Accepted for publication in *Monthly Weather Review*, 25 April 2006

Inferential, non-parametric statistics to assess quality of probabilistic forecast systems

Aline de H. N. Maia[1], Holger Meinke[2], Sarah Lennox[2] and Roger Stone[2,3]

[1] Embrapa Meio Ambiente, PO Box 69, Jaguariúna, SP, Brazil;
ahmaia@cnpma.embrapa.br

[2] Queensland Department of Primary Industries and Fisheries, Emerging Technologies, APSRU, PO Box 102, Toowoomba, Qld 4350, Australia;
holger.meinke@dpi.qld.gov.au, sarah.lennox@dpi.qld.gov.au, roger.stone@dpi.qld.gov.au

[3] University of Southern Queensland, Faculty of Sciences, Toowoomba, Qld 4350, Australia

Abstract

Many statistical forecast systems are available to interested users. In order to be useful for decision-making, these systems must be based on evidence of underlying mechanisms. Once causal connections between the mechanism and their statistical manifestation have been firmly established, the forecasts must also provide some quantitative evidence of `quality'. However, the quality of statistical climate forecast systems (forecast quality) is an ill-defined and frequently misunderstood property. Often, providers and users of such forecast systems are unclear about what 'quality' entails and how to measure it, leading to confusion and misinformation. Here we present a generic framework to quantify aspects of forecast quality using an inferential approach to calculate nominal significance levels (p-values) that can be obtained either by directly applying non-parametric statistical tests such as Kruskal-Wallis (KW) or Kolmogorov-Smirnov (KS) or by using Monte-Carlo methods (in the case of forecast skill scores). Once converted to p-values, these forecast quality measures provide a means to objectively evaluate and compare temporal and spatial patterns of forecast quality across datasets and forecast systems. Our analysis demonstrates the importance of providing p-values rather than adopting some arbitrarily chosen significance levels such as $p < 0.05$ or $p < 0.01$, which is still common practice. This is illustrated by applying non-parametric tests (such as KW and KS) and skill scoring methods (LEPS and RPSS) to the 5-phase Southern Oscillation Index classification system using historical rainfall data from Australia, The Republic of South Africa and India. The selection of quality measures is solely based on their common use and does not constitute endorsement. We found that non-parametric statistical tests can be adequate proxies for skill measures such as LEPS or RPSS. The framework can be implemented anywhere, regardless of dataset, forecast system or quality measure. Eventually such inferential evidence should be complimented by descriptive statistical methods in order to fully assist in operational risk management.

1. Introduction

Climate variability affects the performance of many climate sensitive systems. Agricultural systems are particularly impacted by climate variability that often results in reduced production volume or quality. Decision makers, be they farmers, policy makers or agribusiness managers, need to devise sound, adaptive risk management strategies in order to improve overall systems performance and to avoid potentially disastrous system failures such as bankruptcy, environmental collapse or famine. Such sound agricultural risk management requires objective assessments of alternative but uncertain outcomes. In highly variable climates, seasonal climate forecasting in combination with simulation models of farming systems has therefore become an important tool for risk assessments and the evaluation of management options (Hammer *et al.* 2000; Sivakumar *et al.* 2000; Ferreyra *et al.* 2001; Meinke and Stone 2005).

Hence objective criteria regarding the performance of forecast systems are required (Hartmann *et al.* 2002). We assert that for appropriate risk management, statistical forecasts must be based on evidence of underpinning mechanisms. Without a plausible explanation for the observed variability in predictors it would be inappropriate to use such forecasts in decision-making. Once some mechanistic basis has been established, other quality attributes of the forecast need to be examined. With this paper we aim to contribute to this process. Information about quality and uncertainty is as important as the forecast itself in order to establish the necessary credibility amongst users. What exactly are these attributes and how should they should be measured? A WMO report (2005) emphasises that only probabilistic forecast systems[1] should be considered for risk management. We concur. The report

[1] So far, most operational, probabilistic forecast systems that connect with decision-making tools such as agricultural simulation models are based on an 'analogue year' approach, whereby climate series are segregated into classes corresponding to climate indicators such as the Southern Oscillation Index (SOI), El Niño/ Southern Oscillation (ENSO) phases, sea surface temperature (SST) phases or combination of such indicators. These classes constitute 'conditional climatologies' that need to be compared to the unconditional climatology or reference distribution (Meinke and Stone 2005).

lists four key forecast system attributes, namely (a) consistency (whether the forecasts correspond with the forecaster's judgment); (b) quality (whether the forecast correspond with the observations); (c) relevancy (whether what is forecasted is of concern to the user) and (d) value (whether the forecasts are/can be beneficial when used). Each of these attributes deserves further attention. In particular, methods that quantify the quality of probabilistic forecast systems are poorly understood and often misused (Potgieter *et al.* 2003). Therefore, we focus here solely on forecast quality by considering explicitly two aspects that are closely related but often differentiated in the literature: discriminatory ability (DA) and skill.

Here we demonstrate the application of an inferential framework for the evaluation of probabilistic, class-based forecast systems. The analogue-years approach is a frequently used example for such class-based systems and has provided valuable information for decision-makers in many world regions (e.g. Messina *et al.* 1999; Singels *et al.* 1997; De Jager *et al.* 1998; Meinke and Hochman 2000; Nelson *et al.* 2002; Podestá *et al.* 2002; Selvaraju *et al.* 2004).

According to Stone *et al.* (2000), DA is the ability of the forecast system to partition the unconditional probability distribution (also referred to as 'climatology') of the variable of interest (e.g. rainfall, temperature, yield, drainage, runoff) into conditional distributions corresponding to each class or phase within the forecast system (such as, the consistently negative, consistently positive, falling, rising and neutral phases of the SOI-phase system). They emphasise that DA '*does not necessarily imply the level of forecasting skill that would be determined from a test on independent data of forecast model performance*'. DA represents the additional knowledge about future states arising from some forecast system over and above the total variability of the prognostic variable (climatology in the case of our study here). Note that *discriminatory ability* as defined by Stone *et al.* (2000) is different to forecast discrimination (Wilks 1995; Murphy 1993). DA is concerned with distributions of observations only and does not attempt to make any comparison between forecasts and observations (in contrast to *forecast discrimination*).

Most skill measures were originally designed to quantify changes in the agreement between observed and predicted **values** (accuracy) of deterministic forecasts with some attempts to incorporate probabilistic properties (Mason 2004). Skill measures are supposed to account for changes in accuracy, relative to using the reference system as a framework (Murphy 1993; Potgieter *et al.* 2003). However, in order to appropriately evaluate probabilistic forecast systems based on analogue years, better skill measures are required to appropriately account for the probabilistic nature of these systems (Potgieter *et al.* 2003).

DA is associated with variability of the observations among classes. For such forecast systems, there is no single, predicted value corresponding to each observation. Instead, the forecast consists of a set of possible values represented by empirical distributions functions (CDF) derived from previously observed values. The lack of clear distinction between sets of predicted and observed values and the probabilistic nature of those predictions needs to be taken into account when developing and applying skill measures.

