Assessment of probabilistic forecast ‘skill’ using p-values

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

[1]Embrapa Meio Ambiente, P.O. Box 69, Jaguari?na, SP, Brazil, CEP 4350. www.cnpma.embrapa.br/
Email ahmaia@cnpma.embrapa.br
[2] Queensland Department of Primary Industries and Fisheries, P.O. Box 102, Toowoomba, Qld 4350.
www.dpi.qld.gov.au/
Email holger.meinke@dpi.qld.gov.au; sarah.lennox@dpi.qld.gov.au

Abstract

The establishment and communication of climatological forecast ‘skill’ are complex issues requiring simple approaches. The major issues are: (a) inappropriate use of significance testing to quantify signal intensity, (b) skewed probability distributions of time series of bio-physical data (such as rainfall, crop or pasture production) rendering parametric skill measures based on Normal distribution inadequate, (c) for a spatial assessment of forecast skill, the use of skill measures derived from parametric tests require location-by-location checking of assumptions about underlying distributions, making the process cumbersome and expensive, (d) the level of significance required for forecast skill to be useful depends on the user and the application rather than on an arbitrary, pre-determined significance level and (e) signal intensity varies temporally and spatially. Hence, we propose the use of p-values derived from non-parametric tests such as the Log-Rank test as direct indicators of signal intensity. This method does not require any knowledge of the underlying data structure, nor does it require any arbitrarily chosen level of significance. 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 signal intensity. We illustrate this method by assessing the ability of a three-way ENSO classification in forecasting winter rainfall across Australia.

Media summary

To quantify the signal intensity (‘skill’) of forecast systems we have mapped p-values arising from non-parametric tests applied to rainfall recording stations across Australia.

Key words

Log-Rank test, skill measures, spatial analysis, seasonal forecasting, climate, ENSO.

Introduction

Sound agricultural risk management requires objective assessment of alternative probabilistic outcomes. In highly variable climates, seasonal climate forecasting in combination with simulation models of farming systems has therefore become a powerful tool for risk assessments and the evaluation of 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 (‘conditional’ distributions) corresponding to climate indicators such as the Southern Oscillation Index (SOI), El Ni?o/ Southern Oscillation (ENSO) phases or Sea Surface Temperature (SST) phases (Meinke and Stone 2004). This approach has been used worldwide and has provided valuable information for decision makers managing climate-sensitive systems (Messina et al. 1999; Phillips et al. 1999; Potgieter et al. 2002).

These conditional distributions are frequently presented as cumulative distribution functions (CDFs) or their complement, probability of exceedance functions (POEs). They are a simple and convenient way to represent probabilistic information arising from a time series. Such representation is particularly valid for time series that exhibit 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 serial auto-correlation thus allowing the CDF/POE representation to convey seasonal climate forecast information (e.g. Selvaraju at al. 2004).

For the evaluation of forecast ability, a statistical hypothesis testing framework serves two purposes: (a) to establish ?significance? to assist in the determination of causal links between classification systems and the forecast quantity (eg. rainfall) and (b) to quantify the signal intensity of the forecast system. We define signal intensity as the contribution of the classification system (ie. 'forecast' system) to the overall variability of the response variable, such as rainfall, temperature, yield, drainage, runoff (eg only). Given that the impact of ENSO and their causal links on rainfall are well established, no further significance testing is required. However, we suggest instead mapping p-values as a summary measure of signal intensity to quantify the spatial impact and variability of the classification system.

Many parametric and non-parametric statistical procedures can be used to quantify signal intensity arising from different classes of CDFs or POEs. Parametric tests require assumptions regarding the nature of the underlying distribution corresponding to each strata. To establish if these assumptions have been violated, exploratory analyses need to be performed. For a spatial assessment, data from each location need to be assessed prior to applying such parametric tests, thereby making the process cumbersome and expensive.

Non-parametric procedures have the advantage of not requiring any prior knowledge regarding the nature of the underlying distribution. In order to quantify the overall segregation effect, we need to test the null hypothesis that the conditional distribution originated from the same population as the unconditional distribution. There are many parametric and non-parametric procedures available to test such hypothesis, e.g. Snedecor's F (F), Likelihood ratio (LR), Kolmogorov-Smirnov multisample (KSM) and Log-Rank (LGRK) test (Mantel 1966; Montgomery 2001; Conover 1980; Kalbfleish and Prentice 1980). Here we outline how the results from such tests can be presented in order to convey spatial variability of signal intensity arising from probabilistic forecast systems. For this purpose, we use the LGRK test as an example – others might be equally appropriate.

