TECHNICAL MEMORANDUM

615

A New Equitable Score Suitable for Verifying Precipitation in NWP

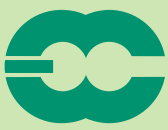
M.J. Rodwell, D.S. Richardson
and T.D. Hewson

Research Department

In Press in *Quart. J. Roy. Meteorol. Soc.*

February 2010

This paper has not been published and should be regarded as an Internal Report from ECMWF.
Permission to quote from it should be obtained from the ECMWF.



European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

©Copyright 2010

European Centre for Medium-Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. This publication is not to be reprinted or translated in whole or in part without the written permission of the Director. Appropriate non-commercial use will normally be granted under the condition that reference is made to ECMWF.

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error, omission and for loss or damage arising from its use.

Abstract

A new equitable score is developed for monitoring precipitation forecasts and for guiding forecast system development. To accommodate the difficult distribution of precipitation, the score measures error in 'probability space' through use of the climatological cumulative distribution function. For sufficiently skillful forecasting systems, the new score is less sensitive to sampling uncertainty than other established scores. It is therefore called here 'Stable Equitable Error in Probability Space' (SEEPS). Weather is partitioned into three categories: 'dry', 'light precipitation' and 'heavy precipitation'. SEEPS adapts to the climate of the region in question so that it assesses the salient aspects of the local weather, encouraging 'refinement' and 'discrimination' and discouraging 'hedging'. To permit continuous monitoring of a system whose resolution is increasing with time, forecasts are verified against point observations. With some careful choices, observation error and lack of representativeness of model grid-box averages are found to have minimal impact. SEEPS can identify key forecasting errors including the over-prediction of drizzle, failure to predict heavy large-scale precipitation, and incorrectly locating convective cells. Area-averages are calculated taking into account the observation density, so that all sub-regions are treated more equally. A gain of $\sim$2 days, at lead-times 3–9 days, over the last 14 years is found in extratropical scores of forecasts made at the European Centre for Medium-range Weather Forecasts (ECMWF). This gain is due to system improvements, not the increased amount of data assimilated. SEEPS may also be applicable for verifying other quantities that suffer from difficult spatio-temporal distributions.

1 Introduction

Routine verification is crucial in numerical weather prediction (NWP) for monitoring progress, setting targets, comparing forecasts by different centres and for guiding development decisions. Through these various roles, verification scores for the large-scale flow have helped drive impressive improvements in NWP performance. An example of these improvements is that an 8 day (D+8) ECMWF forecast for the Northern Extratropics in 2008 has the same average spatial anomaly correlation skill (for 500 hPa geopotential heights, Z500) as a D+5$\frac{1}{2}$ forecast had in 1980.

Contours in Fig. 1 show (a) observed (*i.e.* analysed) and (b) D+4 forecast Z500 verifying at 12 UTC on 23 August 2008. The correspondence is indicative of the improvements in large-scale skill. However, it is clear that Z500 is not sufficient to characterise the entire flow. Precipitation (shaded), for example, is rather poorly predicted over Europe. This emphasises the need to monitor other aspects of the forecast; for example aspects of direct relevance to the user community and aspects representative of diabatic processes. It is difficult, however, to make development decisions based on many scores. Ideally decisions should be based on some minimal number of scores that concisely summarise a system's performance. Since precipitation is user-relevant and a consequence of diabatic processes, it would appear to be a natural choice.

Precipitation is a difficult quantity to verify for numerous reasons. Firstly, it is rather sparsely observed by surface observations and imperfectly estimated by radar (where available) and satellite at present. Secondly, a point observation may not be representative of a model grid-box average. Thirdly, precipitation has a difficult spatio-temporal distribution with, often, a large number of dry days and occasional very extreme events (notice the non-linear colour scale in Fig. 1). Any precipitation score must contend with these issues.

Considerable research has focused on developing precipitation scores. For example, Du et al. (2000), following Hoffman et al. (1995), partitioned precipitation forecast error into components associated with large-scale advection, magnitude and a residual. Casati et al. (2004) partitioned error by intensity and spatial scale. Such decompositions are essential to truly understand the nature of forecast error but can be overly complex for the purposes of monitoring. Importantly, they do not necessarily give useful guidance for high-level development decisions. For instance, doubling precipitation everywhere constitutes a major change to a forecast system, but it would not alter values of the intensity-scale score of Casati et al. (2004). Other research has centred on

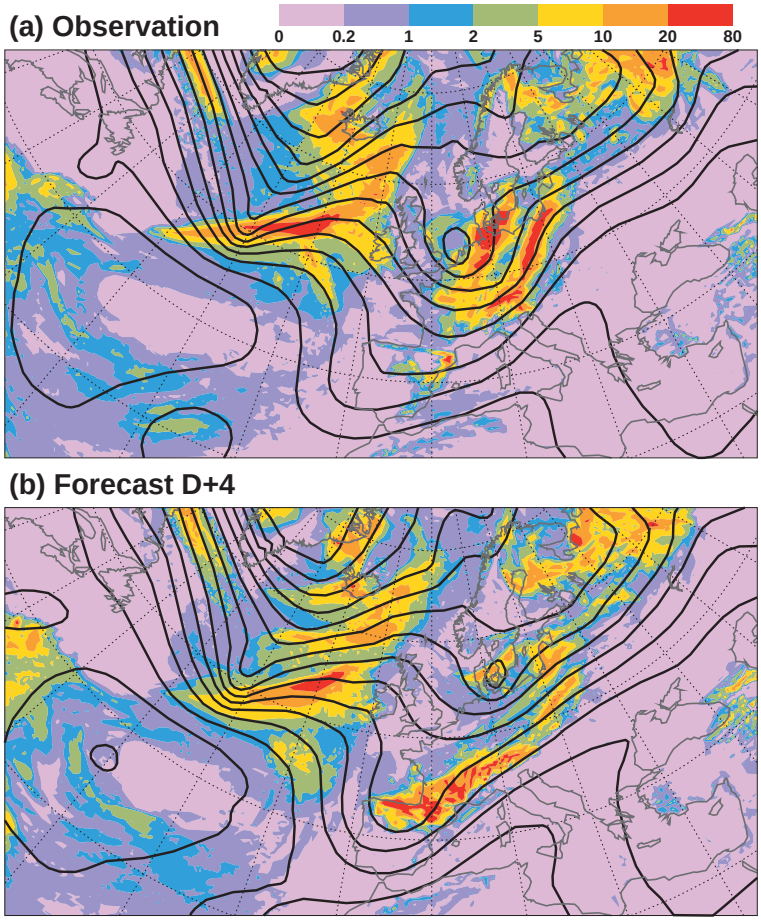


Figure 1: 500 hPa geopotential height field (Z500, contoured with interval 50m) and 24-hour accumulated precipitation (shaded, mm). (a) 'Observations': analysed Z500 and short-range (D+0–D+1) forecast precipitation centred at time 12 UTC on 23 August 2008. (b) Forecast: D+4 forecast Z500 and D+3$\frac{1}{2}$–D+4$\frac{1}{2}$ forecast precipitation verifying at the same time.

the verification of extreme precipitation (e.g. Stephenson et al., 2008). This is highly desirable from the users' perspective, but sampling uncertainties render it a difficult task.

Here the aim is to develop a new score that concisely quantifies NWP performance in the prediction of precipitation and steers development in the correct direction. The desirable attributes of such a score can be summarised as

(a) Monitoring Progress

- A single score that assesses forecast skill for dry *and* wet weather.
- Verification against point observations in order to permit continuous monitoring of a system whose resolution is increasing with time, and to satisfy the typical user interested in a small geographic area.
- To detect performance changes, sensitivity to sampling uncertainty should be minimised, while maintaining the ability to differentiate between 'good' and 'bad' forecasts.
- For area and temporal averages to be meaningful, it should be possible to combine scores from different climate regions and different times of the year.

(b) Aiding decision-making

- To facilitate the identification of model error, there should be a clear link between the score and the error in the forecast.
- A score should encourage developments that permit a forecast system to predict the full range of possible outcomes.
- A better score should indicate a 'better forecast system'.

Two key approaches are used as a starting-point for the present study. The first represents a method discussed by Ward and Folland (1991). They transformed seasonal-mean precipitation anomalies into 'probability space' through the application of the observed cumulative distribution function. This results in a score known as Linear Error in Probability Space (LEPS). The transformation handles, in a natural way, the difficult distribution of precipitation and makes a score much less sensitive to extreme values. The LEPS approach seems attractive for the routine scoring of moderate daily precipitation accumulations if the problem of the existence of dry days can be overcome. The second approach is the application of 'equitability' constraints (Gandin and Murphy, 1992) that place upper and lower bounds on the expected skill scores for perfect and unskillful forecasting systems, respectively. Defined bounds facilitate the comparison and combination of scores from climatologically different regions and from different times of the year. If a score is *in*equitable, it is possible for an unskilled forecast system to score better than a forecast system with some skill. This is clearly undesirable.

The data used here are described in section 2. Section 3 reviews some established scores and comments further on 'equitability' and 'error in probability space'. The new score is developed in section 4. Section 5 compares this score with other established scores in terms of sampling uncertainty and susceptibility to hedging. Section 6 discusses some parameter settings and section 7 gives a summary of the new score. Section 8 applies the score to some case-studies. Area-mean scores, that take account of observation density, are presented in section 9. Section 10 investigates the detection of system improvements. The impacts of observation error and lack of representativeness are quantified in section 11. Conclusions are given in section 12.

2 Data

2.1 Observational Data

2.1.1 Daily SYNOP data from the GTS: 1980–2008

The data used for the point-verification of precipitation are 'SYNOP' observations. Other sources of data, such as retrievals from radar or satellite may be suitable in the future, and could equally be used with the score developed here. Another alternative is to use short-range forecasts of precipitation, as shown in Fig. 1a. The (in)adequacy of such short-range forecasts can be gauged from the scores against real observations presented here. The SYNOP observations used are those that are exchanged in near real-time over the Global Telecommunications System (GTS) and stored at ECMWF in 'BUFR' archives. Verification against these data, which are not assimilated at ECMWF, should provide an independent evaluation of performance and a valuable 'anchor' to the system (Casati et al., 2008). Daily observations of 24-hour accumulated precipitation for the period 1980–2008 are used here. The hope is that a 24-hour temporal average will alleviate to some extent the problem of the lack of representativity of grid-box spatial averages.

