P-VALUES AS ANALYTICAL TOOLS IN PROBABILISTIC FORECAST ASSESSMENTS

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

ABSTRACT - Much has been written about 'quality' of probabilistic forecasts. Often, providers and users of such forecasts are unclear about what 'quality' entails, leading to confusion and misinformation. Here we try to provide some guidance and suggest a general approach to communicate aspects of probabilistic forecast quality related to discriminatory ability (*DA*) and skill (*S*). In our opinion, these two components of forecast quality should be considered independently. *DA* represents the additional knowledge about future states arising from some forecast system (FS) over and above the total variability of the prognostic variable while *S* quantifies changes in the agreement between observed and predicted values when using a specific FS instead of a FS based on "climatology" only. The major concerns are: generally poor distinction between *DA* and *S*; inappropriate use of significance testing to quantify *DA* and use of *DA* and *S* measures that do not account for the series lengths and/or number of classes of the *FS*. To address all of these issues, we propose the use of p-values derived from non-parametric tests as direct measures of *DA* and *S*. We illustrate this approach by quantifying *DA* and *S* of the Southern Oscillation Index applied to forecasting rainfall across Australia.

INTRODUCTION

Probabilistic seasonal climate forecasting in combination with simulation models of farming systems are powerful tools for rural risk assessments and the evaluation of alternative management options. A simple and intuitive way of connecting climate forecasts with such models is an 'analogue year' approach, whereby historical climate series are segregated into 'year or season types' resulting in sub-series, strata or classes, represented by their respective conditional distributions or probability of exceeding functions (POES). Those classes are derived from climate indicators such as the Southern Oscillation Index (SOI), El Niño/ Southern Oscillation (ENSO) phases or Sea Surface Temperature (SST) phases (Meinke and Stone 2005). However, methods that quantify the quality of such probabilistic forecast systems are poorly understood and often misused. Here we focus on two aspects related to climate forecast quality: discriminatory ability and skill (see definition in abstract). High discriminatory ability does not necessarily imply high skill and their respective quantifications require different statistical procedures (Stone et al., 2000). We therefore propose a simple, general framework to quantify both of these quality attributes that is based on p-values associated with non-parametrical statistical procedures. P-values range between 0 and 1 and are inversely proportional to the degree of evidence against the hypothesis of no-class effect. Thus, lower p-values indicate higher magnitude of the system attribute (*DA* or *S*). The magnitude of evidence takes into account the length of the series and the number of classes of the chosen classification system (Maia et al., 2004).

MATERIAL AND METHODS

We used aggregated, long-term June-August (JJA) rainfall from 590 stations across Australia to demonstrate the utility and adequacy of p-values as measures of discriminatory ability (case study I) and skill (case study II). The 5-phase SOI forecast system considered here is based on an 'analogue year' approach (Stone et. al, 1996; Meinke and Stone, 2005). Years were categorised into analogue sets according to their similarity regarding oceanic and/or atmospheric conditions as measured by SOI phases just prior to the 3-months period. Hence, the 590 JJA rainfall series (unconditional POEs) were segregated into five sub-series (conditional POEs) according to this SOI phase classification, resulting in 2950 sub-series with variable record lengths.

Case study I. Discriminatory ability of a 5-phase SOI forecast system over Australia.

For a spatial analysis of DA, we used long-term rainfall data from all 590 stations. The Kruskal-Wallis (KW) non-parametric test (Conover, 1980) was performed in order to quantify the magnitude of the SOI classification effect on medians of the JJA rainfall conditional POEs. P-values associated with this test are a direct measure of discriminatory ability. Hence, we mapped the KW p-values in order to show spatial patterns of the discriminatory ability for JJA rainfall over Australia arising from the SOI phase system.

Case study II. P-values associated to LEPS skill score.

To demonstrate how p-values can be used to quantify the skill of forecast systems, we calculated the *LEPS* skill score (*LEPS_SS*) for the SOI phase system at one location (Dalby; 1889-2003). *LEPS_SS* was chosen as an example for a widely used skill measure in climatological research (Potts et. al, 1996). That measure is based on LEPS (linear error in the probability space) and was applied to quantify the agreement between Dalby JJA observed and predicted values for the SOI forecast system relative to the agreement between those values for the forecast system based on "climatology" only.

Hence, we calculated the *LEPS*_SS (range:-100 to +100) based on categories defined by rainfall terciles for the observed Dalby JJA data set (observed *LEPS_SS*) as outlined by Potts et al. (1996). Using randomisation techniques (Manly, 1981) we generated 5000 random allocations of the observed JJA rainfall amounts to five 'synthetic' SOI classes. *LEPS_SS* were then calculated for each random allocation, allowing us to derive an empirical null distribution for that skill measure. The *LEPS_SS* null distribution represents the set of possible *LEPS_SS* values under the hypothesis of 'no SOI classification effect' for this dataset. The p-value associated with the original *LEPS* skill score (observed *LEPS_SS*) was calculated as the relative frequency of LEPS skill scores that exceed observed *LEPS_SS*.

