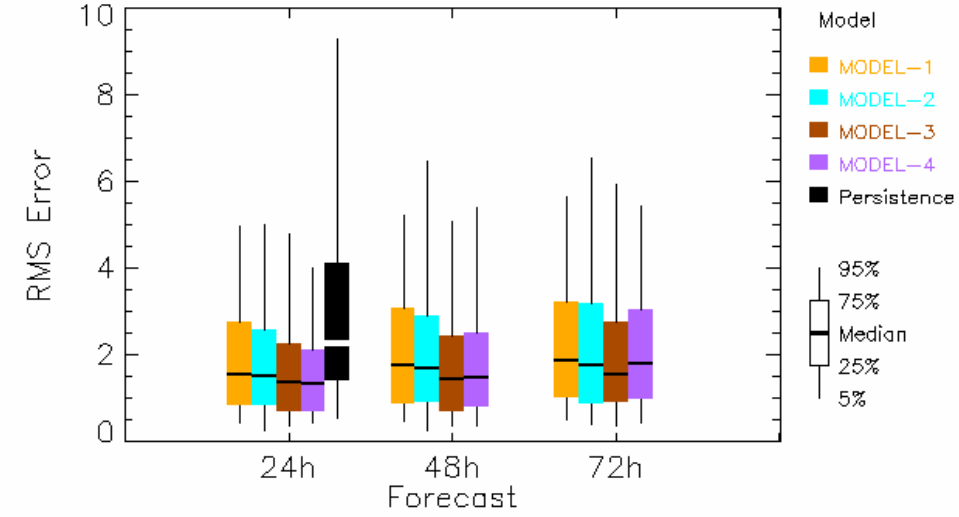
3. Pooled value of score for multiple seasons as a function of model



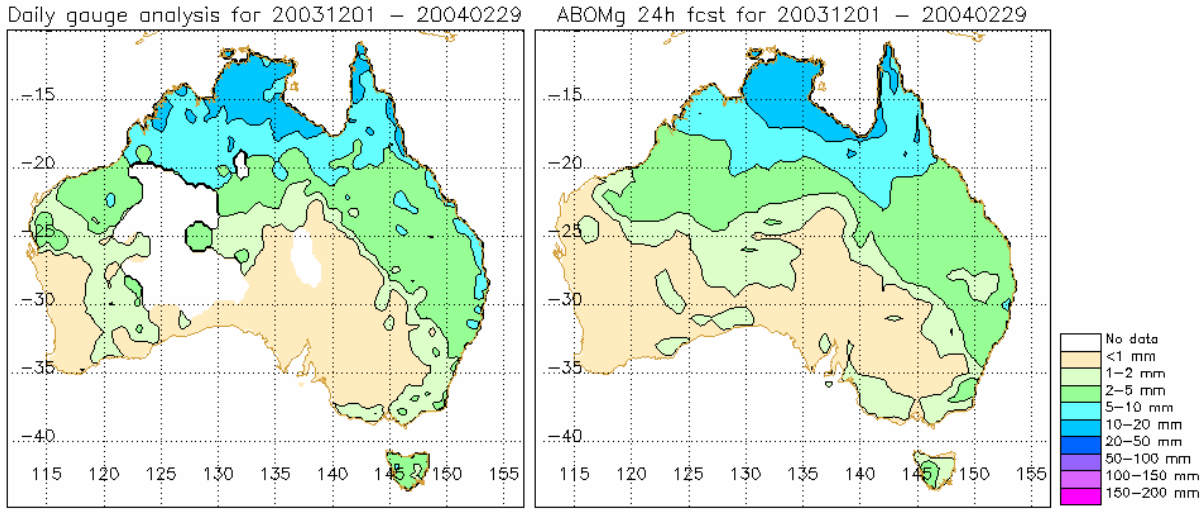
4. POD vs FAR for multiple models as a function of rain threshold



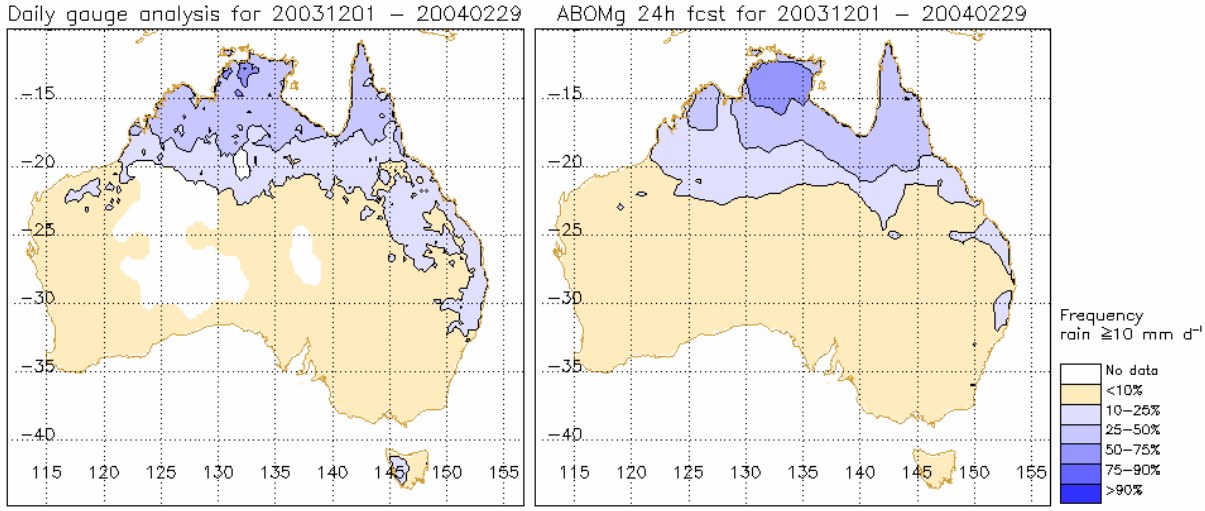
5. Box plot of daily values of score for multiple models as a function of lead time