Since precipitation is an accumulated quantity, it is generally necessary to derive the required 'observed' accumulations from the raw reports. For example, under European reporting practices, the 6-hour 0–6 UTC accumulation is derived by subtracting the previous 6-hour 18–0 UTC accumulation from the 12-hour 18–6 UTC accumulation. European 24-hour 0–0 UTC accumulations are then deduced by combining this derived 6-hour 0–6 UTC accumulation with the subsequent reported 12-hour 6–18 UTC and the 6-hour 18–0 UTC accumulations.

Because reporting practices vary throughout the World, a general algorithm has been developed that can produce almost all derivable 6, 12 and 24-hour accumulations from the raw observations for periods ending at any hour of the day, regardless of local reporting practice. The algorithm dramatically increases the number of available accumulations. For example, the number of 24-hour accumulations, world-wide, is increased from $\sim$500 to $\sim$4000 for periods ending 0,6,12 and 18 UTC. There are also reported (and derivable) accumulations ending at other times of the day. For example, India reports at 3, 9, 15 and 21 UTC. The results presented here mainly focus on accumulations ending at 12 UTC.

Because forecast error will be measured in 'probability space', quality control can be more relaxed than for (*e.g.*) correlation scores or scores for extreme weather. Here, reported (and derived) 24-hour accumulations are required to be merely $<$1m.

2.1.2 Climatology of daily SYNOP data: 1980–2008

Climatologies for all stations are based on the reported observations and derived accumulations discussed in section 2.1.1. At least 150 daily accumulations are required for a station to be accorded a climatology for a given month. This equates to $\sim$5 years of observations (5x30=150).

With the intention of following the 'LEPS' approach of measuring error in probability space, climatological cumulative distribution functions are derived for these stations. Figure 2 shows the cumulative distribution functions for a range of such stations and months based on 24-hour 12–12UTC accumulations. These cumulative distribution functions have a different structure to those presented for seasonal-mean data by Ward and Folland (1991). In particular, they do not start at zero probability (y-axis) but rather at a value corresponding to the fraction of days with zero reported precipitation (for the month in question). Baoshan, China (Fig. 2a) in July is frequently wet, with only 10% of days being 'dry'. Formosa, Argentina (Fig. 2f) is dry 80% of the time in

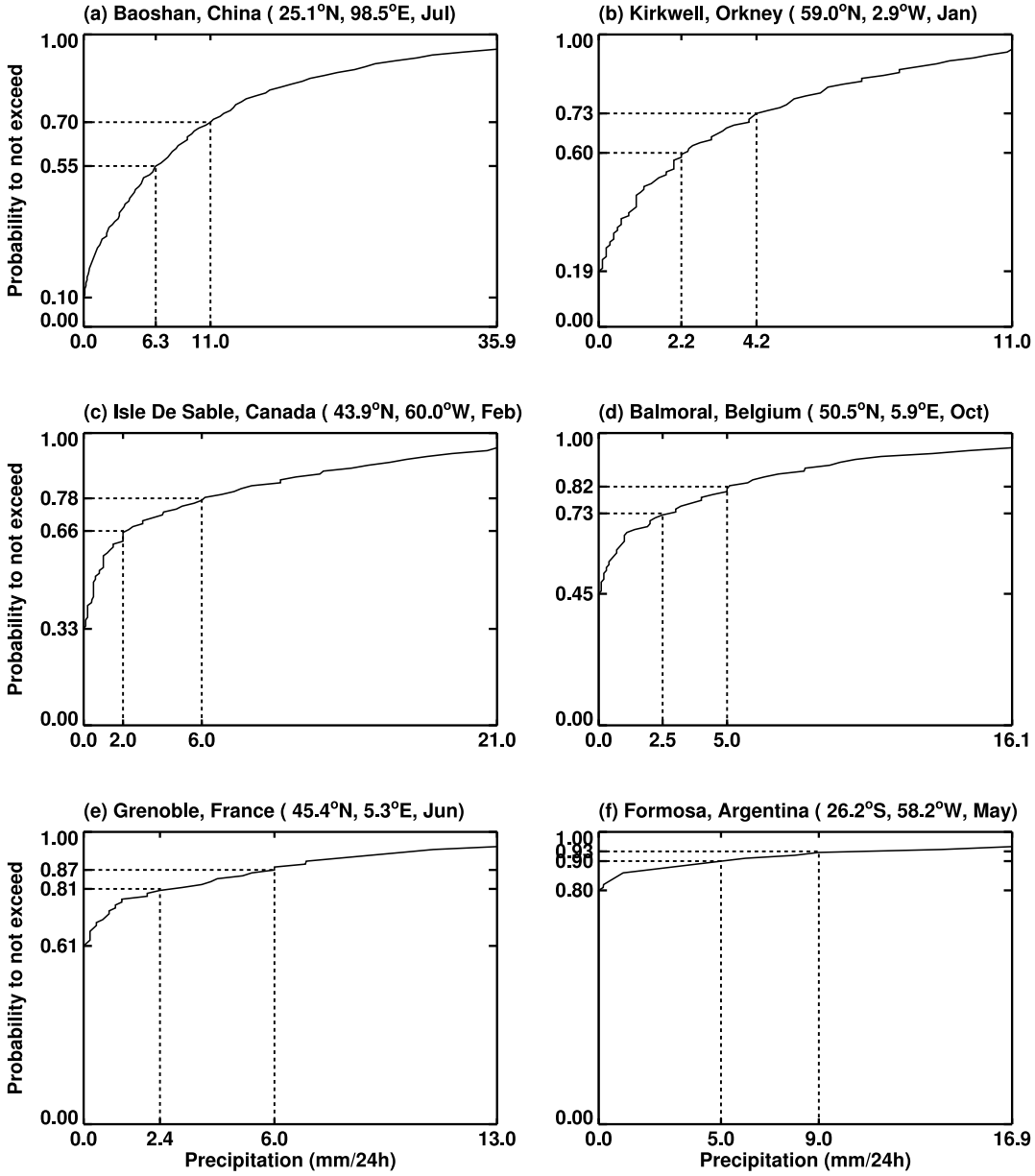


Figure 2: Cumulative distributions for selected SYNOP stations and months based on 12–12UTC 24-hour precipitation accumulations for 1980–2008. The extreme right of each graph corresponds to the 95th percentile of the distribution. Dotted lines indicate the sub-division of the wet days in the ratio 1:1 and 2:1.

May. Figure 2 will be referred to further in subsequent sections.

2.1.3 High-density gridded observations: 2007

Gridded precipitation observations, based on a high-density network of European stations (Ghelli and Lalaurette, 2000), are available from 2002. Cherubini et al. (2002) used this as forecast verification data. Here, point data are required for verification but the gridded 24-hour 6–6UTC accumulations for 2007 (the most recent year available) are used to represent a 'perfect model'. Scoring this perfect model will provide an upper bound for a skill-score that takes SYNOP observation error and representativity into account.

2.2 Forecast Data: 1995–2008

The ECMWF operational 12UTC high resolution ('deterministic') forecast is used to obtain 24-hour 12–12UTC accumulated precipitation forecasts for leadtimes 1–10 days, for the period 1995–2008. These data are matched to all the available SYNOP stations on any given day using the nearest grid-point approach. The alternative approach of bilinear interpolation between the four grid-points surrounding an observation (e.g. Cherubini et al., 2002) was thought more likely to exacerbate the lack of representativity of point data. No account is made for discrepancies between model orographic height and station height and no distinction is made between land-points and sea-points. This should ensure that trends in model performance, including the impact of resolution changes, are not removed from the data.

The operational forecasts are compared with a parallel set of forecasts (for the same period) made within the 'ERA Interim' re-analysis project (Simmons et al., 2007). ERA-Interim uses a single model cycle run at constant resolution. Comparison is also made against a parallel set of test forecasts for a (previously) experimental model cycle for the period 1 April to 8 September 2009.

2.3 What does 'dry' mean?

It is important from the atmospheric physics perspective to assess a forecast's ability to distinguish between wet and dry conditions. However, the definition of 'dry' needs to be applicable to all regions of the World, where reporting practices vary, and should allow a consistent comparison with forecast data. The solution has been to base the definition on the WMO publication 'Guide to Meteorological Instruments and Methods of Observation' (WMO-No. 8, ISBN 978-92-63-10008-5). In part I, chapter 6, the guide states that:

- "Daily amounts of precipitation should be read to the nearest 0.2 mm and, if feasible, to the nearest 0.1 mm".
- "Less than 0.1 mm (0.2 mm in the United States) is generally referred to as a trace".

Based on the second statement the definition of 'dry' must clearly include all forecast (and reported) values strictly less than 0.2 mm. However, with the possibility of rounding, an observation of 0.16 mm could be recorded as 0.2 mm in some parts of the World and simply recorded as 'trace' in other regions. Hence the definition of 'dry' used here is all accumulations $\leqslant$ 0.2 mm. Note that, for rounding to the nearest 0.1 mm, an observation of 0.24 mm would be recorded as 0.2 mm and thus now classified as 'dry'. For compatibility, forecast data is therefore also rounded here to the nearest 0.1 mm prior to classification.

There is a potential caveat in this definition of 'dry'. For regions where rounding is to the nearest 0.2 mm, observations in the interval $[0.25, 0.3)$ mm will be classified as 'dry', while forecast values in this interval will be classified as not 'dry'. Other definitions of 'dry' ("any value < 0.05 mm", "any value < 0.1 mm") have also been tried but the chosen definition seems preferable in that it is as-compatible-as-possible with WMO standards and has a higher (more easily observable) threshold. However, it appears that there is little difference in the trends in area-mean scores whichever definition is used.

3 Review of previous scores

3.1 Continuous scores

Continuous (at opposed to categorical) scores of precipitation have previously been considered. For example, the spatial correlation of normalised precipitation (Rodwell, 2005) shows a clear trend of improvement in the prediction of extratropical precipitation at ECMWF. However the contributions to the score from different regions within the area of interest are difficult to assess. In addition, the correlation is sensitive to extreme values, whether real or due to erroneous observations, and this increases the score's uncertainty. Ward and Folland (1991) applied the LEPS approach to continuous (as well as categorical) seasonal-mean precipitation anomalies. The method greatly reduces the sensitivity to extreme values, but it is unclear how this continuous version can be made compatible with the existence of dry weather.

3.2 Categorical scores

This study focuses on the development of a linear categorical score for precipitation. Such scores make use of a scoring matrix, $\{s_{vf}\}$, that defines the score for any given combination of forecast (FC) category, f, and verifying observation (Obs) category, v. The general scoring matrix for an n-category forecast is shown in Table 1. Here, the observed (climatological) probabilities for the n categories are p_1, p_2, $\cdots$, p_n (with $\sum_{v=1}^{n} p_v = 1$). If a set of observation/forecast pairs have a sample distribution of $\tilde{p}_{vf}$, then the sample-mean score over the set of forecasts is given by

$$\tilde{S} = \sum_{v,f} \tilde{p}_{vf} s_{vf} \quad . \tag{1}$$

A tilde ($\tilde{}$) is used here to denote sample-mean values, as opposed to expected (*i.e.* population-mean or climatological-mean) values or constants.