Skill scores developed to quantify hindcast skill (e.g. LEPS skill score and RPSS) of probabilistic forecast systems that produce categorical forecasts (probabilities of belonging to predefined intervals or classes) are now in common use, in spite of their limitations. Class-based forecast systems do not readily lend themselves to such categorical evaluations without the loss of at least some valuable information by reducing the full probabilistic nature of the forecast systems to some broad bands of categories such as intervals defined by terciles (Potgieter *et al.* 2003). However, changes in agreement between observations and predicted probabilities for the predefined classes are directly related to DA, i.e. divergences between the empirical, conditional CDFs corresponding to each forecast system class and the unconditional CDF arising from 'climatology'. Hence, DA measures can be used as indirect skill measures for class-based forecast systems.

Inferential methods proposed here only quantify the degree of evidence against a null-hypothesis of either 'no DA' or 'no skill'. This information is essential but not sufficient for sound risk management. Decision makers also require complementary knowledge about the magnitude of expected change in the forecast variable. To quantify this magnitude requires descriptive measures (e.g. distance among cumulative distribution functions; magnitude of differences among conditional median or mean values etc) rather than inferential statistical methods. However, an in-depth evaluation and discussion of such descriptive measures is beyond the scope of this paper. The objective of this paper is to provide a generic, inferential framework for hypothesis testing that will add value to descriptive assessments of forecast quality.

The proposed inferential approach is based on distribution-free statistical methods that include both traditional non-parametric tests (e.g. Kolmogorov-Smirnov test) and computationally intensive methods based on non-parametric Monte Carlo techniques (e.g. bootstrapping and randomization tests). P-values derived from those distribution-free procedures are used to quantify evidences of 'true' DA and skill. This approach can be applied when the underlying probability distributions are unknown (it is not necessary to specify a particular distribution such as normal, gamma, etc), and it does not require any arbitrarily chosen level of significance. P-values range between 0 and 1 and are inversely proportional to the degree of evidence against the hypothesis of 'no class effect'. This approach takes into account the length of the time series, the number of classes of the chosen classification system and the intra-class variability. Further, given adequate spatial coverage, p-values can be mapped using interpolation methods, providing a powerful and intuitive means of communicating the spatial variability of DA and skill. We illustrate this approach by quantifying DA and skill of the 5-phase Southern Oscillation Index (SOI) classification applied to forecasting rainfall across Australia, Republic of South Africa and India.

2. Material and methods

We used 3-monthly rainfall totals from 3 sample stations (one each from Australia, The Republic of South Africa and India) and gridded (0.5 degree by 0.5 degree) rainfall data for each of these countries to demonstrate the use of p-values to measure aspects of discriminatory ability and skill of seasonal forecast systems. The 3 sample stations were Echuca (Australia), Bangalore (India) and Bloemfontein (The Republic of South Africa). Rainfall data for Echuca was obtained from the SILO Patched Point Dataset (PPD; Jeffrey et al, 2001), while Bangalore and Bloemfontein rainfall data are generated from GHCN Data (NOAA; Version 2: Peterson and Easterling 1994; Easterling and Peterson 1995; Easterling *et al.* 1996). Gridded data from the Hadley/CRU 0.5x0.5 global rainfall grid (New *et al.* 2000) was used for the spatial analyses of Australia, India and South Africa due to a lack of recent Indian rainfall records.

All 3 sample stations had at least 94 years of daily rainfall records. They represent vastly different climatic and agricultural regions.

The SOI 5-phase forecast system (SOI-5 FS) considered here is based on an analogue year approach and is underpinned by a sound, physical understanding of the ENSO cycle (Stone *et al.* 1996). The system has been used extensively for decision-making in Australia (e.g. Hammer et al. 2000) and elsewhere (e.g. Hill et al. 2000, 2004; Selvaraju et al. 2004). Years were categorised into five analogue sets according to their similarity regarding oceanic and/or atmospheric conditions as measured by SOI phases just prior to the 3-months forecast period. Hence, the rainfall time series were segregated into sub-series corresponding to each SOI class (consistently negative, consistently positive, falling, rising and neutral), resulting in 5 sub-series with variable record lengths. These rainfall time series were represented by their respective cumulative distribution functions (CDFs) or their complement, probability of exceedance functions (POEs): a conditional CDF (or POE) for each class and an unconditional CDF (or POE) for 'climatology'. Cumulative probabilities are a simple and convenient way to represent probabilistic information arising from a time series that exhibits no or only weak auto-correlation patterns. However, if the time series shows moderate to strong auto-correlation patterns, a CDF/POE summary will result in some loss of information. Yearly sequences of rainfall data from a specific month or period exhibit only weak auto-correlation, thus allowing the CDF/POE representation to convey seasonal climate forecast information (e.g. Selvaraju *et al.* 2004). Figure 1 provides an example of rainfall categorisation based on the SOI classes for the three sample locations.

a. Inferential statistical methods to quantify DA and skill

The proposed inferential framework can be applied in conjunction with any statistical test or skill measure. As an example we implemented the approach using:

(i) Two nonparametric statistical tests to quantify DA, namely the Kruskall-Wallis (KW) test for comparing medians and the multi-sample Kolmogorov-Smirnov (multi-sample KS) test for comparing CDFs (Conover 1980; Stone *et al.* 1996; Stone *et al.* 2000).

(ii) Randomization tests for quantifying evidences of skill as measured by two descriptive skill scores, LEPS (linear error in the probability space, LEPS score; Potts *et al.* 1996) and RPSS (ranked probability skill score; Epstein 1969), which is an operational forecast evaluation procedure used by the International Research Institute (Goddard *et al.* 2003).

The KW test is a generalization of the Wilcoxon-Mann-Whitney test applied to three or more groups (Stokes *et al.* 2000). It accounts for overall divergences among medians of conditional CDFs[2] and is the non-parametric test equivalent to the F-test used in analysis of variance. KW accounts for divergences among locations only, while the multi-sample KS test, based on maximum vertical distances among CDFs, takes a whole distribution approach to testing. The KS test is therefore able - unlike the KW test - to detect differences due to both spread and/or shape of distributions.

We used p-values associated with the observed KW and multi-sample KS statistics as evidences against the 'no discriminatory ability' null hypothesis. To demonstrate the universal applicability of this inferential framework, we have included the two popular skill measures LEPS and RPSS in spite of our reservations regarding their ability to adequately represent the probabilistic features of class-based forecast systems (see earlier). At the very least, the examples show how quantitative descriptive measures can be converted into inferential measures, thus providing a much sounder basis for comparative assessments.

The use of statistical tests to assess spatial variability of DA was proposed by Stone (1992), who produced contour maps of significance levels of KW tests applied to SOI grouping on rainfall medians over Australia. Our approach differs in one important aspect: we do not propose the use of any predetermined cut-off levels for evidence against the null hypothesis. Instead of choosing significance levels and verifying if DA or skill is significant or not, we provide the nominal significance levels (p-values). This avoids the loss of potentially valuable information and provides informed users with the opportunity to form their own opinion whether or not the evidence is sufficient to influence decision-making. Stone *et al.* (2000) has already used the KW p-values to quantify discriminatory ability of GCM-derived analogue forecast systems.