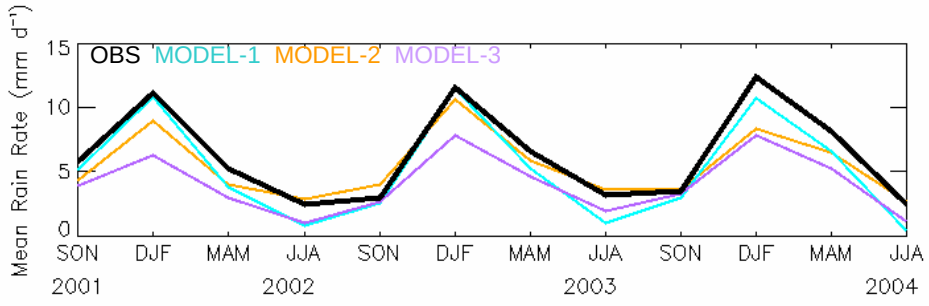
6. Maps of forecast and observed mean seasonal rainfall



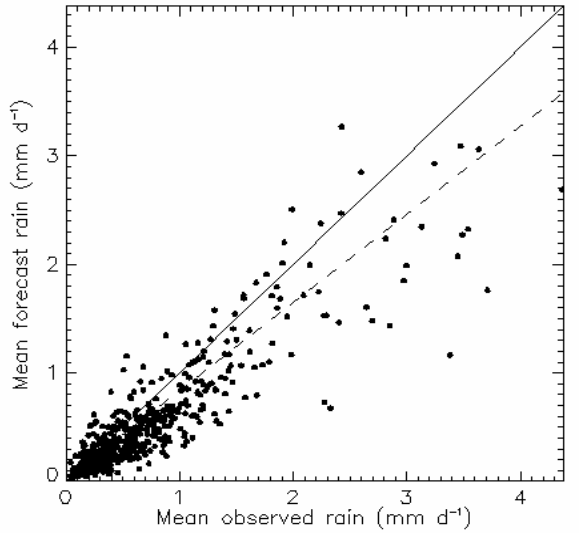
7. Maps of forecast and observed frequency of rain exceeding 10 mm d^{-1}.



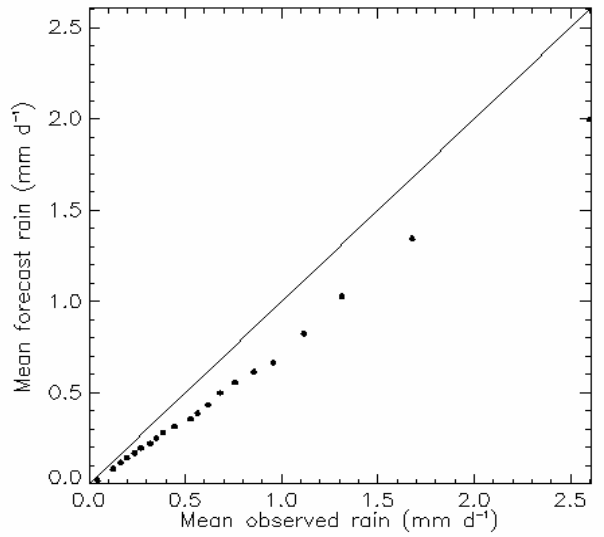
8. Seasonal time series of forecast and observed mean rainfall



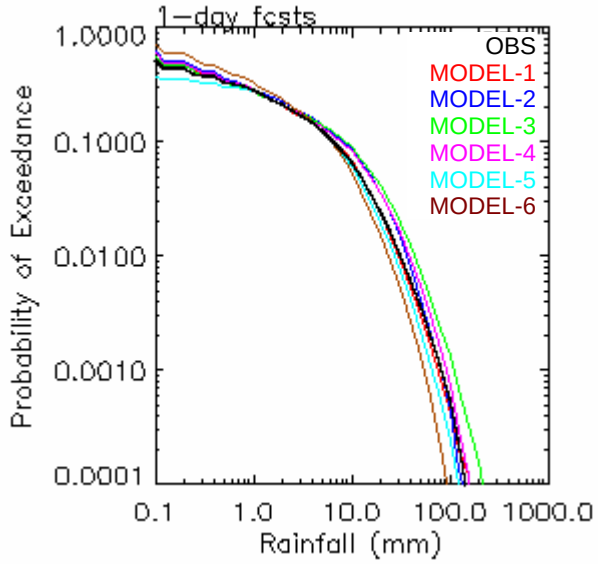
9. Scatter plot of forecast versus observed rainfall. The dashed line shows the best fit to the data when normalized using a square root transformation.



