Adv. Geosci., 20, 19–23, 2009
www.adv-geosci.net/20/19/2009/
© Author(s) 2009. This work is distributed under
the Creative Commons Attribution 3.0 License.

Advances in
Geosciences

Relationship between forecast precipitation relative errors and skill scores: the case of rare event frequencies

N. Tartaglione

Department of Physics, University of Camerino, Camerino, Italy

School of Mathematical Sciences, University College Dublin, Dublin, Ireland

Received: 30 September 2008 – Revised: 9 March 2009 – Accepted: 10 March 2009 – Published: 12 March 2009

Abstract. This paper addresses the problem of the relationship between skill scores and forecast rainfall relative errors. The problem is approached by using synthetic time series of rainfall data representing the observations. It is assumed that the magnitude of the relative error is known. The forecasts are constructed by adding errors to the observations. We use a threshold to dichotomise forecasts and observations to obtain the skill scores. We perform 1000 simulations for each error magnitude in order to obtain the mean values and uncertainties of the scores.

We consider two different precipitation regimes, and we show the influence of these regimes on the precipitation. We find that the relationship between forecast errors and skill scores is strongly influenced by the event frequencies, which in turn depend on the precipitation regime. We find that only when the event frequency of the two regimes is made similar by changing the threshold, the relationship between the scores and relative errors is similar. This suggests that a comparison between two forecast precipitation datasets should account for the difference (if any) in precipitation regimes.

1 Introduction

The computation of forecast scores has become more widespread in recent years for many reasons (Jolliffe and Stephenson, 2003), whereas score uncertainty is rarely evaluated (Jolliffe, 2007). Even when uncertainty is computed, the meaning behind such scores is not always understood (Mason, 2008). In fact, computing scores gives rise to the question: "does the score value indicate a good forecast?" This question has recently been considered by Mason (2008), who discussed the probability that useless forecasts may have scored simply by chance. One of the qualities of a score is the effectiveness, defined as the property of a score to follow the differences between observation and forecast (Mason, 2008). Another question that arises is whether the dependence of scores on errors is related to precipitation regimes. The aim of this paper is to try to answer these questions.

Since forecast errors are unknown, one way to answer the question is to use synthetic data. In this paper we use time series instead of gridded data, for reasons given in the data and methods section (Sect. 2). Results are presented in Section 3 and conclusions are drawn in Sect. 4.

2 Scores, data and methods

In this section we first describe the scores. Then we describe the data and methods used to evaluate the relationship between scores and forecast relative errors.

2.1 Scores used in verification

In a dichotomous forecast, a contingency table shows the frequency of "yes" and "no" forecasts and occurrences. Table 1 gives the combinations of hits (a), misses (b), false alarms (c) and correct negatives (d).

The analysis presented here is limited to two scores, the equitable threat score (ETS) and the Hanssen-Kuipers skill score (KSS). Their mathematical formulations are expressed below:

$$\text{ETS} = \frac{a-e}{a+b+c-e} \tag{1}$$

where

$$e = \left(\frac{(a+b)\,(a+c)}{a+b+c+d} \right) \tag{2}$$

is the probability of having hits by chance. ETS is commonly used to evaluate forecast skill, especially precipitation (Accadia et al., 2003; Hamill, 1999; Hamill and Juras, 2006).

Correspondence to: N. Tartaglione
(nazario.tartaglione@unicam.it)

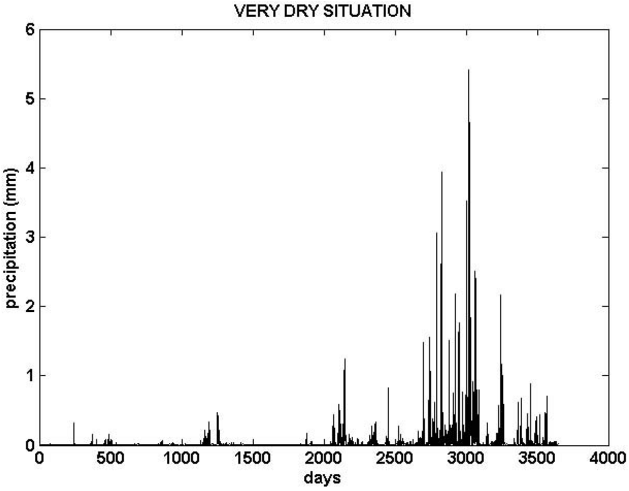

Fig. 1. Precipitation time series for the “very dry” situation.