[2] Kruskal-Wallis is a nonparametric test for the null hypothesis that the distribution of an ordinally scaled response is the same in two or more independently sampled populations. It is sensitive to the alternative hypothesis that there is a location difference between at least a pair of populations (Stokes *et al.* 2000). It requires the assumption of same population variances when used for comparing distributions. In our case studies, we are using KW for comparing medians, thus homogeneity of variances is not required.

Both LEPS skill scores and RPSS quantify the departure between a categorical forecast and observations in the cumulative probability space (Zhang and Casey 2000). Here we used calculations for LEPS skill scores and RPSS, which are based on agreement between actual rainfall values and predicted probabilities for intervals defined by the empirical terciles of the corresponding cross-validated forecast distributions.

Every empirical, descriptive skill measure, including LEPS skill scores and RPSS, needs to be complimented by some measure of uncertainty before the information can be confidently applied in decision making (Potts *et al.* 1996; Zhang and Casey 2000; Jolliffe 2004). Beyond assessing the skill magnitude (observed skill score), it is critically important for users of forecast systems to know the probability of such skill arising by chance, in order to avoid making decisions based on artificial or perceived skill. This probability is used to assess the true class effect, considering the time series size (record length) and other sources of variability, not explained by the classification system used.

However, appropriate null-distributions for such assessments are not readily available when using the standard LEPS skill scores or RPSS. Zhang and Casey (2000) therefore proposed the construction of 'statistical distributions' using quasi-random experiments in order to assess the significance of forecast skill. They generated 95 'quasi-random ensembles' from the rainfall time series (1900-1995) at each station by shifting the observations 1 year ahead at a time and replacing, after each iteration, the first observation with the last. A single 'statistical distribution' was then derived for each skill measure, using the 95 skill values arising from each grid location in Australia. Those 'statistical distributions' were used as 'null distributions' for performing skill significance tests. However, those distributions are not appropriate for assessing significance or calculating p-values because they do not adequately represent the set of possible values of the skill measure under the hypothesis of 'no skill'. Further, existing spatial variability of skill is misinterpreted as 'quasi-random variation'. Combining skill values from different locations into one distribution ignores the existence of well-established, spatial correlation patterns. Hence, we propose a location-by-location assessment, which allows the construction of true null distributions of each skill measure at each location based on randomisation techniques (Monte-Carlo analyses).

For each location, we calculated p-values associated with LEPS skill scores and RPSS using Monte Carlo methods by randomly allocating all the observed rainfall data 5000 times to the five SOI sub-classes and calculating cross-validated skill score values for each random allocation in order to derive empirical null distributions for both skill measures. Such distributions represent the set of possible skill score values under the null hypothesis of 'no skill'. Considering the alternative hypothesis of skill score for forecast system > skill score for climatology, p-values associated with observed skill scores for both measures were calculated as the relative frequency of skill scores that exceeded the respective observed skill scores (see example for 3 locations, Fig. 2).

b. Spatial and temporal assessments of forecast quality

To demonstrate the usefulness of p-values for spatial analyses, we mapped the KW and LEPS skill score p-values[3] from all grid points for the periods analysed. This allowed us to examine spatial patterns of forecast quality associated with the chosen forecast system. For a temporal assessment of DA and skill we investigated month-by-month changes in p-values associated with KW, multi-sample KS, LEPS skill scores and RPSS at the three sample locations (Fig. 3).

[3] Mapping values arising from KS and RPSS yielded near-identical results and were therefore omitted.

3. Results and discussion

Here we presented a generic inferential framework for quantifying forecast quality attributes associated with probabilistic forecast systems, namely, skill and discriminatory ability. We discuss that the distinction between the two concepts becomes blurred, in the case of probabilistic forecast systems based on analogue years approach. We selected KW and KS to quantify DA because of their non-parametric nature that allows us to apply these statistical tests without the need for distributional assumptions. We selected LEPS and RPSS because they are two frequently used scoring systems (Mason 2004).

Skill scores associated with these systems are not very informative on their own, difficult to interpret and need to be accompanied by some measure of uncertainty to be useful (Hartmann *et al.* 2002; Potgieter *et al.* 2003; Jolliffe 2004). Any other statistical test or skill measure can be used with this generic inferential framework (e.g. for DA, Median or Log-Rank tests; for skill, measures such as Brier skill scores and many more, see Potgieter *et al.* 2003; Mason 2004). Here KW, multi-sample KS, LEPS and RPSS merely serve as examples to demonstrate the overall approach, the choice of the quality measure depends on the objective of the study. Our choice of LEPS or RPSS does not constitute an endorsement of these measures - they must be adequate for testing the hypothesis under investigation[4]. Using a statistical hypothesis testing approach, we converted observed skill scores into corresponding nominal significance levels (p-values). This accounts for differences in record length and number of classes and thus enables (i) investigation of temporal variability in forecast quality (e.g. length and timing of the 'autumn predictability barrier' of ENSO based forecast systems), (ii) objective comparisons of DA/skill among sites or assessments of spatial patterns of DA/skill over regions, (iii)

[4] When different statistical tests are available for the same hypothesis, a power analysis would provide objective criteria for choosing the most appropriate test. This could be achieved by analysing a large number of Monte Carlo samples (sets of conditional distributions) drawn from synthetically constructed distributions with known properties (divergences among conditional CDFs).

assessment of congruence of DA/skill magnitude as measured by different skill scores and (iv) performance evaluation of different forecast systems at a location and/or regionally.

a. Converting skill scores into p-values

In our examples, the relative location of the observed skill scores (Fig. 2, dark, thick line) on the skill scores' empirical null distributions indicates the degree of evidence against the hypothesis of 'no skill'. The higher the observed skill score value relative to the null-distribution, the greater the empirical evidence of true skill of the forecast system. For instance, the LEPS and RPSS skill score p-values of the SOI-5 FS to predict JAS Echuca rainfall were both < 0.001. This indicates highly significant and similar forecast skill, regardless of skill score used. The LEPS skill score (RPSS) p-values for JAS at Bangalore, 0.009 (0.027) and for NDJ at Bloemfontein 0.002 (< 0.001) also indicate highly significant and similar forecast skill (Fig. 3). The results show that the conversion of skill scores into p-values can overcome the issue raised by Mason (2004) of some skill measures, such as Brier and RPSS, having negative skill score values that can still be indicative of forecast system skill (as demonstrated by low p-values shown for Bangalore, Bloemfontein and Echuca). This goes some way towards overcoming the lack of equitability of some scoring systems also addressed by Mason (2004).

Goddard and Dilley (2005) commented that Monte Carlo re-sampling would not be appropriate to assess nominal significance levels (p-values) for RPSS because '*forecasts drawn at random relative to the observed sequence of years typically yield a RPSS worse than that of climatology*'. We disagree. As explained by Mason (2004), such negative scores are an inherent feature of RPSS and negative scores can occur even when true forecast skill exists. As the expected value of RPSS under the null hypothesis is influenced by the forecast system employed, empirical null distributions generated using Monte Carlo techniques can contain a high frequency of negative values (Mason 2004) as shown in Fig 2. Such a negative bias of the RPSS expected value (in contrast to LEPS) does therefore not invalidate the use of Monte Carlo techniques for establishing p-values associated with skill scores.