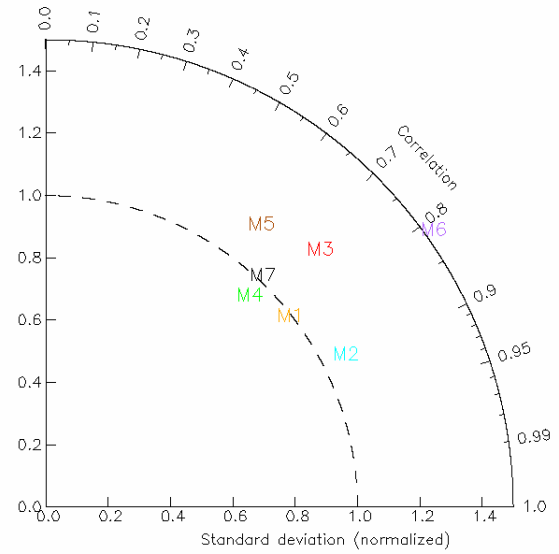
10. Quantile-quantile plot of forecast versus observed rainfall. Quantiles are given in 5% increments.



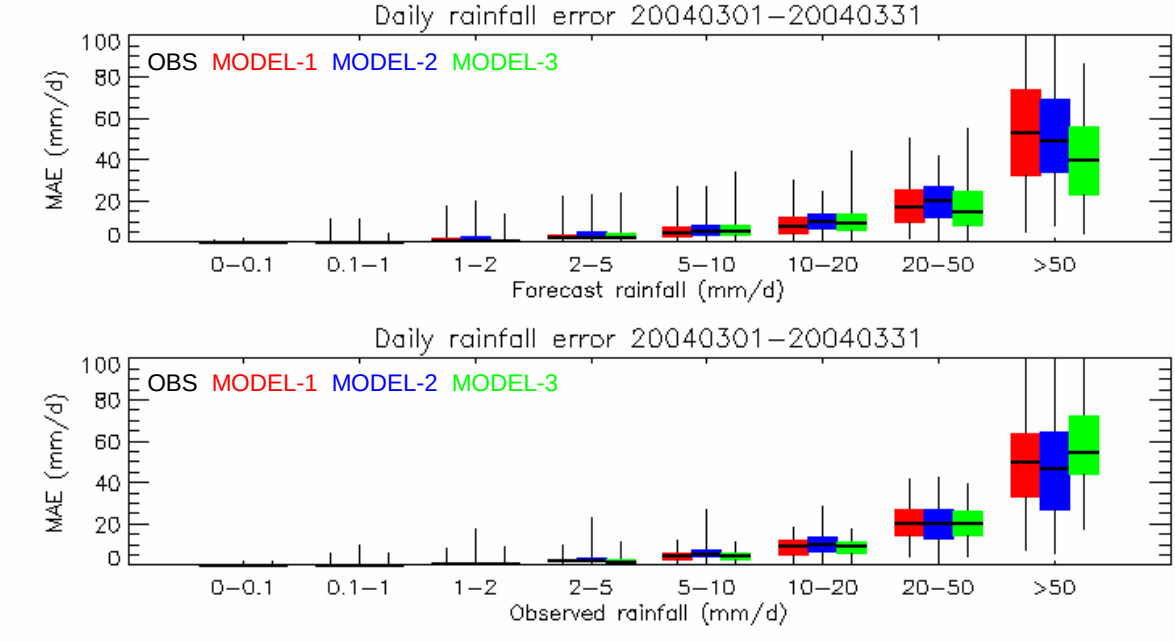
11. Exceedance probability for forecast and observed rainfall



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.



Appendix 4. Membership of WWRP/WGNE Joint Working Group on Verification (JWGV)

Barb Brown (chair), National Center for Atmospheric Research, USA
Frédéric Atger, Météo-France, France
Harold Brooks, National Severe Storms Laboratory, USA
Barbara Casati, Recherche en Prévision Numérique, Canada
Ulrich Damrath, Deutscher Wetterdienst, Germany
Beth Ebert, Bureau of Meteorology Research Centre, Australia
Anna Ghelli, ECMWF
Pertti Nurmi, Finnish Meteorological Institute, Finland
David Stephenson, University of Reading, UK
Clive Wilson, The Met Office, UK
Laurie Wilson, Recherche en Prévision Numérique, Canada