The definition of $\{s_{vf}\}$ will affect the relative sensitivities of the score to the range of possible forecasting errors, such as the over-prediction of drizzle or the under-prediction of heavy rain. In general, one would hope that a better score would indicate a 'better forecast system'. What is meant by a 'better forecast system' is partly subjective - the relative emphases placed on drizzle and heavy rain could reflect the user's subjective notion of the inconvenience associated with these precipitation features, for example. These subjective aspects will be discussed in section 8. Below, some objective criteria that have been used in previous scores are highlighted.

3.2.1 2-category scores

The 'Hit-Rate', or 'Probability of Detection' is an example of a score where a better score value does not necessarily indicate a better forecast system. The Hit-Rate is defined as $H/(H+M)$ where H is the number

			Obs			
	Probability		p_1	p_2	$\cdots$	p_n
	Category		v			
			1	2	$\cdots$	n
FC	f	1	s_{11}	s_{21}	$\cdots$	s_{n1}
		2	s_{12}	s_{22}	$\cdots$	s_{n2}
		$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
		n	s_{1n}	s_{2n}	$\cdots$	s_{nn}

Table 1: General scoring matrix, $\{s_{vf}\}$, for an n-category score. 'FC' refers to the forecast, 'Obs' refers to the verifying observations and the values $\{p_v\}$ refer to the observed climatological probabilities of the categories.

of correctly forecast events (hits) and M is the number of events that were not predicted (misses). Note that $H+M$ is the number of events that actually occurred. For a perfect forecasting system, Hit-Rate$=1$. However, the converse is not true. A single line program that always predicted the event would have Hit-Rate$=1$ but is clearly not a perfect forecasting system. Decisions made on the basis of the Hit-Rate alone could lead to a forecasting system that issued far too many forecasts that the event would happen. What is missing from this score is a penalty for predicting the event when it did not happen (a false alarm) and a bonus for correctly predicting that the event would not occur (a correct negative).

As the sample size increases (so that the observed sample distribution tends to the climatological distribution, $\{\tilde{p}_v\} \to \{p_v\}$), the Hit-Rate (for hits on category 1) will tend to $S = \tilde{p}_{11}/p_1$. Gandin and Murphy (1992) highlighted the desirability of separating the forecasting and scoring tasks by making $\{s_{vf}\}$ independent of $\{\tilde{p}_{vf}\}$ (but possibly dependent on the climatological distribution $\{p_v\}$). Table 2 makes this separation for the Hit-Rate (valid for large sample size).

			Obs	
	Prob		p_1	p_2
	Cat		v	
			1	2
FC	f	1	$\frac{1}{p_1}$	0
		2	0	0

Table 2: Scoring matrix for the Hit-Rate score (for category-1 hits).

The main contribution of Gandin and Murphy (1992) was to introduce some 'equitability' constraints that, if applied to the definition of the scoring matrix, ensure a score does not suffer in the way that the Hit-Rate does. These constraints give a perfect forecasting system an expected skill value 1 and give all constant and random forecasting systems an expected skill 0:

$$\begin{aligned} &\text{Perfect FC:} && \textstyle\sum_v p_v s_{vv} = 1 \\ &\text{Constant FC:} && \textstyle\sum_v p_v s_{vf} = 0 \quad \forall f \\ &\text{Random FC:} && \textstyle\sum_{v,f} q_f p_v s_{vf} = 0 \quad , \end{aligned} \tag{2}$$

where '$\forall$' means 'for all' and $\{q_f\}$ represents any distribution of forecast categories (with $\sum_f q_f = 1$ and

$q_f > 0 \; \forall f$).

A score whose scoring matrix satisfies these constraints is known as an 'equitable score'. Strictly speaking, a score is equitable if any linear transformation of its scoring matrix (of the form $m\{s_{vf}\}+c$, where m, c are real) satisfies (2). Such a transformation simply re-bases and scales a score without altering its properties. In the present study, this transformation has already been applied in the scoring matrices displayed. The implications of equitability can be summarised as follows. If a score is *in*equitable, and it accords different scores to two unskillful (*e.g.* random) forecast systems, then adding some skill to the system with the poorer score could still leave it appearing worse than the other unskillful system. Equitability removes this possibility. By heavily penalising systems that produce a constant forecast (such as for the climatologically most likely category), equitability also encourages 'refinement' (whereby the forecast distribution becomes equal to the climatological distribution; $\{q_f\} = \{p_v\}$, Murphy and Winkler, 1987). Refinement is discussed further in subsequent sections.

Notice that the Hit-Rate (as defined in Table 2) does not satisfy the constant and random constraints and, thus, is not equitable.

As Gandin and Murphy (1992) point out, the random forecast constraint readily follows from the constant forecast constraints so equations (2) actually represent $n+1$ constraints on the n^2 values of an n-category scoring matrix. Since $n+1 < n^2$ (for $n >= 2$), other constraints or optimisations are required to fully define the scoring matrix. These constraints can ensure that the score possesses other desirable properties. Imposing symmetry (as well as equitability) on a 2-category scoring matrix, for example, is enough to completely constrain the matrix. Hence it can readily be shown that the only symmetric, equitable, 2-category scoring matrix (independent of $\{\tilde{p}_{vf}\}$) is the one shown in Table 3.

			Obs	
	Prob		p_1	p_2
	Cat		v	
			1	2
FC	f	1	$\frac{p_2}{p_1}$	-1
		2	-1	$\frac{p_1}{p_2}$

Table 3: Symmetric scoring matrix for the Peirce Skill Score.

For $\{\tilde{p}_v\} = \{p_v\}$ and $\tilde{p}_{11} + \tilde{p}_{12} = \tilde{p}_1 = p_1$, etc, the corresponding score can be written as

$$\tilde{S} = \frac{\tilde{p}_{11}}{p_1} - \frac{\tilde{p}_{21}}{p_2} \quad . \tag{3}$$

This is the Hit-Rate minus the False-Alarm-Rate (for large sample size) first defined by Peirce (1884). It is called here the Peirce Skill Score, although it has been named differently over the years (such as the "Hanssen-Kuipers Discriminant", the "Kuipers' Performance Index" and the "True Skill Statistic"). Unlike the Hit-Rate alone, the Peirce Skill Score does include a penalty for false alarms and cannot be artificially increased by over-predicting the event.

If the two observed categories have equal climatological probabilities ($p_1 = p_2$), then the diagonal elements of the Peirce Skill scoring matrix satisfy what will be called here the 'strong perfect forecast constraints':

$$\text{Strong Perfect FC:} \quad s_{vv} = 1 \quad \forall v \quad , \tag{4}$$

and the score for a perfect forecast will always be 1. Constraint (4) is stronger than the perfect forecast constraint in (2), which states only that the *expected* score should be 1. Satisfying (4), even in situations of unequal $\{p_v\}$, is found here to be a desirable attribute and will be discussed further.

Note that if the Peirce Skill Score is written in terms of the sample distribution: $\tilde{S} = \tilde{p}_{11}/\tilde{p}_1 - \tilde{p}_{21}/\tilde{p}_2$, the sample-mean score for a perfect forecast system would be 1, even if (4) is not satisfied. The disadvantage of defining a score by the sample distribution is that the scoring matrix is undefined for small samples and unstable unless the sample is sufficiently large. Therefore the use of the climatological distribution will be preferred here.

3.2.2 n-category scores

For a categorical score that assesses both the prediction of dry weather and precipitation quantity, more than 2 categories are required. Here the attributes of some established n-category scores are discussed.

A simple n-category score is the Heidke Skill Score (Heidke, 1926). This Score is based on the identity matrix, I_n, and therefore rewards a hit on any category equally and penalises all misses equally, regardless of the class of category error. The Heidke Skill scoring matrix for the 3-category score is shown in the form $(3I_3 - 1)/2$ in Table 4.

			Obs		
	Prob		p_1	p_2	p_3
	Cat			v	
			1	2	3
FC	f	1	1	$-\frac{1}{2}$	$-\frac{1}{2}$
		2	$-\frac{1}{2}$	1	$-\frac{1}{2}$
		3	$-\frac{1}{2}$	$-\frac{1}{2}$	1

Table 4: Scoring matrix for a 3*-category Heidke Skill Score.*

The Heidke Skill scoring matrix satisfies the strong perfect forecast constraint (4) and, for equi-probable categories ($p_v = \frac{1}{3}\ \forall v$ in the case of 3 categories), it also satisfies the equitability constraints (2).

Barnston (1992) modified the Heidke Skill Score, for equi-probable climatological categories, so that the penalty for an incorrect forecast was linearly dependent on the class of the category error. Barnston then made further adjustments to restore equitability. The 3-category scoring matrix is given in Table 5. It's dependence on the class of error is apparent although linearity and symmetry are compromised by the equitability adjustment. Note that the scoring matrix does not satisfy the strong perfect forecast constraint (4).

The LEPS approach of measuring error in 'probability space' (Ward and Folland, 1991) was introduced in section 1. The dotted lines in Fig. 2 show how the climatological cumulative distribution, P, is used to calculate this error. For example, if it rained 5.0mm at Balmoral, Belgium in October (Fig. 2d) when the forecast was for 2.5mm, then the linear error in probability space would be $f - v \equiv P(5.0) - P(2.5) = 0.82 - 0.73 = 0.09$. In general, the aim was a categorical score defined by the absolute linear error in probability space

$$s^L_{vf} = |f - v| \quad , \tag{5}$$

where v and f are the observed and forecast categories, respectively and L refers to 'LEPS'. After subsequent

			Obs		
	Prob		$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
	Cat			v	
			1	2	3
FC	f	1	$\frac{9}{8}$	0	$-\frac{9}{8}$
		2	$-\frac{3}{8}$	$\frac{3}{4}$	$-\frac{3}{8}$
		3	$-\frac{9}{8}$	0	$\frac{9}{8}$

Table 5: Scoring matrix for a 3-category Barnston Skill Score with equi-probable climatological categories.

adjustments including those for equitability, the scoring matrix for the 3-category LEPS Skill Score (Potts et al., 1996) with equi-probable climatological categories is given in Table 6. Notice that the final scoring matrix is not entirely linear.