In this context, we need to flag the issue of how such null distributions can be constructed. For class-based forecast systems using conditional distributions, random reallocation of climate data to these classes of conditional probabilities is the intuitively obvious method. A truly probabilistic assessment of GCM-based forecasts is more difficult to obtain. Allen and Stainforth (2002) criticise the 'probabilistic' outputs generated by GCMs through altering initial and boundary conditions without explicitly accounting for the climate's response. They argue that climate forecasts are intrinsically five-dimensional, spanning space, time and probability, a fact not accounted for when compiling subjective GCM-based probability distribution of forecasts. More attention to formal uncertainty analyses is required, including much more rigorous sensitivity testing based on many more elaborate ensemble runs, before reliable GCM-based probabilistic trajectories of future climate states can be provided (Meinke *et al.* 2004). For now, the methods suggested by Stone *et al.* (2000) provide a way to develop class-based forecast systems from GCM outputs, allowing an immediate application of the generic inferential framework we developed here.

b. Quantifying temporal patterns of forecast quality attributes

Particularly with ENSO-based forecast system, forecast quality attributes will vary temporally due to the well-known, seasonal life cycle of the ENSO phenomenon. Therefore, location-specific temporal analyses are necessary to evaluate when the forecast system is sufficiently informative to influence decision making. Here we show the usefulness of the generic inferential framework to simultaneously assess temporal patterns. These tests revealed some interesting insights regarding predictability and dynamics of seasonal rainfall patterns that should be investigated in more detail elsewhere (Fig. 3):

(i) Regardless of test or skill score, p-values at all locations followed similar time courses, albeit with some exceptions;

(ii) The autumn predictability barrier (Clarke and Shu 2000; Clarke and van Gorder 2003) around MAM is clearly evident at all locations.

(iii) For Echuca the typical ENSO-lifecycle is evident, with a strong impact on winter and spring rainfall;

(iv) Bloemfontein, a region that is seasonally dry in winter, shows ENSO impact for the beginning of the rainy season around October, while Bangalore shows a strong impact for the (northern) summer monsoon (MJJ to ASO) and a small peak for the much weaker winter monsoon (OND).

At the 3 locations, all measures broadly identified similar trends, but differed in detail. The fact that there is broad congruence between p-values regardless of test or measure employed demonstrates our earlier assertion that discriminatory ability can be used as a surrogate for skill for class-based, probabilistic forecasts. Further, it shows the generic inferential framework's ability to reconcile vastly different measures (Fig 3).

Although discriminatory ability and skill tests generally follow fairly consistent patterns (temporally as well as spatially; Figures 3 & 4), there may be situations when results may differ substantially (e.g. JJA rainfall for Echuca; Figure 3). Results where the p-values from discriminatory ability tests (such as KW and the multi-sample KS) are much smaller than the p-values from skill tests can be attributed to procedural differences between the two types of tests. KW/KS test for differences in at least 1 phase (or class) distribution. There will be instances when a single phase distribution is significantly different from all the other distributions, which do not differ from each other, resulting in a low p-value. For a skill test, this is unlikely to yield a low p-value as such tests are designed to compare observed data with forecast data. Even though these circumstances will arise occasionally, Figures 3 and 4 show that p-values are generally very consistent between tests, broadly indicating the same temporal and spatial trends. This is particularly the case when p-values are low, indicating a convergence of results when real differences between distributions exist. Generally discriminatory ability tests seem to be slightly more emphatic, but our results show that they serve as reliable proxies for skill measures when evaluating class-based forecast systems. Here we would like to add a cautionary note: there is a temptation to 'over-interpret' temporal patterns in p-values as presented in Fig. 3. By definition, p-values indicate the evidence against the null-hypothesis – a value of 0.2 means that in one out of five cases we would falsely reject the null-hypothesis. While the broad temporal patters from all tests are similar, differences are inevitable and it is probably not helpful to discuss whether or not p-values of 0.3 versus 0.8 constitute evidence of 'skill' or otherwise (cf. Echuca, JJA).

It is up to the informed user to decide the appropriate level of 'significance' (i.e. the degree of evidence against the hypothesis of 'no class effect') before using the information in decision making (Nicholls 2001). We also note that Nicholls (2001) argues against 'blanking out ' of areas on maps where some feature does not reach statistical significance, a practice often seen in atmospheric science. This practice can lead to the loss of potentially valuable information. Therefore, we argue against the use of any artificial cut-off levels to determine whether or not the p-values of the tests indicate sufficiently high evidence. Instead, we provide all nominal significance levels (p-values) and concur with Nicholls (2001), who questions the appropriateness of commonly used cut-off levels, such as p<0.05 or p<0.01. These cut-offs are no more than a convention that reduce continuous probabilistic information to a dichotomous response, thereby ignoring valuable information contained in the nominal significance levels. Rosnell and Rosenthal (1989), cited by Nicholls (2001), noted that '*...surely, God loves the .06 nearly as much as the .05*'.

In our examples, using traditional cut-off levels for significance testing would result in conflicting conclusions for some sites and seasons. For instance, at Bangalore, India, the JAS ENSO signal would be considered either as being 'significant' or 'not significant' if a 5% cut-off was adopted, depending on the chosen test or measure. For JAS p-values ranged from 0.09 (KW) to 0.01 (LEPS). This is in spite of a strong and well-established ENSO impact at this location and the fact that p-values for all tests are moderate to low, but not all are below 0.05 (Fig. 3). Should risk managers in the Bangalore region ignore ENSO-based forecasts during this season? Alternatively, should they base their decisions on a single measure using a pre-determined cut-off for significance? Risk managers must decide for themselves whether or not the evidence is strong enough to influence their decisions.

We hypothesise that differences among p-values coming from different measures *(ceteris paribus)* might be caused by differences in the ability of each test or measure to detect divergences among conditional distributions regarding the target attribute (e.g. median, variance, whole CDF). For example, for time series that contain more than 50% of rainfall values equalling zero in all classes (all class medians are zero) p-value arising from a median test would yield a value of one, while a test comparing conditional CDFs could produce a low p-value, depending on the differences in the right tails of conditional CDFs. Therefore, lack of agreement between tests or measures can sometimes be explained by differences in the underlying hypotheses tested or existing power differences between tests. Understanding the differences between skill and or DA tests is important and although these differences have not been addressed within this paper, it is the subject of ongoing research. Based on our research here we argue that non-parametric DA measures such as KW and KS are in most cases adequate surrogates for skill measures and little if any additional information can be gained from using skill measures originally designed for deterministic forecasts.

c. Quantifying spatial patterns of DA and skill over a region

As we have shown, quality measures of forecast systems vary temporally and spatially. The spatial patterns of DA and skill for the SOI-5 FS based on KW and LEPS p-values (Fig. 4) were consistent with known, typical ENSO impacts. Again, DA and skill measures showed similar spatial patterns, regardless of season - high divergence among conditional CDFs is highly likely to lead to improved agreement between `predicted' and `observed' values as captured by a measure of skill. However, high DA does not necessarily imply high skill (Stone *et al.* 2000) due to, for instance, possible changes in skill over time, a factor not accounted for when calculating DA, but considered during the cross-validation procedure when calculating skill. Hence, it is not unexpected that there appears to be a general tendency for p-values associated with DA to be slightly lower than those associated with skill.