The use of the LGRK test to compare rainfall POE was proposed by Maia and Meinke (1999). This test is a non-parametrical tool widely used in clinical trials to evaluate possible strata effects on time-to-failure distributions, also known as survival distributions (Kalbfleish and Prentice 1980).

Statistical significance tests have been criticised for using arbitrary levels of significance (Nicholls 2001). This is particularly pertinent in relation to forecast skill measures, since even a moderately 'skilful' forecast (low signal intensity) can result in high value, depending on how the forecast is applied in operational risk management (Meinke and Stone 2004). Therefore, we suggest to clearly differentiate between (a) the establishment of significance to assist in the determination of causal links between classification systems and the forecast quantity (eg. rainfall) and (b) the quantification of signal intensity arising from the forecast system. To achieve this, we can select an appropriate statistical test for the hypothesis of 'no skill' and subsequently use p-values associated with the corresponding test statistic as a quantitative measure for signal intensity (e.g. Stone 1992). A p-value ranges between 0 and 1 and is inversely proportional to the degree of evidence against the hypothesis of no-strata effect. Thus lower p-values indicate higher signal intensity. The magnitude of evidence takes into account the length of the series. We therefore propose mapping p-values as a tool for assessing spatial and temporal variability of probabilistic forecast ability in climatology, agriculture and natural resource management. Here we provide a simple example of this approach.

Materials and methods

For illustration purposes we used 3-monthly total rainfall records (1901-2002) of June-August (JJA) winter rainfall from 64 high-quality rainfall recording stations across Australia. From the 102 years analysed, 24 were classified as El Ni?o, 22 as La Ni?a and 56 as 'Other'. The classification system was derived from a combination of ocean and atmospheric data sets (Potgieter et al. 2004). The 64 JJA rainfall series (POEs) were segregated into three sub-series (conditional POEs) according to this ENSO classification, resulting in 192 sub-series with variable lengths. Series with a high frequency of zeros (e.g. in seasonally dry areas such as Australia's northern regions) will not follow either Gamma, Log Normal or Normal distributions (Lennox 2003) and therefore skill measures derived from parametric tests such as p-values associated with F or LR tests based on those distributions are not adequate. Hence, we calculated the percent of JJA periods with zero rainfall amount for each sub-series and evaluated the goodness-of-fit to the above referred distributions using p-values arising from Chi-square tests. We also applied Bartelett's Kolmogorov-Smirnov white noise tests for autocorrelations (Durbin 1967) using the SPECTRA procedure of SAS System (SAS 1998). LGRK tests (Mantel 1966) were performed using the SAS Procedure PROC LIFETEST (SAS 1998) in order to quantify the magnitude of any ENSO class effect on the JJA POE functions. The p-values associated with LGRK were recorded.

Results and discussion

Of the 192 sub-series analysed, 47% had at least one year with zero JJA rainfall. The percent of sub-series showing low goodness-of-fit to Normal (58%), Log Normal (38%) or Gamma (28%) distributions was high and always greater than the overall percent of sub-series with at least one zero JJA rain. This finding confirms that the lack of fit is not solely due to the binary nature of rainfall as expressed in dry versus wet days (Dunn 2003). This provides further evidence that distributions other than those widely used for describing rainfall series should be considered if we adopt a parametrical approach e.g. Tweedie distributions (Jφrgensen 1987) as proposed by Lennox et al. (2004).

P-values are not readily available for KSM tests. Further, KSM tests rely on maximum distances, whereas the LGRK captures divergences over the entire distribution. Most statistical software packages have special modules to perform non-parametrical survival analysis, including the computation of p-values associated with the LGRK test. This greatly facilitates spatial assessments of signal intensity.

As long as an appropriately dense spatial network of data is available, continuous p-value maps can be constructed using interpolation methods, e.g. Model Based Geostatistics (MBG) (Diggle et al. 1998). The main advantage of using MBG instead of other widely used interpolation methods is that hypotheses about coherence of spatial patterns can be tested and spatial confidence limits for predictions can be generated.

For demonstration purposes we have spatially interpolated the p-values derived from LGRK shown in Fig. 1a and provided a corresponding 'signal intensity map' for all of Australia (Fig. 1b). Based on LGRK for each station, our analysis shows typical ENSO impacts that have been frequently reported (e.g. Nicholls and Wong 1991; Drosdowsky 1993). As expected, ENSO impacts are most noticeable throughout the eastern part of Australia confirming the appropriateness of the suggested approach. This signal intensity measure based on p-values also provides a means to objectively compare different forecast systems – a critically important issue in order to improve risk management practices related to climate variability.

We acknowledge that for large parts of Australia, the station density of our data set is too sparse for an detailed assessment of signal intensity. However, we are currently investigating a much larger data set and have included Fig. 1b solely to demonstrate the utility of the approach, rather than providing an accurate assessment.