			Obs		
	Prob		$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
	Cat			v	
			1	2	3
FC	f	1	$\frac{4}{3}$	$-\frac{1}{6}$	$-\frac{7}{6}$
		2	$-\frac{1}{6}$	$\frac{1}{3}$	$-\frac{1}{6}$
		3	$-\frac{7}{6}$	$-\frac{1}{6}$	$\frac{4}{3}$

Table 6: Scoring matrix for a 3-category LEPS Skill Score with equi-probable climatological categories.

It could be argued that the motivation behind the Barnston Skill Score was also to measure error in probability space and the main difference between the two scores is in the method by which equitability is achieved. (Potts et al., 1996) note that the LEPS score is 'doubly equitable' in that the equation

$$\text{Constant Obs:} \quad \sum_f p_f s_{vf} = 0 \quad \forall v \tag{6}$$

is also satisfied. This means that the expected skill score for constant observation is also 0. However, this is only realised in general for a model with no skill, but which still manages to produce a perfect distribution of categories ($\{q_f\} = \{p_v\}$). The apparent benefit of 'double equitability' is that the LEPS scoring matrix is symmetric although it does not satisfy the strong perfect forecast constraint (4). Note that, for daily precipitation, it is not possible to make categories equi-probable if 'dry' weather is to be defined as a category in itself.

Gerrity (1992) demonstrated how, for unequal probabilities, an equitable, symmetric, n-category score could be constructed from $n-1$ 2-category scores of the form of the Peirce Skill Score (Table 3). The method, which involves taking the arithmetic mean of these Peirce scores, effectively constrains all the remaining degrees of

freedom in the scoring matrix. The 3-category scoring matrix $\{s^G_{vf}\}$ for any $\{p_v\}$ is given by

$$\{s^G_{vf}\} = \frac{1}{2}\left\{\begin{array}{ccc} \frac{1-p_1}{p_1}+\frac{p_3}{1-p_3} & \frac{p_3}{1-p_3}-1 & -2 \\ \frac{p_3}{1-p_3}-1 & \frac{p_1}{1-p_1}+\frac{p_3}{1-p_3} & \frac{p_1}{1-p_1}-1 \\ -2 & \frac{p_1}{1-p_1}-1 & \frac{p_1}{1-p_1}+\frac{1-p_3}{p_3} \end{array}\right\}. \tag{7}$$

This score would allow 'dry' days to be defined as a single category. The scoring matrix in (7) for variable $\{p_v\}$ will be discussed later. Substituting $p_1 = p_2 = p_3 = \frac{1}{3}$ into (7) gives the 3-category scoring matrix for equi-probable climatological categories (Table 7). As with the Barnston and LEPS skill scores, the Gerrity scoring matrix does not satisfy the strong perfect forecast constraint (4).

			Obs		
	Prob		$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
	Cat		v		
			1	2	3
FC	f	1	$\frac{5}{4}$	$-\frac{1}{4}$	-1
		2	$-\frac{1}{4}$	$\frac{1}{2}$	$-\frac{1}{4}$
		3	-1	$-\frac{1}{4}$	$\frac{5}{4}$

Table 7: Scoring matrix for a 3*-category Gerrity Skill Score with equi-probable climatological categories.*

4 Stable Equitable Error in Probability Space

The desirable attributes of a score are dependent on the problem at hand. Barnston (1992) was interested in optimising comparability with the correlation score and in reducing sensitivity to the number of categories. Neither of these aspects seem particularly useful for monitoring progress or aiding decision-making in NWP. However, measuring error in probability space and ensuring equitability should help deliver some of the desirable attributes listed in section 1. To allow 'dry' weather to be a category in itself, the score must accommodate categories with variable probabilities. To aid the detection of performance changes, a score's sensitivity to sampling uncertainty needs to be minimised, while maintaining its ability to differentiate between good and bad forecasts. With the aim of minimising this sampling uncertainty, the stronger perfect forecast constraints will also be imposed. This is the starting point for the development here of a new categorical equitable score for verifying precipitation in NWP. Since the proposed score is based on error in probability space, it is formulated as an 'error score' rather than a 'skill score'.

The lack of complete linearity in the LEPS scoring matrix indicates that it is not possible to derive a completely linear score that is consistent with the equitability constraints. Hence a less constrained structure than (5) is initially proposed for an n-category error score. The first category represents 'dry' weather and the remaining categories represent equally-likely bins with successively heavier precipitation. In order to ensure that the score's error matrix, for a given location and time of year, is constant and well-defined (even for a single forecast) it is defined by the *climatological* cumulative distribution for each month. The first category thus

has climatological probability p_1. The other categories have probabilities $p_i = (1-p_1)/(n-1) \quad \forall i > 1$. The proposed structure is given by

$$s_{vf} = \begin{cases} |f-v|\,a & +\delta_{1f}(c-a) & \text{if } v > f \\ |f-v|\,b & +\delta_{v1}(d-b) & \text{if } v < f \\ 0 & & \text{if } v = f \end{cases} \quad , \tag{8}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. If $v > f$, the error increases linearly by a value $a > 0$ for each extra category separating the forecast, f, and verifying observation, v. The error increases by the value $b > 0$ if $v < f$. It is not imposed that a should be equal to b so this represents a form of 'semi-linearity' in probability space. Note that since p_1 is not necessarily equal to the other p_i, a different increment is used between categories 1 and 2. This increment is $c > 0$ if $v > f$ and $d > 0$ if $v < f$. Again, c and d are not specified to be equal. The hope was that this less constrained error matrix may be consistent with the equitability constraints. Note that (8) is consistent with the strong perfect forecast constraints for an error score,

$$\text{Strong Perfect FC (error):} \quad s_{vv} = 0 \quad \forall v \quad . \tag{9}$$

The equitability constraints for an error score (ignoring the redundant random constraint) can be written as

$$\begin{aligned} &\text{Perfect FC (error):} && \textstyle\sum_v p_v s_{vv} = 0 \\ &\text{Constant FC (error):} && \textstyle\sum_v p_v s_{vf} = 1 \quad \forall f \quad . \end{aligned} \tag{10}$$

The perfect forecast constraint in (10) is automatically satisfied because (9) is. However, it is not possible to satisfy the constant forecast constraints in (10) if $n > 3$. This is because the combination of constant forecast constraints: $(f=2) - 2(f=3) + (f=4)$ implies that $(a+b)p_3 = 0$ and this is not possible since $a, b, p_3 > 0$.

A 3-category score is possible (see below) and this will be the focus of the study. The structure of the error matrix for this score, consistent with (8), is given in Table 8. The climatological probability for the ('dry') category, $p_1 \in (0,1)$, will be dependent on location and month of year. The two remaining categories are termed here 'light precipitation' and 'heavy precipitation'. Their climatological probabilities, p_2 and p_3, respectively will define the precipitation threshold (in mm) between them (through application of the climatological cumulative distribution function).

			Obs		
	Prob		p_1	p_2	p_3
	Cat			v	
			1	2	3
FC	f	1	0	c	$c+a$
		2	d	0	a
		3	$d+b$	b	0

Table 8: Error matrix for a new 3-category score. Here p_1, p_2, p_3 represent the climatological probabilities of 'dry weather', 'light precipitation' and 'heavy precipitation', respectively, with $p_1 + p_2 + p_3 = 1$.

The 3 remaining (constant forecast) constraints in (10) are used to write b, c and d in terms of a:

$$\begin{aligned} b &= \frac{p_3 a}{1-p_3} \\ c &= \frac{1-p_3 a}{1-p_1} \\ d &= \frac{1-p_3 a}{p_1} \quad . \end{aligned} \tag{11}$$

Notice that, in general, $b \neq a$ and $c \neq d$. In addition, the initial concept of 'semi linearity' is no-longer evident when $n = 3$ although there is some clear consistency between error and probability differences. For example, in Table 8, $(s_{32} - s_{22}) = (s_{31} - s_{21}) = a$, both of which relate to the same difference in probability space between observed categories 3 and 2. Similarly, $(s_{13} - s_{23}) = (s_{12} - s_{22}) = d$, both of which relate to the difference in probability space between observed categories 1 and 2. There is also consistency in terms of differences in probability space between forecast categories: $(s_{21} - s_{22}) = (s_{31} - s_{32}) = c$ and $(s_{13} - s_{12}) = (s_{23} - s_{22}) = b$. Note that (in the case of 3 categories) this consistency is not contingent on constraining p_2 and p_3 to be equal, and so this is no-longer required. By increasing the ratio p_2/p_3 (discussed later), the threshold between 'light' and 'heavy' precipitation can be raised. This has the advantage of setting a harder challenge for the forecasting system but, in the limit, will lead to a 2-category score rather than a 3-category score.

Since b, c and d must all be greater than 0, (11) requires $0 < a < 1/p_3$. Which value of a is it best to use? It is worth examining the error matrices that would arise if a is allowed to take its extreme values. For $a = 0$ (upper error matrix in Table 9), a forecast for category 2 or 3 lead to the same score, regardless of the observed outcome. Moreover it is possible, with this error matrix, for a forecast system that only predicts categories 1 and 2 to obtain a perfect score. This means that there is a still a limit to how much the score can encourage refinement ($\{q_f\} \rightarrow \{p_v\}$, Murphy and Winkler, 1987). Similarly for $a = 1/p_3$ (lower error matrix in Table 9), there is no score difference whether category 1 or 2 is predicted. Hence a value for a strictly within the range $(0, 1/p_3)$ is required. Here, the optimal value for a is found by defining a 'refinement constraint' that maximises the lower-bound on the expected error for any forecast system that never predicts category 1 or never predicts category 3. Before deducing this value of a, it is worth noting that it is unnecessary to penalise a forecast system for never predicting category 2. Such a system would either predict the discontinuous categories 1 and 3, which is unrealistic for a dynamic model, or it would predict a single category which is already heavily penalised by equitability (10).