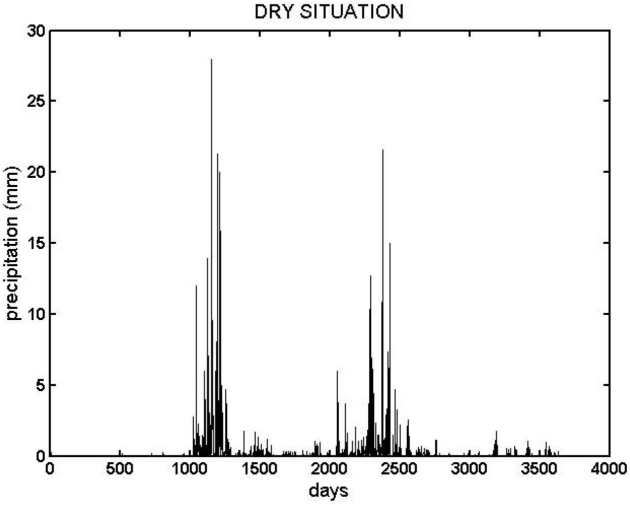

Fig. 2. Precipitation time series for the “dry” situation

The KSS was introduced by Hanssen (1965) and it is usually used in the verification of meteorological fields (Briggs, 2005; Accadia et al., 2003; Woodcock, 1976). KSS is defined as:

$$\mathrm{KSS}=\frac{(ad-bc)}{(a+c)\,(b+d)} \tag{3}$$

Both of these scores are equitable, i.e., they will give the same value for two unskilful forecasts. Both of them are skill scores, i.e., they account for random chance, persistence or climatology. Although equitability should be taken into account, Hamill and Juras (2006) suggested that climatology can affect the ETS. This aspect deserves further investigation, and is examined in Sect. 3.

Table 1. Contingency table used to compute the scores.

		OBSERVED	
		Yes	No
FORECAST	Yes	a	b
	No	c	d

2.2 Observations and forecasts

The approach used in this paper is to generate synthetic data. In particular, we use time series instead of gridded precipitation in order to avoid problems related to the presence of a grid. Among these, the problems of multiplicity (Livezey and Chen, 1983), the double penalty effect (Mass et al., 2002) and different climatology on the same grid can affect skill scores (Hamill and Juras, 2006).

We produced two time series, one referred to as “very dry” (Fig. 1) and the other “dry” (Fig. 2). It is important to understand how the event frequency may impact the scores. More common situations are treated in Tartaglione (2009). In this way we can also assess how errors affect ETS and KSS for rare events.

The time series were produced by means of a multiplicative cascade algorithm (Flores, 2004). From an initial level $i=0$, an initial water mass M is distributed on a number of cells at successive levels, by assigning random numbers and imposing mass conservation. After a number of cascades, we arrive at a daily distribution of water such as that shown in Figs. 1 and 2. The elements of the single time series are uncorrelated with each other.

The distinction between rain and no-rain events was performed by assigning a threshold of 0.5 mm/day. In such a case, the “very dry” situation had 68 events over-threshold and the “dry” situation 267. The event frequencies are ∼0.02 and ∼0.07, which indicate rather rare frequencies. However, we would stress how even small differences between the event frequencies can alter the evaluation of scores, especially when the events are rare. We shall see that the number of over-threshold events plays a key role in determining the value and uncertainty of the scores, as already suggested by Hamill and Juras (2006) and Baldwin and Kain (2006). Our interest here is to understand how scores are related to errors. We construct a simple error model and add the errors to the observations in order to simulate forecasts. We wish to know the magnitude of the errors. Let us define the relative error in the following way:

$$\varepsilon_r = \left| \frac{f - o}{o} \right| \tag{4}$$

where o represents the observations and f is the forecast. In practice, our relative errors span from 0.1 to 1, corresponding to a span in percentage errors of 10% to 100%. The absolute error is defined as the observation minus the forecast on the same day, and is added to the observation according to

$$f = o + (-1)^n \cdot o \cdot \varepsilon_r , \tag{5}$$

where n is the pair (0, 1). The value of n sets the error sign, which was randomly assigned to the elements of the time series. The value of n was obtained by using a uniform random number generator. For each realization, the algorithm chose to assign zero when the random number was less than 0.5 and 1 otherwise. The assignment of the error signs was performed 1000 times. In this way we can perform a statistical evaluation of scores and compute the score uncertainties.

Since for each sample we had almost 50% of 0 (or $+$) and 50% of 1 (or $-$), the forecast result was unbiased in the long run, i.e., the bias oscillated around the value 1. Once the scores were computed, the 1000 samples of ETS and KSS were plotted. This yielded Gaussian distributions, for which we computed the mean and the standard deviation.

A note of caution is required here. We performed some experiments with different time series and obtained similar results. However, particular distributions of the precipitation might lead to very specific situations, which should be treated carefully.

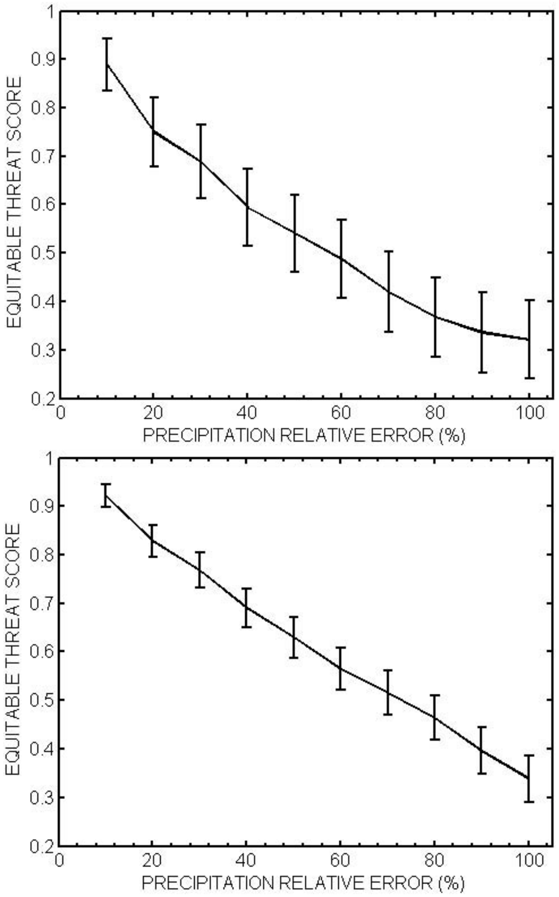

Fig. 3. Equitable threat score for the "very dry" (upper panel) and "dry" (lower panel) situations, with a threshold of 0.5 mm/day.

3 Results

We show the trend of ETS and KSS as a function of the relative errors for the two time series considered, in Figs. 3 and 4 respectively. For the two time series examined here, the "very dry" situation gives lower errors than the "dry" situation, when a fixed score is considered. This means that when we are evaluating a forecast in a precipitation regime that has a low event frequency, the resulting score will indicate a low forecast error. Hence, for a given relative error, the expected score should be higher in the "very dry" situation relative to the "dry" situation.

It is interesting to note the length of the error bars in the two figures. The error bars represent the score uncertainties, and for the "very dry" situation they are longer than for the "dry" one. The number of over-threshold events seems to affect the score value and the score uncertainty. This means that a single verification, such as a real forecast verification, might give a result that is within the error bar. To overcome this problem, the score uncertainties are computed in various ways (see Jolliffe (2007) for a review of methods of assessing uncertainty).

The error bars are particularly long for the KSS (Fig. 4). Their length increases with the forecast percentage relative error, whereas the ETS uncertainty remains constant. We also note that the linear relationship between mean score and forecast relative error tends to disappear for high values of relative error. This is particularly evident for the KSS.