For rainfall data, the use of p-values associated with likelihood tests based on Tweedie distributions could also be considered (Lennox et al. 2004). This would provide a suitable means to address the binary nature of rainfall events as well as the skewness of rainfall amounts. However, as with all parametrical approaches, this would require a location-by-location assessment of goodness-of-fit to, for example, Gamma, Log Normal or Normal distributions, which is one of the major disadvantages of parametric approaches.

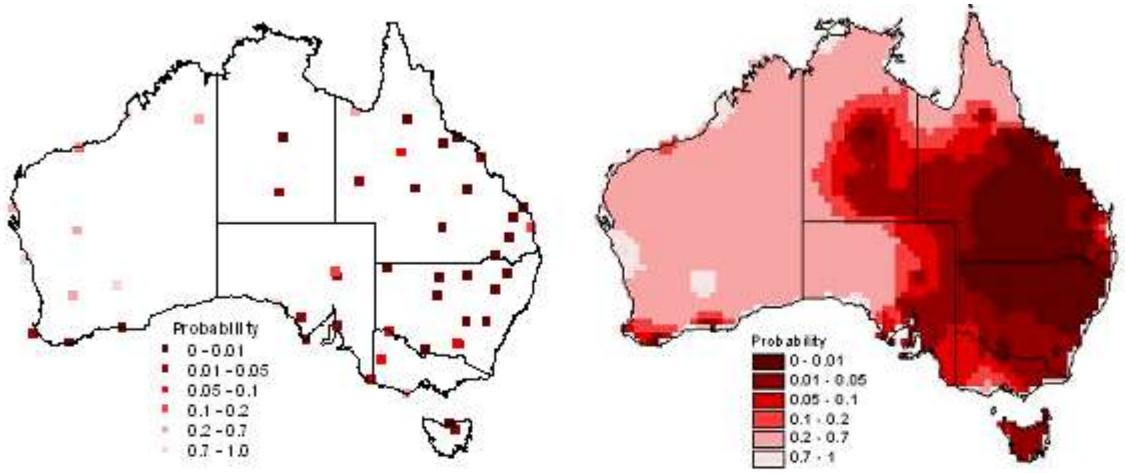

Fig. 1. ENSO signal intensity based on p -values derived from the Log-Rank test applied to compare conditional POEs of 3-monthly JJA rainfall records (1900-2002) from 64 rainfall recording stations across Australia.

References

Conover W J (1980). ‘Practical Nonparametric Statistic’. John Wiley & Sons, New York.

Diggle P J, Tawn J A and Moyeed R A (1998). Applied Statistics 47, 299-350.

Drosdowsky W (1993). Int. J. Climatol. 13, 111-149.

Dunn P K (2003). Working Paper, Series SC-MC-0305. Faculty of Sciences, University of Southern Queensland, Australia. .

Durbin J (1967). Bulletin of the International Statistics Institute 42, 1039-1049.

Kalbsfleish J D and Prentice R L (1980). ‘The Statistical Analysis of Failure Time Data’. John Wiley & Sons, New York.

Lennox S M (2003). B Sc (Hons) thesis, University of Southern Queensland, Australia.

Lennox S, Dunn P K, Power B D and deVoil P (2004). These proceedings.

Jφrgensen B (1987). Journal of the Royal Statistical Society, Series B, 49, 127-162.

Mantel N (1966). Cancer Chemother. Rep. 50, 163-70.

Maia A H N and Meinke, H (1999). 2nd International Symposium on Modelling Cropping Systems, Lleida, Spain, 103-104.

Meinke H and Stone R C (2004). Climatic Change, in press.

Messina C D, Hanssen, J W and Hall, A J (1999). Ag. Systems 60, 197-212.

Montgomery D C (2001) ‘Design and Analysis of Experiments’ 5th ed., John Wiley & Sons, New York.

Nicholls N (2001). Bull. Amer. Meteor. Soc. 81, 981-986.

Nicholls N and Wong K K (1991). J. Climate, 3, 163-170.

Phillips J, Rajagopalan B, Cane M and Rosenzweig C (1999). Int. J. Climatology 19, 877-888.

Potgieter A B, Hammer G L and Butler D (2002). Aust. J. Agric. Res. 53, 77-89.

Potgieter A B, Hammer G L, Meinke H, Stone R C and Goddard L (2004). J. Climate, in press.

SAS Institute Inc.(1998) ‘SAS STAT Users Guide’ Version 6.12, NC, Cary. SAS Institute Inc.

Selvaraju R, Meinke H and Hansen J (2004). These proceedings.

Stone R C (1992). Unpublished Ph D Thesis, University of Queensland, Brisbane, Australia.