The lowest expected score for a system that never predicts category 3 (1) is achieved when it always correctly predicts the occurrence of categories 1 and 2 (2 and 3) and it additionally predicts category 2 on the fraction p_3 (p_1) of times that category 3 (1) occurs. This leads to an expected score of $p_3 a \quad (p_1 d = 1 - p_3 a$, by (11)). The lower-bound for the expected error for a 2-category system is therefore $\min(p_3 a, 1 - p_3 a)$, which is maximised at $\frac{1}{2}$ when $a = \frac{1}{2p_3}$. Choosing this value of a should rewarded a system for attempting to predict the full-range of possible outcomes. Using this value of a and the corresponding values of b, c and d (all at their mid-range values) the final error matrix for the new score, $\{s^S_{vf}\}$, is given by

$$\{s^S_{vf}\} = \frac{1}{2} \begin{Bmatrix} 0 & \frac{1}{1-p_1} & \frac{1}{p_3} + \frac{1}{1-p_1} \\ \frac{1}{p_1} & 0 & \frac{1}{p_3} \\ \frac{1}{p_1} + \frac{1}{1-p_3} & \frac{1}{1-p_3} & 0 \end{Bmatrix}. \tag{12}$$

			Obs		
	Prob		p_1	p_2	p_3
	Cat			v	
			1	2	3
FC	f	1	0	$\frac{1}{1-p_1}$	$\frac{1}{1-p_1}$
		2	$\frac{1}{p_1}$	0	0
		3	$\frac{1}{p_1}$	0	0
FC	f	1	0	0	$\frac{1}{p_3}$
		2	0	0	$\frac{1}{p_3}$
		3	$\frac{1}{1-p_3}$	$\frac{1}{1-p_3}$	0

Table 9: Error matrices for the two sets of extreme values of a, b, c and d as a function of p_1, p_2 and p_3, the climatological probabilities of 'dry' conditions, 'light precipitation' and 'heavy precipitation', respectively.

In anticipation of its reduced sensitivity to sampling error, this score will be called here 'Stable Equitable Error in Probability Space' (SEEPS).

5 Comparison with other scores

Here a comparison is made with the skill scores reviewed in section 3.2.2. As with SEEPS, these scores will be assumed here to be defined in terms of the climatological distribution $\{p_v\}$ rather than the sample distribution. The comparison requires a SEEPS 'skill-score'. This can readily be produced by calculating $1-$SEEPS, which clearly satisfies the equitability constraints (2). For 3 equi-probable categories (as is the case for Isle De Sable, Fig. 2c, with $p_2/p_3 = 1$, for example), the scoring matrix for the SEEPS Skill Score is given in Table 10.

			Obs		
	Prob		$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
	Cat			v	
			1	2	3
FC	f	1	1	$\frac{1}{4}$	$-\frac{5}{4}$
		2	$-\frac{1}{2}$	1	$-\frac{1}{2}$
		3	$-\frac{5}{4}$	$\frac{1}{4}$	1

Table 10: Error matrix for a 3-category SEEPS Skill Score with equi-probable climatological categories.

5.1 Refinement

The maximum skill possible for a forecast system that never predicts category 1 or never predicts category 3 can readily be calculated for the (equi-probable category) scoring matrices in tables 4, 5, 6, 7 and 10). Interestingly, this maximum is the same ($\frac{1}{2}$) for all scores. As with SEEPS, the 3-category Gerrity Skill Score has this maximum value for all $\{p_v\}$.

5.2 Uncertainty

If a score remains sensitive to sampling uncertainty as the expected skill score of the system approaches its upper-bound, then it will become increasingly difficult to detect further operational performance gains. Since SEEPS satisfies the strong perfect forecast constraint (9), it is insensitive to sampling uncertainty for a hypothetical perfect forecast system (unlike the Barnston, Gerrity and LEPS skill scores). Here the aim is to determine whether the strong perfect forecast constraint makes a material difference to sampling uncertainty for a less-than-perfect system. To do this, the standard deviation of each score is calculated as a function of expected skill.

To obtain a forecast system with variable skill, a conditional distribution $p_{v|f}$, the probability of verifying observation category v given a forecast for category f, is defined by

$$p_{v|f} = (1-\gamma)p_v + \gamma\delta_{vf} \quad , \tag{13}$$

where γ is a 'forecast system performance' parameter (see below). Note that the forecast distribution, $\{q_f\}$ is assumed to be the same as that of the observed climatology, $\{p_v\}$ and thus written as $\{p_f\}$. The definition of $p_{v|f}$ in (13) is consistent with this assumption since

$$\begin{aligned} \textstyle\sum_f p_f p_{v|f} &= \textstyle\sum_f p_f\left((1-\gamma)p_v + \gamma\delta_{vf}\right) \\ &= \left((1-\gamma)p_v \textstyle\sum_f p_f\right) + \gamma p_v \\ &= p_v \quad . \end{aligned} \tag{14}$$

The range of γ in (13) is $0 \leqslant \gamma \leqslant 1$. It can be seen that, for $\gamma = 0$, $p_{v|f} = p_v \; \forall (v,f)$ so that the forecast system is completely unskillful (Murphy and Winkler, 1987). For $\gamma = 1$, $p_{v|f} = 1$ if $v = f$ and $p_{v|f} = 0$ otherwise. This corresponds to a perfect forecast system.

Scores for this forecast system can be compared for the case of 3 equi-probable categories (Tables 4, 5, 6, 7 and 10). With the conditional distribution defined in (13), the expected skill for all these scores (indeed any equitable skill score satisfying (2)) is simply γ:

$$\begin{aligned} S &= \textstyle\sum_{v,f} p_{vf} s_{vf} \\ &= \textstyle\sum_{v,f} p_f p_{v|f} s_{vf} \\ &= \textstyle\sum_{v,f} p_f\left((1-\gamma)p_v + \gamma\delta_{vf}\right) s_{vf} \\ &= (1-\gamma)\textstyle\sum_f \left(p_f \sum_v p_v s_{vf}\right) + \gamma \sum_v p_v s_{vv} \\ &= (1-\gamma)\textstyle\sum_f \left(p_f \times 0\right) + \gamma \times 1 \\ &= \gamma \quad , \end{aligned} \tag{15}$$

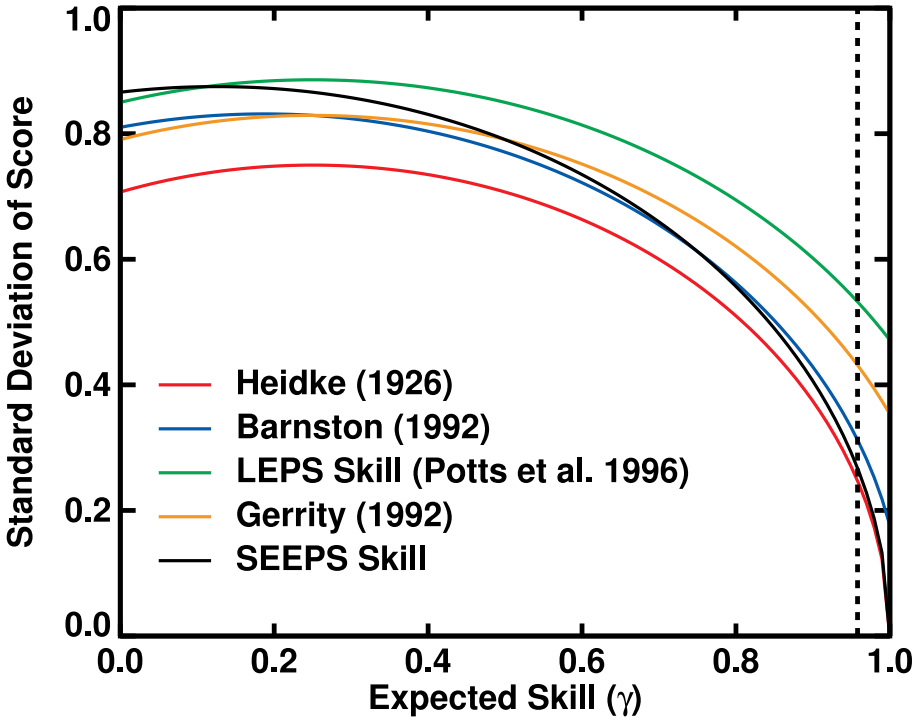


Figure 3: Expected standard deviation of a range of 3-category forecast scores as a function of expected skill, γ. Equi-probable categories are used for each score indicated in the key. The dashed line indicates the skill of a perfect forecast system that takes into account observation error and the lack of a grid-box's representativity of a point observation (as discussed in section 11).

where the equitability constraints (2) have been invoked in the penultimate line. The expected standard deviation of a equitable score with scoring matrix $\{s_{vf}\}$ can thus be written as

$$\sigma(\gamma) = \sqrt{\sum_{v,f} (\{s_{vf}\} - \gamma)^2 \, p_{v|f} \, p_f} \quad . \tag{16}$$

Figure 3 shows $\sigma(\gamma)$ for each score. (To obtain the standard deviation, and thus confidence intervals, of a sample-mean with sample-size n, simply divide σ by $\sqrt{n}$). The standard deviation of SEEPS is less than the standard deviation of the Gerrity Skill Score for $\gamma > \frac{1}{2}$. It is less than that of LEPS for $\gamma > \frac{1}{9}$ and less than that of the Barnston Skill Score for $\gamma > \frac{3}{4}$. It is never less than that of the Heidke Skill Score but this reflects the fact that the Heidke Skill Score does not differentiate between class-1 and class-2 category errors. Since the Gerrity Skill Score is the only other score defined and equitable for all $\{p_v\}$, it is the comparison with this score that is most relevant. Since present forecast skill is already better than $\frac{1}{2}$ at short lead-times (see later), SEEPS must be considered preferable. The higher standard deviation of SEEPS for $\gamma < \frac{1}{2}$ is less relevant and will become even more so in future.

The standard deviations of the SEEPS and Gerrity skill scores are plotting in Fig. 4 as a function of (γ, p_1) for $p_2/p_3 = 1$. As the system's performance improves, it can be seen that the Gerrity Skill Score becomes even more sensitive to sampling uncertainty when p_1 diverges from $\frac{1}{3}$. Hence the Gerrity Skill Score is likely to be unstable for skillful systems and less able to detect operational trends.