We do not demand that a score has a linear dependence on the errors, but one of the properties of a good score is the effectiveness. Citing Mason (2008): "*An effective score is one which monotonically improves as the distance (however it is measured) between the forecast and the observation decreases*". Thus, we can imagine that this monotonic behaviour should not be in a functional sense, but in a statistical one. The KSS, composed of mean value and uncertainty, is certainly not monotonic for high values of the forecast relative errors, even though its mean is apparently monotonic.

We have analyzed two situations which are climatologically different, but verified the results using the same threshold (0.5 mm/day). What happens when the threshold of the "dry" situation is lowered? The values of ETS and KSS as a function of the percentage relative error, when the threshold of the "very dry" situation is 0.1 mm/day, are shown in Fig. 5.

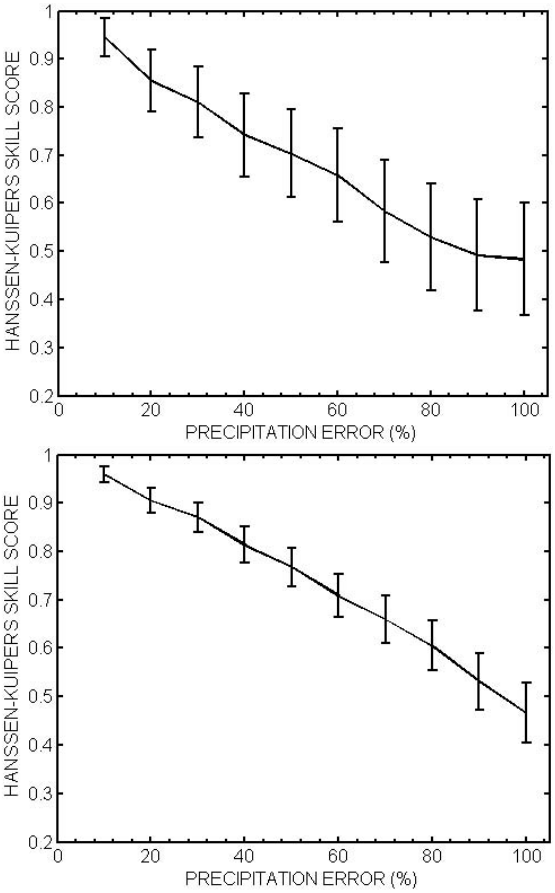

Fig. 4. Same as Fig. 3, but for the Hanssen-Kuipers Skill Score.

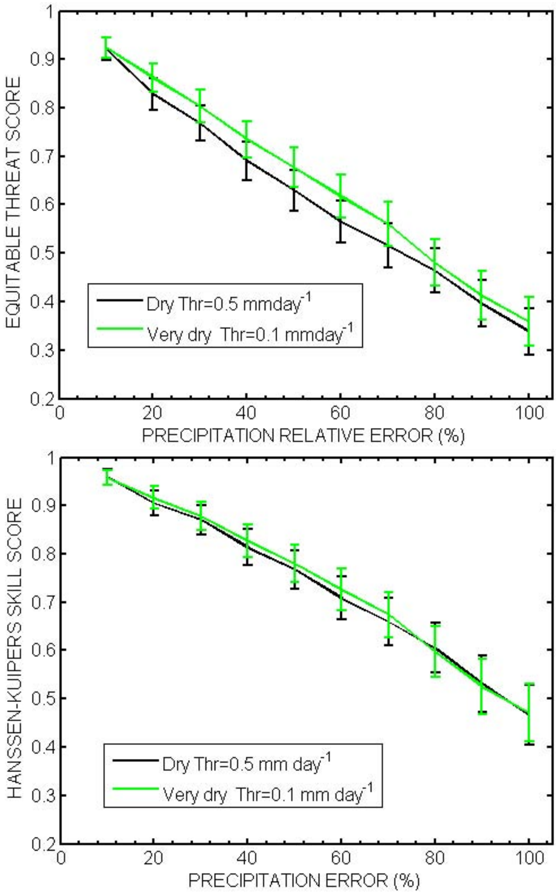

Fig. 5. Equitable threat score and Hanssen-Kuipers score, with the two situations considered having similar event frequency.