[1] Embrapa Meio Ambiente, P.O. Box 69, Jaguariúna, SP, Brazil, CEP 13820-000 (ahmaia@cnpma.embrapa.br)
[2] Queensland Department of Primary Industries and Fisheries, P.O. Box 102, Toowoomba, Australia, Qld 4350.

RESULTS AND DISCUSSION

Discriminatory ability

The spatial pattern of DA for the 5-phase SOI system based on KW p-values was consistent with typical ENSO impacts across Australia (Fig. 1). It shows strong impact of ENSO on winter rain for Southern and Eastern Australia with weaker and less consistent DA for Western Australia (Northern Australia is seasonally dry at this time of the year).

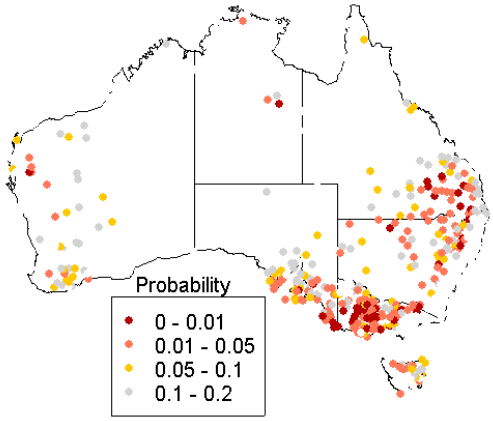

Figure 1. Discriminatory ability (*DA)* of 5-phase SOI system for JJA rainfall across Australia, as measured by Kruskal-Wallis p-values.

Non-parametric tests for comparing conditional probability distributions (e.g. Kolmogorov-Smirnov, Log-Rank) rather than a comparison of medians as in our case study, could also be used to quantify discriminatory ability.

Skill

LEPS_SS, like any other empirical measure, requires uncertainty assessments in order to adequately quantifying the magnitude of forecast systems skill (Jolliffe, 2004). Beyond assessing the skill magnitude as a point estimate (observed *LEPS_SS*), it is important for users of forecast systems to know the probability of exceeding the observed skill score 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 current classification system (intra-class variability).

Using a statistical hypothesis testing framework, we converted observed *LEPS_SS* into corresponding nominal significance levels (p-values), allowing us to compare skill arising from a particular classification system, regardless of differences in record length and intra-class variability. This enables objective comparisons of forecast systems skill among sites or assessments of spatial patterns of forecast skill over regions. The same approach applies to other skill measures.

The relative location of the observed *LEPS_SS* (Fig.2, red line) on the *LEPS_SS* empirical null distribution indicates the magnitude of skill of the probabilistic forecast system being evaluated. The higher the observed *LEPS_SS* value, the greater the empirical evidence of true skill of the forecast system. In our example, the p-value of the 5-phase SOI system to predict JJA Dalby rainfall was 0.0014 (*LEPS_SS* = 9.6), indicating highly significant forecast skill.

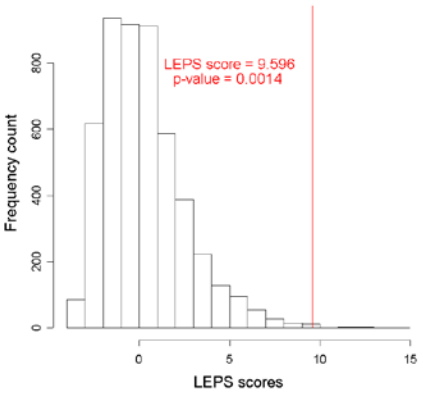

Figure 2. Empirical null distribution for tercile LEPS skill scores based on the 5-phase SOI forecast system applied to Dalby (SE Qld, Australia) JJA rainfall (bars), compared to *LEPSobs* (red line).

Here we demonstrated an intuitively simple, but powerful method to objectively quantify discriminatory ability and skill of probabilistic forecast systems. Forecast quality measures based on p-values can also provide the means to compare different probabilistic forecast systems according to objective quality criteria – a key issue to further improve risk management in climate-sensitive agricultural systems.

REFERENCES

Conover, W. J. *Practical Nonparametric Statistics*, John Wiley & Sons, New York, 1980.

Jolliffe, I. T. Estimation of uncertainty in verification measures. Proceedings of the International Verification Methods Workshop, Montreal, Canada, 2004.

Maia, A. H. N., Meinke, H., Lennox, S. Assessment of probabilistic forecast 'skill' using p-values. *Proceedings of the 4th ICSC*, Brisbane, Australia, 2004.

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

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

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

Stone, R. C., Smith, I., McIntosh, P. Statistical methods for deriving seasonal climate forecasts from GCM´s. In: Hammer, G.L., Nichols, N., Mitchell, C. (Ed.). *Applications of seasonal climate forecasting in agricultural and natural ecosystems.* Kluwer Academic Publishers, London, 2000, p.145.

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