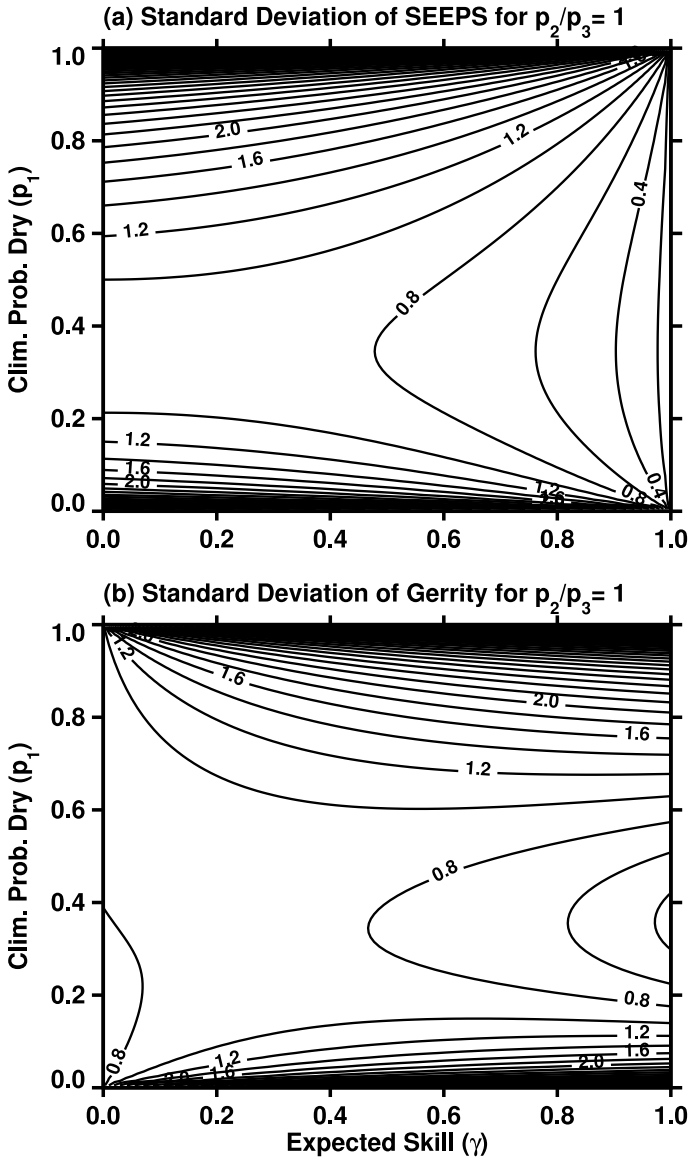


Figure 4: Standard deviation of scores as a function of (γ, p_1) for $p_2/p_3 = 1$. (a) SEEPS. (b) Gerrity Skill Score.

5.2.1 *Relationship between SEEPS and Gerrity scores*

Comparing tables 7 and 10, it can be seen that the columns of the Gerrity and SEEPS skill scores, for equiprobable categories, differ only by a constant (dependent only on v). By comparing (7) and (12) it can readily be shown that this is true in general (for all $\{p_v\}$) so that

$$\begin{aligned} s^G_{vf} &= (1-s^S_{vf}) + \lambda(v) \qquad \forall v, f \\ \tilde{S}^G &= \tilde{S}^S + \textstyle\sum_v \tilde{p}_v \lambda(v) \\ &\equiv \tilde{S}^S + \Lambda(\{\tilde{p}_v\}) \quad , \end{aligned} \tag{17}$$

where $\tilde{S}^G$ and $\tilde{S}^S$ are sample-mean Gerrity and SEEPS skill scores, respectively. Equation (17) implies that both skill scores respond identically to the forecast system's performance and only differ by a term Λ, dependent on the observed sample distribution $\{\tilde{p}_v\}$. For a perfect forecast system, $\tilde{S}^S = 1$ and $\sigma^S = 0$ so the Gerrity Skill Score includes all the sampling uncertainty associated with $\Lambda(\{\tilde{p}_v\})$. Consistent with the result above for equiprobable categories, it can be shown that, for any $\{p_v\}$, the Gerrity Skill Score is more sensitive to sampling uncertainty than SEEPS when assessing systems which have expected skill $> \frac{1}{2}$. SEEPS is more sensitive when the expected skill is $< \frac{1}{2}$.

It is interesting to discover that the derivation here of SEEPS, that addresses key requirements for the monitoring of precipitation forecasts, produces a score so similar to the rather elegantly constructed Gerrity Skill Score. Indeed, both skill scores have identical expected values (since $\Lambda(\{p_v\}) = 0$). Their only difference, however, is important and renders SEEPS more stable for assessing forecast systems with sufficient expected skill.

The SEEPS error matrix can (also) be derived as the mean of two 2-category error matrices of the form shown in Table 11. If defined by the sample distribution, the sample-mean error based on this matrix is identical to (1−) the sample-mean Peirce Skill Score. Hence both tables 3 and (1−) 11 can be considered as valid choices for representing the linear equitable scoring matrix for the Peirce Skill Score. After imposing equitability, Gerrity (1992) constrained the last degree of freedom by requiring symmetry. Here it has be demonstrated that symmetry is not a useful attribute for the scoring matrix and the last degree of freedom is, instead, constrained by requiring that all perfect forecasts have zero error. It is possible that the error matrix in Table 11 could be used to define a series of n-category scores with reduced sensitivity to sampling uncertainty, although not possessing the structure originally proposed in (8) for $n > 3$.

			Obs	
	Prob		p_1	p_2
	Cat		v	
			1	2
FC	f	1	0	$\frac{1}{p_2}$
		2	$\frac{1}{p_1}$	0

Table 11: The 2-category equitable error matrix for a score that SEEPS can be built from.

The comparison of uncertainty in section 5.2 is valid for comparing the scores' abilities to detect operational performance trends, but not for assessing their abilities to detect performance differences when two forecast systems are used to predict the same set of observations. This is clear from (17) since taking a difference will

eliminate the $\Lambda(\{\tilde{p}_v\})$ term. For equi-probable categories, the Barnston Skill Score has a similar relationship to SEEPS as in (17). LEPS, however, is more uncertain than these scores for all γ.

5.3 Hedging

Hedging is said to have occurred whenever a forecaster's judgement and forecast differ (Murphy, 1978). In the context of system development, the prevention of hedging should mean that there is always a physical basis for any change in a forecasting system - so that 'judgement' and forecast both change in unison. Changes in a forecast system alter the joint distribution $\{p_{vf}\}$ $(=\{p_{v|f}q_f\})$. A score will inhibit hedging if it cannot be improved by making changes to $\{p_{vf}\}$ in the absence of additional physical insight. Changes to $\{p_{vf}\}$ can be broken-down into a number of steps in which a fraction of forecasts for one given category, f_1, are changed to another category, f_2. Hence, to determine if a score can be hedged, it is only necessary to assess whether it can be improved by making a single such step. Fundamental to the hedging assessment is the recognition that, in the absence of physical insight, it is not possible to choose which forecasts for category f_1 will be changed and so those changed must have the distribution of the original system; $\{p_{v|f_1}\}$. Using (12), the change in SEEPS error that occurs when a fraction $\delta q/q_1$ (>0) of the forecasts for category 1 are changed to category 2 ('1→2') is given by

$$\begin{aligned}
\delta\text{SEEPS} &= \delta q \textstyle\sum_v p_{v|f=1}(s_{v2}-s_{v1}) \\
&= \frac{\delta q}{2}\left(\frac{p_{v=1|f=1}}{p_1}-\frac{p_{v=2|f=1}+p_{v=3|f=1}}{1-p_1}\right) \\
&= \frac{\delta q}{2}\left(\frac{p_{v=1|f=1}}{p_1}-\frac{1-p_{v=1|f=1}}{1-p_1}\right) \\
&= \frac{\delta q}{2p_1(1-p_1)}(p_{v=1|f=1}-p_1) \quad ,
\end{aligned} \tag{18}$$

Hence SEEPS is reduced only if $p_{v=1|f=1} < p_1$, and thus only if the original forecast system is very poor in terms of its prediction of category 1. Indeed, the likelihood of dry weather, given that dry weather is forecast, would have to be less than the climatological chance of dry weather. Using similar mathematics, the change 2→1 only decreases the SEEPS if $p_{v=1|f=2} > p_1$. Again, this would only be true for a particularly bad forecast system for which it were more likely to be dry when the forecast predicted light rain than the climatological chance of dry weather. Similarly, 3→2 only reduces the SEEPS if $p_{v=3|f=3} < p_3$ and 2→3 only reduces the SEEPS if $p_{v=3|f=2} > p_3$. Changes between non-adjacent categories (1→3 and 3→1) are less plausible for a dynamical forecast model. Ignoring this possibility, it has therefore been shown that SEEPS can only be hedged if the forecast system is very poor in the first place. Note that this result is true for all $\{p_v\}$. The result is also independent of the refinement constraint (similar mathematics holds for any a with $0 < a < 1/p_3$).

The relationships, present in the SEEPS error matrix, $(s_{21}-s_{22}) = (s_{31}-s_{32})$ and $(s_{13}-s_{12}) = (s_{23}-s_{22})$ are sufficient for this inhibition of hedging. For the same reason, the 3-category Gerrity Skill Score cannot be hedged for any $\{p_v\}$ and, for $p_1 = p_2 = p_3 = \frac{1}{3}$, neither can the Barnston Skill Score. (These results are also clear from the equivalence of expected SEEPS, Gerrity and Barnston skill scores). Numerical experimentation (for $p_1 = p_2 = p_3 = \frac{1}{3}$) shows that SEEPS and these two scores cannot be hedged, even when non-adjacent changes are included, if the conditional distribution is constrained by $p_{v|v} \geqslant p_v$ $\forall v$ and $p_{v|f} \leqslant p_v$ $\forall f \neq v$. However, LEPS can be hedged even under these constraints.

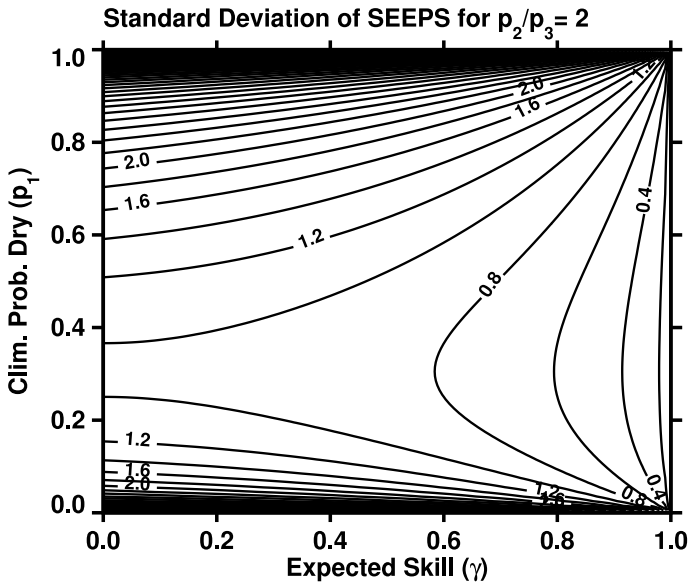


Figure 5: As Fig. 4a but for $p_2/p_3 = 2$.

6 SEEPS parameter settings

Equation (12) shows that, as the probability of 'dry' weather p_1 (or 'wet' weather $1-p_1$) gets close to 0, elements of the SEEPS error matrix become extreme (because they involve reciprocals of p_1 and $1-p_1$). This necessitates the need for bounds on the acceptable range of p_1. There is also a need to define p_2/p_3. Figure 2 shows how the threshold between 'light' and 'heavy' precipitation rises when p_2/p_3 is increased from 1 to 2. For Grenoble, France in June (Fig. 2e), for example, the threshold increases from 2.4 to 6.0 mm. The higher thresholds would set a more challenging task for a forecasting system if they do not greatly increase sensitivity to sampling uncertainty.