The trends of the two scores are very close to each other, demonstrating that the event frequency is fundamental in the evaluation of skill scores. This argument was also pointed out by Baldwin and Kain (2006). These results indicate that comparing two datasets by means of scores obtained from a contingency table, and using two different thresholds, leads to a different evaluation of the two datasets, even though they have the same relative error.

4 Conclusions

In this work we have shown that the two skill scores, the ETS and the KSS, depend on the forecast relative errors. Using a Monte Carlo experiment, we computed a variety of skill ETS and KSS values for the same magnitude of the forecast relative error. This was performed on two different precipitation time series, which represented the observations for two different climatological situations.

The ETS and KSS scores depend on the threshold used to compute the score. In fact, the scores were computed by assigning a threshold for the event occurrence. This threshold was initially fixed to 0.5 mm/day for both situations considered. The score values and uncertainties are dissimilar for the two considered situations. They become similar when the threshold of the “very dry” situation is reduced to give an event frequency similar to the “dry” situation. This suggests that a comparison between two different climatological situations should be performed using different thresholds. The results confirm previous findings by Hamill and Juras (2006) and Baldwin and Kain (2006) on the importance of climatology and event frequencies on the evaluation of forecast precipitation by means of skill scores.

Acknowledgement. A special thank you goes to two reviewers of this paper for their suggestions and comments. Part of this work has been funded by “Agenzia Spaziale Italiana” under the contract “Progetto Pilota Protezione Civile dale Alluvioni: il nowcasting”.

Edited by: S. C. Michaelides
Reviewed by: two anonymous referees

References

Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A., and Speranza, A.: Sensitivity of Precipitation Forecast Skill Scores to Bilinear Interpolation and a Simple Nearest-Neighbour Average Method on High-Resolution Verification Grids, Wea. Forecasting, 18, 918–932, 2003.

Baldwin, M. E. and Kain, J. S.: Sensitivity of several performance measures to displacement error, bias, and event frequency, Wea. Forecasting, 21, 636–648, 2006.

Briggs, W., Pocernich, M., and Ruppert, D.: Incorporating Misclassification Error in Skill Assessment, Mon. Weather Rev., 133, 3382–3392, 2005.

Flores, C.: Multiplicative cascade models for rain in hydrometeorological disasters risk management, 35. ASTIN-Kolloquium, Bergen, Norway, 6–9 June 2004 (available at http://www.actuaries.org/ASTIN/Colloquia/Bergen/Flores.pdf), 2004.

Hanssen, A. W. and Kuipers, W. J. A.: On the relationshipship between the frequency of rain and various meteorological parameters. Koninklijk Nederlands Meteorologisch Institut, Meded. Verhand., 81, 2–15, 1965.

Hamill, T. M.: Hypothesis tests for evaluating numerical precipitation forecasts, Wea. Forecasting, 14, 155–167, 1999.

Hamill, T. M. and Juras, J.: Measuring forecast skill: is it real skill or is it the varying climatology?, Q. J. Roy. Meteor. Soc., 132, 2905–2923, 2006.

Jolliffe, I. T.: Uncertainty and inference for verification measures, Wea. Forecasting, 22, 637–650, 2007.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practitioner's Guide in Atmospheric Science, Wiley: Chichester, 240 pp., 2003.

Livezey, R. E., and Chen, W.: Statistical Field Significance and its Determination by Monte Carlo Techniques, Mon. Wea. Rev., 111, 46–59, 1983.

Mason, S. J.: Understanding forecast verification statistics, Meteor. Appl., 15, 31–40, 2008.

Mass, C. F., Ovens, D., Westrick, K., and Colle, B. A.: Does Increasing Horizontal Resolution Produce More Skillful Forecasts?, B. Am. Meteor. Soc., 83, 407–430, 2002.

Tartaglione, N.: Relationship between Precipitation Forecast Errors and Skill Scores of Dichotomous Forecasts, Wea. Forecasting, submitted, 2009.

Woodcock, F.: The Evaluation of Yes/No Forecasts for Scientific and Administrative Purposes, Mon. Weather Rev., 104, 1209–1214, 1976.