Parametric approaches were initially developed during a time when computer power was not available. Due to their reliance on distributional assumptions they are a convenient way to quickly perform hypothesis tests and calculate associated nominal significance levels (p-values) using known, analytically-derived null distributions. Initially, most of the known parametric methods were based on the assumption of normality. Nowadays there are increasing numbers of parametric methods available that are based on a range of distribution types, such as Tweedie family (e.g. Tweedie 1984; Jφrgensen 1987), which include the Normal, gamma, and

Poisson distributions as special cases and more. Although this greatly broadens the applicability of parametric approaches, spatial assessments of forecast quality still require case-by-case evaluation before parametric methods can be applied. The historical limitation imposed by the lack of computer power no longer holds and it is therefore no longer necessary to make assumptions about distributions – these can now be constructed via non-parametric Monte Carlo approaches, such as bootstrapping and randomisation techniques. This flexible approach is of particular importance for climate science where data sources are varied, underlying distributions can come in many shapes and predictor/predictand relationships are often non-linear (Von Storch and Zwiers 1999).

4. Conclusions

Using an inferential, non-parametric framework we have evaluated aspects of forecast quality using 3-monthly rainfall forecasts for a range of locations in Australia, The Republic of South Africa and India. The approach taken is generic and independent of location, season, data source, statistical test or skill score and provides intuitively simple, but powerful methods that objectively quantify discriminatory ability and skill of probabilistic forecast systems. Forecast quality measures, once converted to nominal significance levels (p-values), can provide the means for investigating temporal and spatial patterns of discriminatory ability and skill. In addition, this allows comparisons among different probabilistic forecast systems according to objective quality criteria – a key issue to further improve risk management in climate-sensitive agricultural systems. In a subsequent step, this framework should be supplemented with descriptive statistical tools that quantify the magnitude of difference between target forecast quantities derived from forecast probability distributions, in addition to the evidence against the null-hypothesis provided by p-values. From a risk management perspective, the ultimate responsibility for the utility of these tools resides with the decision makers. They must be able to properly interpret the relevance of information obtained using these approaches. This requires an ability to handle intrinsically uncertain information (probabilities) as well as asking the relevant questions (Hoffman and Kaplan 1999).

The most insightful results were obtained when two very different skill scoring systems, namely LEPS and RPSS converged in terms of p-values and the resulting evidence against the null-hypothesis was similar for both scoring systems and, indeed, even for all four DA and skill measures. This provides strong supporting evidence of the general applicability of this generic, inferential framework to explore and quantify forecast quality. We concur with comments made by Zhang and Casey (2000), Potgieter *et al.* (2003) and Mason (2004), who all stated the need to consider a set of measures due to the multidimensional nature of forecast quality.

Finally, the generic inferential framework proposed here might also be useful for evaluating similarities among different sets of quality measures: the approach proposed by Potgieter *et al.* (2003) based on Principal Component Analysis and clustering techniques might reveal more insights when applied to p-values arising from those measures. Further, the value of adapting skill measures based on tercile-based categories in order to provide continuous assessments of forecast quality can now be objectively evaluated. All these issues are subjects of on-going research.

Acknowledgements

This research was financially supported by Embrapa Meio Ambiente, São Paulo, Brazil, the Department of Primary Industries and Fisheries, Queensland, Australia and the Australian Grains Research and Development Corporation (GRDC). We wish to thank Ian Jolliffe and Andries Potgieter for their helpful comments on the manuscript. We also thank an anonymous reviewer for a very comprehensive review that has considerably improved the quality of our manuscript.

References

Allen, M. R., and D. A. Stainforth, 2002: Towards objective probabilistic climate forecasting. *Nature*, **419**, 228.

Clarke, A.J. and L. Shu, 2000: Biennial winds in the far western equatorial Pacific phase-locking El Niño to the seasonal cycle, *Geophys. Res. Lett.*, **27(6)**, 771-774.

Clarke, A.J. and S. van Gorder, 2003: Improving El Niño prediction using a space-time integration of Indo-Pacific winds and equatorial Pacific upper ocean heat content. *Geophys. Res. Lett.*, **30(7)**, 1399-1402, doi:10-1029/2002GLO16673.

Conover, W. J., 1980: *Practical Nonparametric Statistics*. 3d ed. John Wiley & Sons, 584 pp.

De Jager, J. M., A. B. Potgieter and W. J. van der Berg, 1998: Framework for forecasting the extent and severity of drought in maize in the Free State province of South Africa. *Agricultural Systems,* **57**, 351-365.

Easterling, David R., Thomas C. Peterson, and Thomas R. Karl, 1996: On the development and use of homogenized climate data sets. *Journal of Climate,* **9**, 1429-1434.

Easterling, David R. and Thomas C. Peterson, 1995: A new method for detecting and adjusting for undocumented discontinuities in climatological time series. *International Journal of Climatology*, **15**, 369-377.

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

Ferreyra, R.A., G. P. Podestá, C. D. Messina, D. Letson, J. Dardanelli, E. Guevara, and S. Meira, 2001: A linked-modeling framework to estimate maize production risk associated with ENSO-related climate variability in Argentina. *Agricultural and Forest Meteorology*, **107**, 177–192.

Goddard, L. and M. Dilley, 2005: El Niño: Catastrophe or opportunity. *J. Climate*, **18**, 651-665.

Goddard, L., A. G. Barnston, and S. J. Mason, 2003: Evaluation of the IRI's "Net Assessment" seasonal climate forecasts 1997–2001. *Bull. Amer. Meteor. Soc.*, **84**, 1761-1781.

Hammer, G.L., N. Nicholls, C. Mitchell, 2000 (eds): *Applications of seasonal climate forecasting in agriculture and natural ecosystems: The Australian experience.* Kluwer Academic Publishers, 469 pp.

Hartmann, H.C., T.C. Pagano, S. Sorooshian, and R. Bales, 2002: Confidence builders: evaluating seasonal climate forecasts from a user perspective. *Bull. Amer. Meteor. Soc.*, **83**, 683-698.

Hill, S. J. H., J. Park, J.W. Mjelde, W. Rosenthal, H. A. Love, and S. W. Fuller, 2000: Comparing the value of Southern Oscillation Index-based climate forecast methods for Canadian and US wheat producers. *Agric. Forest Met.*, **100,** 261-272.

Hill, H. S. J., J. W. Mjelde, H. A. Love, D. J. Rubas, S. W. Fuller, W. Rosenthal, and G. Hammer, 2004: Implications of seasonal climate forecasts on world wheat trade: a stochastic, dynamic analysis. *Canadian Journal of Agricultural Economics*, **52**(3), 289-312.

Hoffman, F.O. and S. Kaplan, 1999: Beyond the Domain of Direct Observation: How to Specify a Probability Distribution that Represents the "State of Knowledge" About Uncertain Inputs. *Risk Analysis*, **19(1)**, 131-134.

Jeffrey, S.J., J. O. Carter, K. M. Moodie, and A. R. Beswick, 2001: Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling and Software*, **16/4**, 309-330. [Available at http://www.nrm.qld.gov.au/silo/ppd/index.html].