Using the conditional distribution (13), Fig. 4(a) showed the standard deviation of SEEPS as a function of (γ, p_1) for $p_2/p_3 = 1$. Uncertainty increases sharply for extreme values of p_1 and the limiting range $p_1 \in [0.10, 0.85]$ is suggested. Precipitation in more arid climates is effectively considered as 'extreme weather' and neglected to reduce uncertainty in area-mean scores. Note that no (trustworthy) SYNOP station has a climatology with $p_1 < 0.10$. The benefits of $p_2/p_3 = 2$ are considered important enough to sacrifice a small increase in uncertainty (*c.f.* Fig. 4a and Fig. 5). Unless otherwise specified, these are the settings used below. Section 10.1 tabulates real forecast results that tend to confirm these choices.

7 SEEPS: Summary of the score

Table 12 shows SEEPS error matrices for a set of climate regimes where the probability of a 'dry' day (p_1) varies within its desired range [0.10,0.85] and 'light' precipitation is defined to occur twice as often as 'heavy' precipitation ($p_2/p_3 = 2$).

These error matrices are asymmetric. In particular, a forecast for a climatologically likely category which turns out to be incorrect is penalised more heavily than a forecast for an unlikely category which turns out to be incorrect. This is a desirable attribute since it should improve a system's 'discrimination' ($\{p_{f|v}\}$, Murphy and Winkler, 1987; Murphy, 1993) by encouraging developments that allow the model (physics) to

		Obs		
		dry	light	heavy
	prob	**0.10**	0.60	0.30
	dry	0.00	0.56	2.22
FC	light	5.00	0.00	1.67
	heavy	5.71	0.71	0.00
	prob	**0.33**	0.44	0.22
	dry	0.00	0.75	3.00
FC	light	1.50	0.00	2.25
	heavy	2.14	0.64	0.00
	prob	**0.50**	0.33	0.17
	dry	0.00	1.00	4.00
FC	light	1.00	0.00	3.00
	heavy	1.60	0.60	0.00
	prob	**0.67**	0.22	0.11
	dry	0.00	1.50	6.00
FC	light	0.75	0.00	4.50
	heavy	1.31	0.56	0.00
	prob	**0.85**	0.10	0.05
	dry	0.00	3.33	13.33
FC	light	0.59	0.00	10.00
	heavy	1.11	0.53	0.00

Table 12: SEEPS error matrices for a range of dry day probabilities (indicated in bold type) and with the probability of 'light precipitation' being double that of 'heavy precipitation'.

represent all categories, whatever their climatological frequency.

For European stations, p_1 is shown in Fig. 6(a) and (b) for January and July, respectively. As would be expected, summer has more 'dry' days than winter. Northwestern Europe has the fewest 'dry' days throughout the year. Southern Europe in high summer (July and August) is particularly arid with probabilities of a 'dry' day in excess of 0.85.

The threshold (in mm) between the 'light' and 'heavy' precipitation categories is shown in Fig. 6(c) and (d) for January and July, respectively. For Europe, the threshold between the 'light' and 'heavy' precipitation categories is generally between 3mm and 10mm, but can be higher over mountainous regions such as the Alps. Hence the category known as 'heavy precipitation' also incorporates what may be considered to be more 'moderate' events.

By adapting to the underlying climate, SEEPS assesses the pertinent aspects of the local weather. By encouraging refinement and inhibiting hedging, it should provide useful guidance for development decisions.

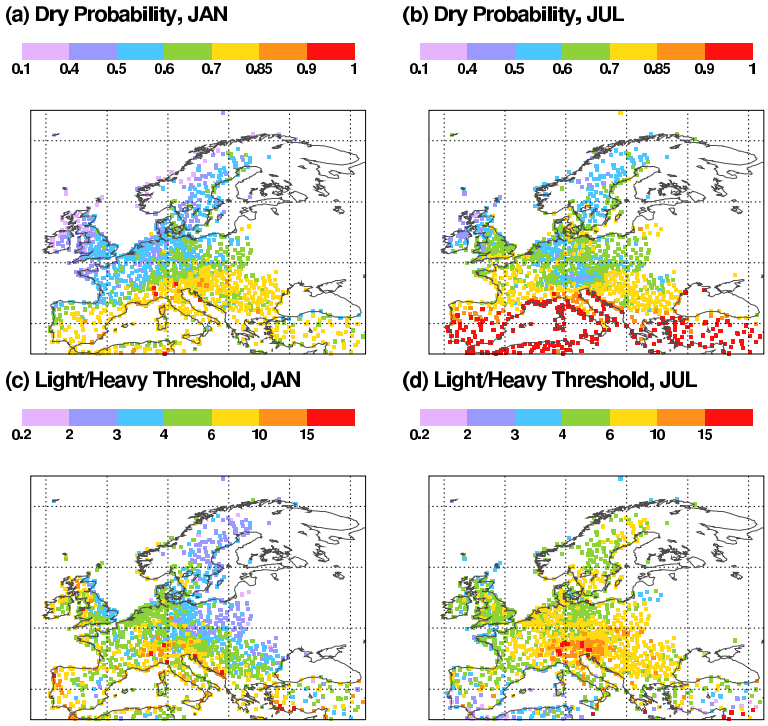


Figure 6: (a) Probability of a 'dry' day for January (b) As (a) but for July. (c) Precipitation amount (in mm) marking the threshold between 'light' and 'heavy' precipitation for January. (d) As (c) but for July. By definition, 'light precipitation' occurs twice as often as 'heavy precipitation'. Results are based on 24-hour precipitation accumulations (12UTC–12UTC) from the 1980–2008 climatology.

8 Case studies: Precipitation errors identified by SEEPS

Before attempting to diagnose trends in area-mean SEEPS scores, it is worth demonstrating some of the precipitation errors that the SEEPS score can identify. Improvements in such errors will, therefore, be reflected in reductions in the SEEPS score.

Fig. 7(a) shows observed 24-hour accumulated precipitation (in mm) on 16 December 2008, and Fig. 7(b) shows the corresponding D+4 forecast precipitation. (D+4 is chosen because of ECMWF's mandate to improve medium-range forecasts). Notice that large parts of northern Europe were predicted to have drizzle but were actually 'dry' (pink). In this case, recorded values were 0.0mm, rather than 0.1 or 0.2 mm. Since this region is generally wet in December (Fig. 7c) and an incorrect forecast for a likely category is strongly penalised, the differences in precipitation categories (*c.f.* Fig. 7d and e) lead to relatively large SEEPS scores (Fig. 7f). This partly explains why the mean European score for this forecast was one of the worst in 2008. Verification at the dry/wet boundary has important physical significance because of the existence of positive feedbacks with latent heating. From the users' perspective, of course, drizzle is also of great relevance. Hence it is desirable that SEEPS can highlight this error.

Note that the station scores in Fig. 7(f) are plotted with variable sizes to indicate their relative weight within an area-mean score. These weights, which depend on the local station network density, are explained in section 9.1.

Another poor European-mean SEEPS score occurred on 23 August 2008. This is the situation presented in Fig. 1. The SYNOP observations (Fig. 8a) show northeast Europe received over 10mm, and up to 57mm, of precipitation associated with a Low centred over Germany. The D+4 forecast (Fig. 8b) had less than 5mm

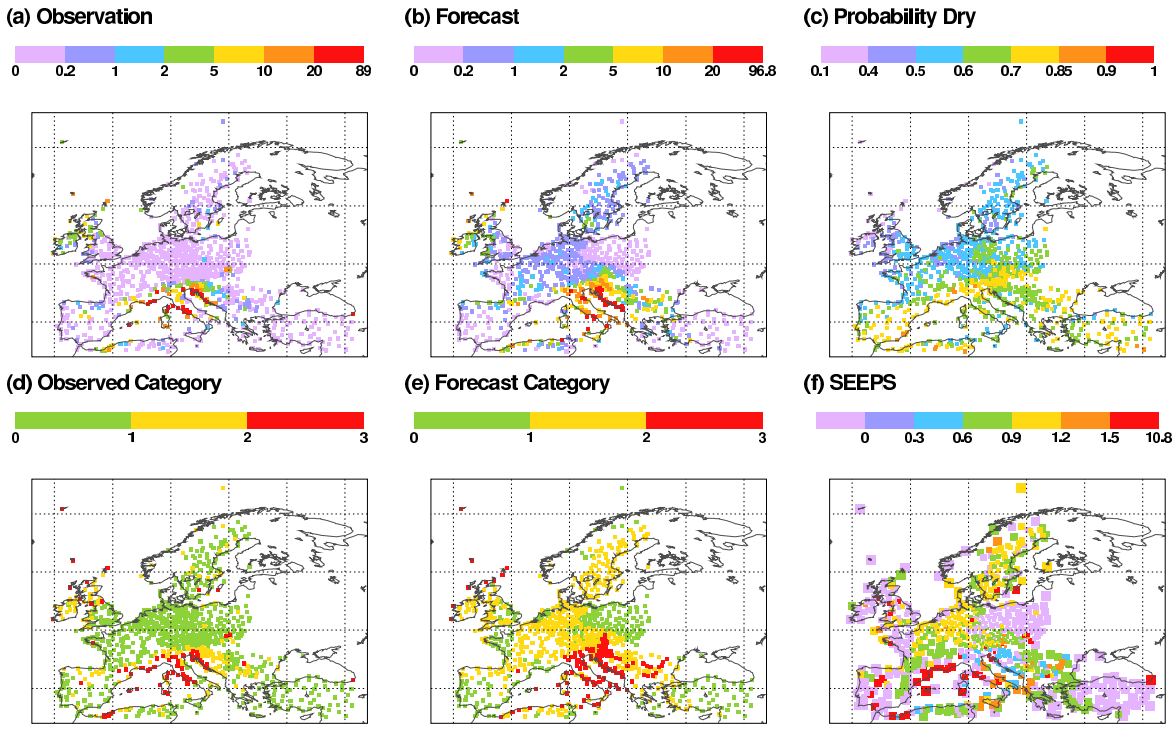


Figure 7: (a) Observed precipitation accumulated over 24 hours 2008/12/15 12UTC to 2008/12/16 12UTC. (b) Forecast precipitation accumulated over leadtimes 72 to 96 hours and valid for the same period as the observations. (c) Probability of a 'dry' day in December, based on the 1980–2008 climatology. (d) Observed precipitation category. (e) Forecast precipitation category. (f) SEEPS. Units in (a) and (b) are mm. Squares in (f) are plotted with areas proportional to the weight given to each station in the area-mean score.