Jolliffe, I.T., 2004: Estimation of uncertainty in verification measures. *Proc. International Verification Methods Workshop*, Montreal, Canada. [Available at http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Workshop2004/home.html]

Jφrgensen, B., 1987: Exponential dispersion models. *Journal of the Royal Statistical Society, Series B*, **49**, 127-162.

Maia, A. H. N., H. Meinke and S. Lennox, 2004: Assessment of probabilistic forecast 'skill' using p-values. *Proc. 4th International Crop Science Congress*, Brisbane, Australia.

Manly, B.F., 1981: *Randomization tests and Monte Carlo methods in biology.* John Willey & Sons.

Mason, S. J., 2004: On using 'climatology' as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **132**, 1891-1895.

Meinke, H. and Z. Hochman, 2000: Using seasonal climate forecasts to manage dryland crops in northern Australia. In: G.L. Hammer, N. Nicholls, and C. Mitchell (eds.), Applications of Seasonal Climate Forecasting in Agricultural and Natural Ecosystems – The Australian Experience. Kluwer Academic, The Netherlands, p. 149-165.

Meinke H. and R. C. Stone, 2005: Seasonal and inter-annual climate forecasting: the new tool for increasing preparedness to climate variability and change in agricultural planning and operations. *Climatic Change*, **70**, 221-253.

Meinke, H., L. Donald, P. deVoil, B. Power, W. Baethgen, M. Howden, R. Allan, and B. Bates, 2004: How predictable is the climate and how can we use it in managing cropping risks? Invited Symposium Paper, *Proc. of the 4th International Crop Science Congress*, 26 Sep – 1 Oct 2004, Brisbane, Australia, CD-ROM. [Available online at www.cropscience.org.au]

Messina, C.D., J. W. Hansen, and A. J. Hall, 1999: Land allocation conditioned on ENSO phases in the Pampas of Argentina. *Ag. Systems*, **60**, 197-212.

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281-293.

Nelson, R.A., D. P. Holzworth, G. L. Hammer, and P. T. Hayman, 2002: Infusing the use of seasonal climate forecasting into crop management practice in North East Australia using discussion support software. *Ag. Systems*, **74**, 393-414.

New, M. G., M. Hulme and P. D. Jones, 2000: Representing twentieth-century space-time climate variability. Part II: Development of 1901-1996 monthly grids of terrestrial surface climate. *J. Climate*, **13**, 2217-2238.

Nicholls, N., 2001. The insignificance of significance testing. *Bull. Amer. Meteor. Soc.*, **82**, 981-986.

Peterson, Thomas C. and David R. Easterling, 1994: Creation of homogeneous composite climatological reference series. *International Journal of Climatology*, **14**, 671-679.

Potgieter, A.B., Y. Everingham, and G. L. Hammer, 2003: Measuring quality of a commodity forecasting from a system that incorporates seasonal climate forecasts. *Int. J. Remote Sensing*, **23**, 1195-1210.

Potgieter, A.B., G. L. Hammer, H. Meinke, R. C. Stone, and L. Goddard, 2005: Three putative types of El Nino revealed by spatial variability in impact on Australian wheat yield. *Journal of Climate*, **18**, 1566-1574.

Potts, J.M., C. K. Folland, I. T. Jolliffe, and D. Sexton, 1996: Revised "LEPS" scores for assessing climate model simulations and long-range forecasts. *J. Climate*, **9**, 34-53.

Podestá, G., D. Letson, C. Messina, F. Rocye, A. Ferreyra, J. Jones, I. Llovet, J. Hansen, M. Grondona, and J. O'Brien, 2002: Use of ENSO-related climate information in agricultural decision making in Argentina: a pilot experience. *Ag. Systems*, **74**, 371-392.

Rosnell, R. L., and R. Rosenthal, 1989: Statistical procedures and the justification of knowledge and psychological science. *Amer. Psychol.*, **44**, 1276-1284.

Selvaraju, R., H. Meinke, and J. Hansen, 2004: Climate information contributes to better water management of irrigated cropping systems in Southern India. *Proc 4th International Crop Science Congress*, Brisbane, Australia.

Singels, A. and A. B. Potgieter, 1997: A technique to evaluate ENSO-based maize production strategies. *South African Journal of Plant and Soil*, **14**, 93-97.

Sivakumar, M.V.K., R. Gommes, and W. Baier, 2000: Agrometeorology and sustainable agriculture, *Agric. Forest Met*, **103**, 11-26.

Stokes, M. E., C. S. David, and G. G. Koch, 2000: *Categorical data analysis using the SAS® System*. 2nd ed. SAS Institute Inc., p. 165.

Stone, R. C., 1992: SOI phase relationship with rainfall probabilities over Australia. Unpublished PhD Thesis, University of Queensland, Brisbane, Australia.

Stone, R. C., G. L. Hammer, T. Marcussen, 1996: Prediction of global rainfall probabilities using phases of the Southern Oscillation Index. *Nature*, **384**, 252-55.

Stone, R. C., I. Smith, and P. McIntosh, 2000: Statistical methods for deriving seasonal climate forecasts from GCM´s. In: *Applications of seasonal climate forecasting in agricultural and natural ecosystems*. Hammer, G.L., N. Nichols, C. Mitchell, Eds., Kluwer Academic Publishers, 135-147.

Tweedie, M. C. K., 1984: An index which distinguishes between some important exponential families. In 'Statistics Applications and New Directions', *Proc. of the Indian Statistical Institute Golden Jubilee International Conference*, Indian Statistical Institute, Calcutta, J. K. Ghosh and J. Roy (eds), 579-604.

Von Storch, H. and F.W. Zwiers, 1999: *Statistical Analysis in Climate Research*, Cambridge University Press, 484 pp.

Wilks, Daniel S., 1995: Statistical Methods in the Atmospheric Sciences, Academic Press, 467 pp.

WMO, 2005: Proceedings of the meeting of the CCl OPAG3 expert team on verification. [Available on line at: http://www.wmo.ch/web/wcp/clips2001/html/Report_ETVerification_080205_2.pdf]

Zhang, H. and T. Casey, 2000: Verification of Categorical Forecasts. *Weather and Forecasting*, **15**, 80-89.

Figure Captions

Figure 1: Time series and probability of exceedance plots for JAS rainfall by May/June SOI phase at (a) Echuca, Victoria (36.17°S, 144.76°E), and (b) Bangalore, India (13.00°N,77.60°E) and for NDJ rainfall by September/October SOI phase at (c) Bloemfontein, Republic of South Africa (29.10°S,26.30°E).

Figure 2: Empirical null distribution for the three-category LEPS skill scores (top row) and RPSS (bottom row) arising from the SOI forecast system for predicting (from left to right) JAS rainfall at Echuca and Bangalore, and NDJ rainfall at Bloemfontein. The LEPS skill scores are defined on the range −1 to +1 and the RPSS are −Inf to +1. Dark, thick lines indicate the location of the observed skill score values. The area to the right of the dark, thick lines correspond to the respective p-value.

Figure 3: Annual patterns of discriminatory ability and skill for Echuca, Bangalore and Bloemfontein based on p-values derived from the Kruskal-Wallis (KW) and Kolmogorov-Smirnov (multi-sample KS), cross-validated RPSS and cross-validated LEPS skill scores.

Figure 4: Discriminatory ability (DA, top row) and skill (bottom row) of the SOI-5 FS based on p-values derived from the Kruskal-Wallis test and cross-validated LEPS skill scores for July-September (JJA) rainfall in Australia and India, and November-January (NDJ) rainfall in the Republic of South Africa. For computational reasons, grids that have a large proportion (>33%) of dry seasons are removed from the LEPS analysis and therefore appear as white grids in the LEPS maps.

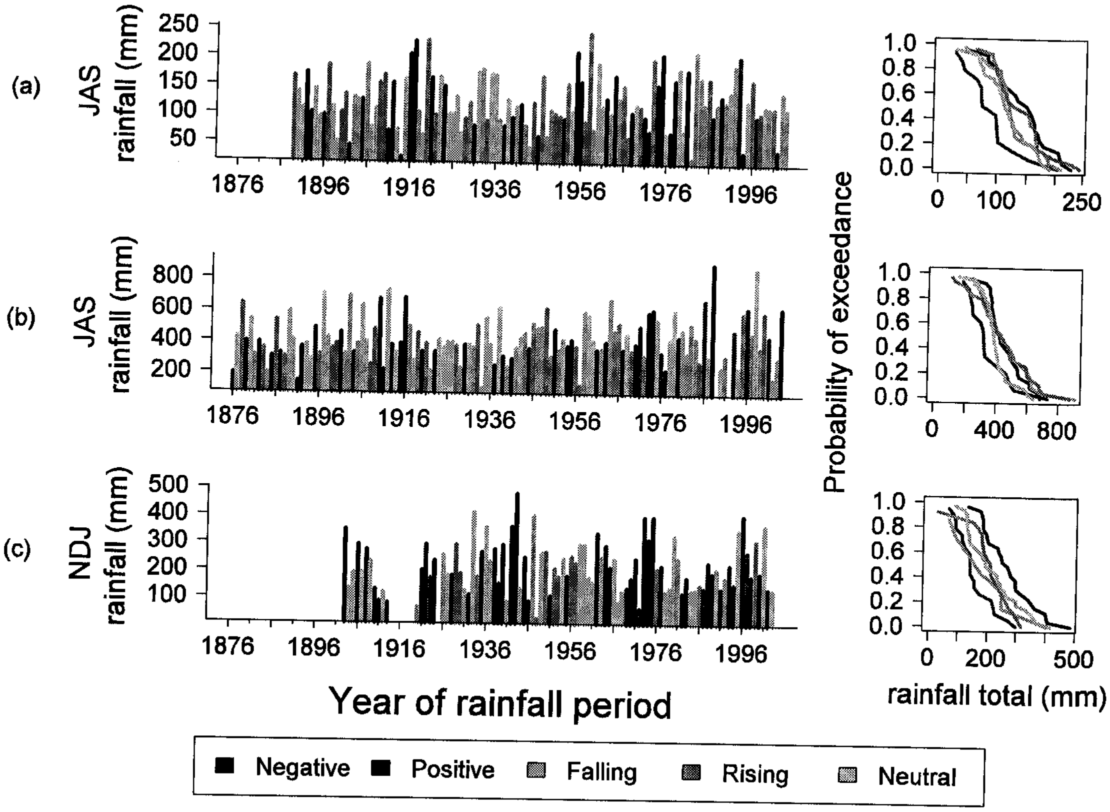

Figure 1: Time series and probability of exceedance plots for JAS rainfall by May/June SOI phase at (a) Echuca, Victoria (36.17°S, 144.76°E), and (b) Bangalore, India (13.00°N,77.60°E) and for NDJ rainfall by September/October SOI phase at (c) Bloemfontein, Republic of South Africa (29.10°S,26.30°E).

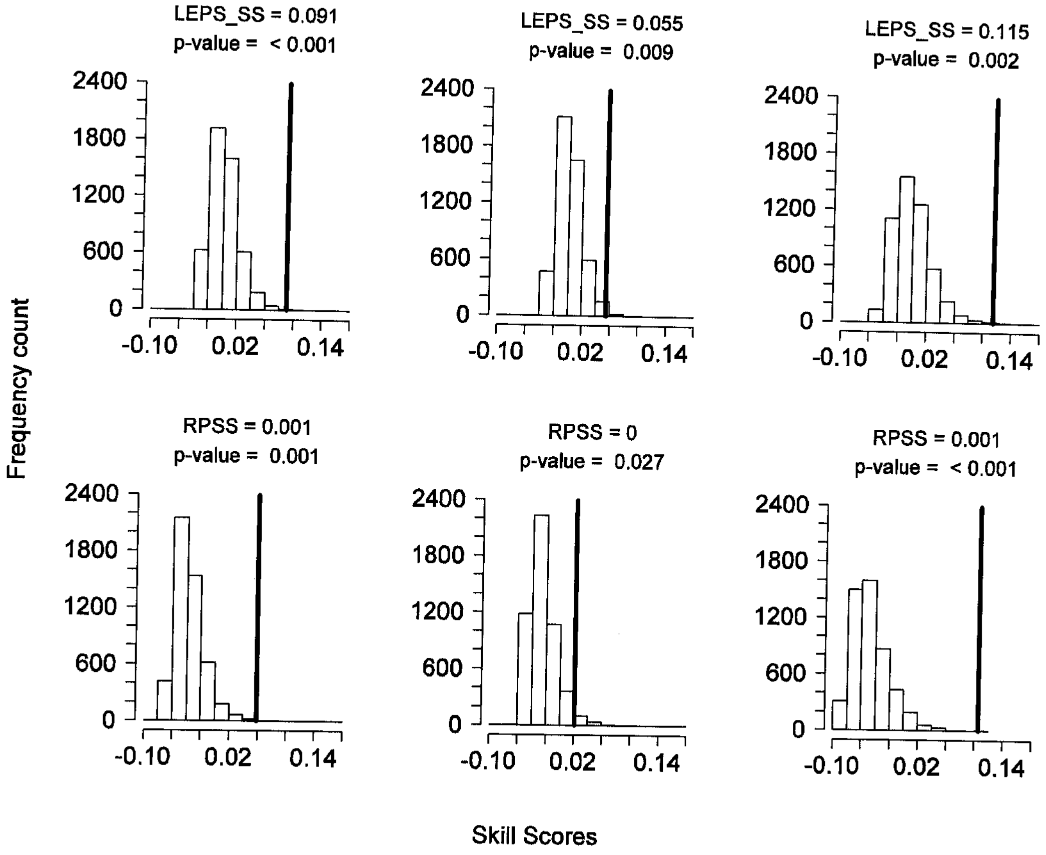

Figure 2: Empirical null distribution for the three-category LEPS skill scores (top row) and RPSS (bottom row) arising from the SOI forecast system for predicting (from left to right) JAS rainfall at Echuca and Bangalore, and NDJ rainfall at Bloemfontein. The LEPS skill scores are defined on the range –1 to +1 and the RPSS are –Inf to +1. Dark, thick lines indicate the location of the observed skill score values. The area to the right of the dark, thick lines correspond to the respective p-value.

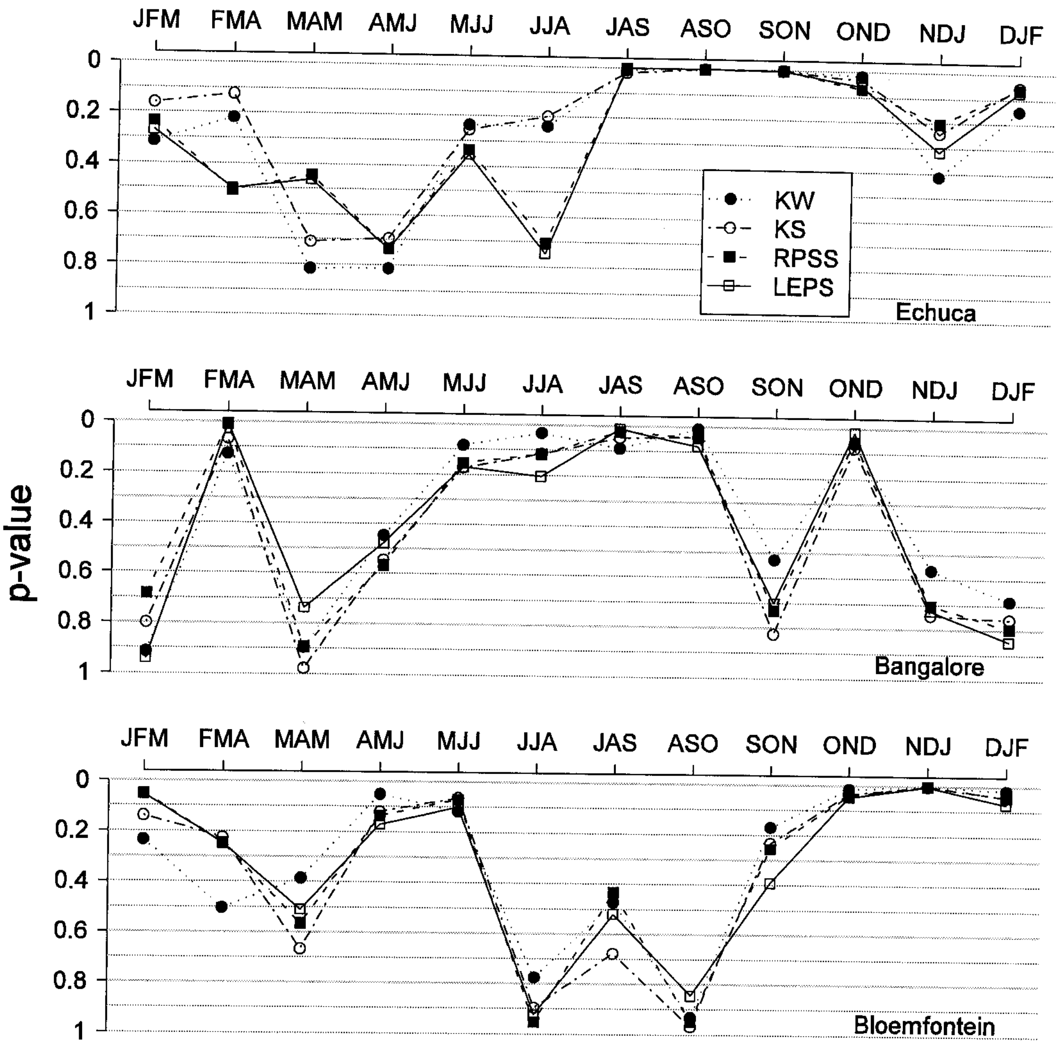

Figure 3: Annual patterns of discriminatory ability and skill for Echuca, Bangalore and Bloemfontein based on p-values derived from the Kruskal-Wallis (KW) and Kolmogorov-Smirnov (multi-sample KS), cross-validated RPSS and cross-validated LEPS skill scores.

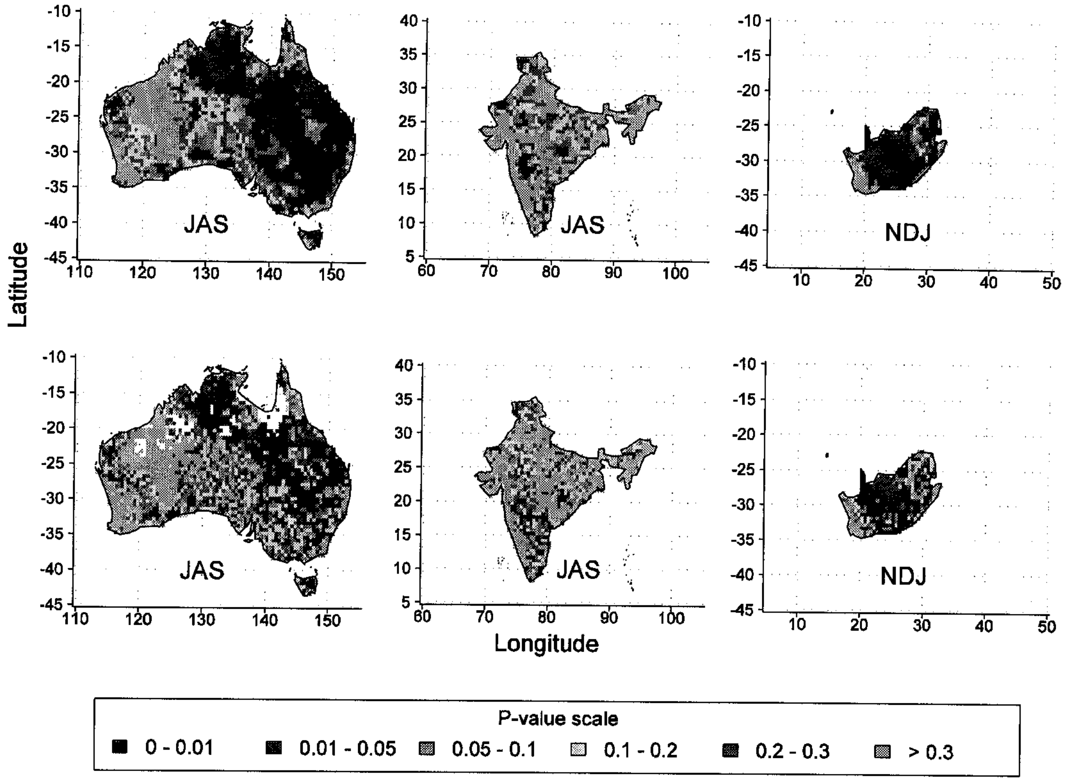

Figure 4: Discriminatory ability (DA, top row) and skill (bottom row) of the SOI-5 FS based on p-values derived from the Kruskal-Wallis test and cross-validated LEPS skill scores for July-September (JAS) rainfall in Australia and India, and November-January (NDJ) rainfall in the Republic of South Africa. For computational reasons, grids that have a large proportion (>33%) of dry seasons are removed from the LEPS analysis and therefore appear as white grids in the LEPS maps.