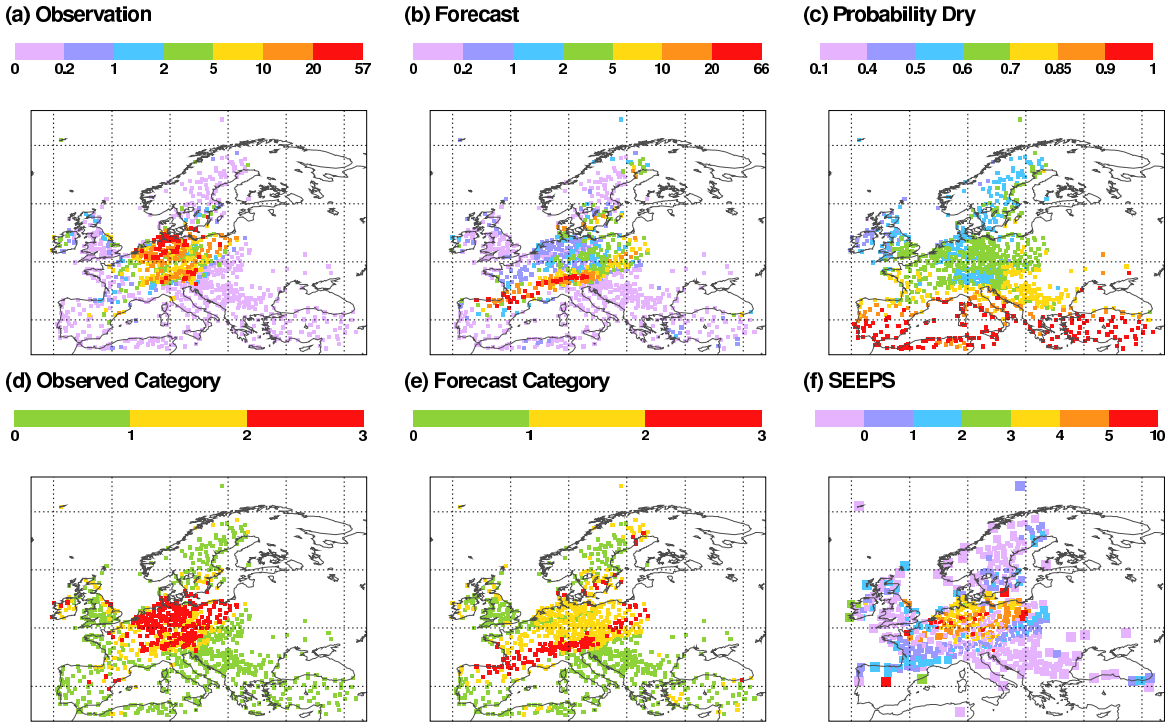


Figure 8: As Fig. 7 but for the prediction of precipitation accumulated over the 24 hours 2008/08/22 12UTC to 2008/08/23 12UTC.

(often less than 1mm) in this region and, instead, predicted convective out-breaks along a front to the south. SEEPS identifies these errors (Fig. 8f) but, since even northern Europe is generally dry in August (Fig. 8c), it is the category differences (*c.f.* Fig. 8c and d) indicating under-prediction that lead to the largest scores (Fig. 8f).

Note that no SEEPS scores are plotted in Fig. 8(f) for the southern Iberian peninsular, northern Africa, and Turkey. This is the unfortunate consequence of avoiding arid climates by insisting that $p_1 \in [0.10, 0.85]$.

The final example of a particularly poor European-mean SEEPS score is that of 9 June 2008 (Fig. 9). This case demonstrates that SEEPS can highlight the mis-location of summertime convection over southern Europe. Although it will be difficult to improve such errors at D+4, it may be possible at shorter lead-times through better forecast initialisation, better model physics and higher resolution.

9 Area-mean scores

9.1 Taking account of station network density

SYNOP stations are not evenly spaced-out over the globe. When area-mean scores are required, it is useful to take the station network density into account in order to prevent sub-regions with high station density dominating the score. Following a methodology used in other areas of meteorology and elsewhere, the station density, ρ_k, in the vicinity of station k is calculated by applying a Gaussian kernel to the network:

$$\rho_k = \sum_l e^{-(\alpha_{kl}/\alpha_0)^2} \quad , \tag{19}$$

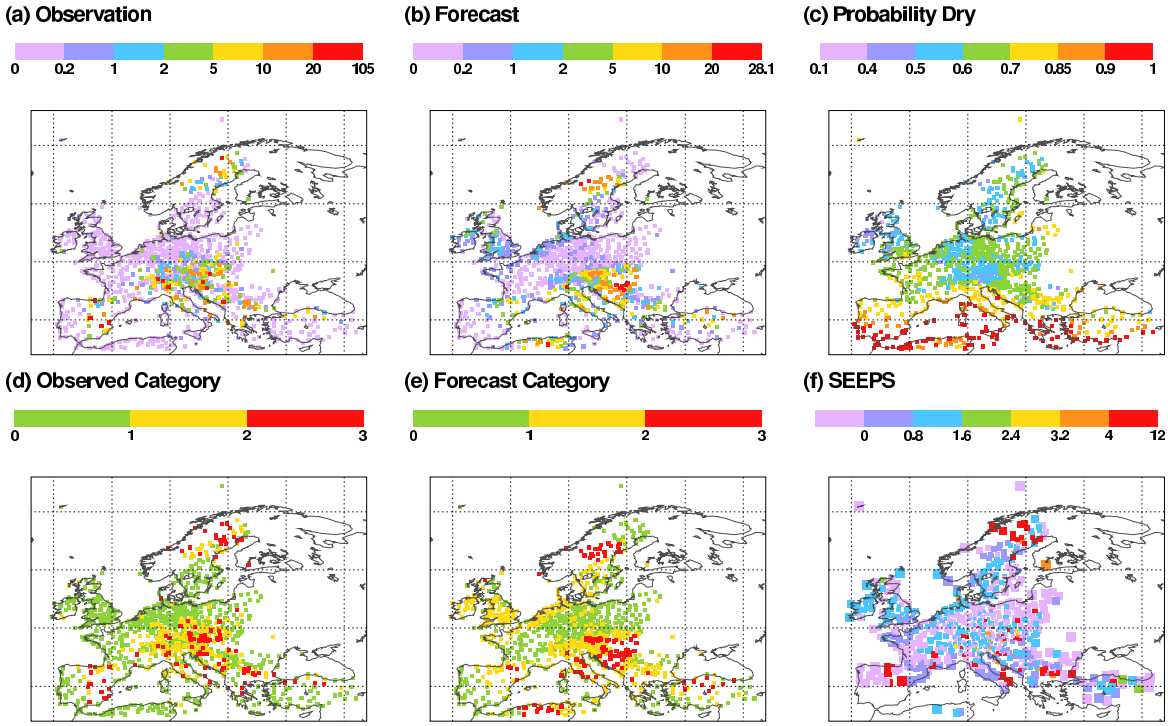


Figure 9: As Fig. 7 but for the prediction of precipitation accumulated over the 24 hours 2008/06/08 12UTC to 2008/06/09 12UTC.

where $\sum_l$ is over all the stations used in the score (on the particular day in question), α_{kl} is the angle subtended at the centre of the Earth between stations k and l, and α_0 is a reference angle. (For each k, stations, l, for which $\alpha_{kl} > 4\alpha_0$ have negligible contribution and are disregarded). Since $\alpha_{kk} = 0$, we have that $\rho_k \geqslant 1 \forall k$. The value of $\alpha_0 = 0.75^o$ (83km) is chosen because it is the smallest possible that ensures approximate equal representation of all sub-regions of Europe.

Writing S_k for the (unweighted) SEEPS score for station k, then the weighting applied to this station, w_k, and the weighted area-mean score, S, are defined by:

$$\begin{aligned} w_k &= \frac{1}{\rho_k} \\ S &= \frac{\sum_k w_k S_k}{\sum_k w_k} \quad . \end{aligned} \tag{20}$$

As discussed in section 8, the areas of the squares in Fig. 7(f), Fig. 8(f) and Fig. 9(f) are proportional to the weights applied to each station. The fact that Europe is reasonably evenly covered with colour demonstrates that, with this density weighting, no sub-region is favoured over any other. Density weighting also ensures that Europe will not dominate so heavily a score of the Extratropics in general. The methodology is currently being developed to conglomerate observations reported at all times of the day. This will mean that, for example, eastern Europe will be much better represented in area-mean scores than indicated in (*e.g.*) Fig. 7(f).

Weighting could also help reduce sampling uncertainty for area-mean scores associated with the spatial correlation of precipitation (and thus scores) in high network-density areas.

By construction, there is an upper limit to the weight any individual station can have. This ensures, for example, that island and coastal stations do not have undue influence on the score.

9.2 The Extratropics

Area-mean scores have been produced, taking the station network density into account, for the period 1995-2008. Plots for the Extratropics (north of 30^oN *and* south of 30^oS), based on ∼2000 stations per day, are shown in Fig. 10. Figure 10(a) shows the annual mean scores based on the 12UTC operational forecasts as a function of leadtime. The colours indicate the years. There is a general progression to lower errors over these 14 years. The black curve shows the most recent year (2008).

The 70% confidence intervals plotted in Fig. 10(a) show the degree of uncertainty in the annual means. They are deduced from the daily scores taking autocorrelation into account following the methodology of von Storch and Zwiers (2001). If one mean lies within the confidence interval of another, then there is no significant difference. If confidence intervals just touch, then mean scores are significantly different at the 14% level, assuming equal variances. It can be seen that it is generally not possible in year y to demonstrate that forecasts are better than in the previous year $y-1$: it takes a few years for improvements to become unequivocal.

Although there have been clear improvements, forecasts are still far from perfect. At D+1 (which is the score for the precipitation accumulated over the first day of the forecast), errors are above 0.4 (skill below 0.6), even for 2008. The poor scores at D+1 indicate that short-range forecasts (like that shown in Fig. 1a) cannot be considered as reliable observations at present. Nevertheless, current SEEPS skill scores for D+1 and D+2 are greater than the critical value of $\frac{1}{2}$ required for SEEPS' sensitivity to sampling uncertainty to be less than that of the Gerrity Skill Score (see Fig. 3 and Fig. 4) and for the refinement constraint (section 4) so benefit development decisions.

It can be seen that by D+10, the SEEPS score is tending towards 1. This is one of the desirable features associated with equitability: by construction, expected SEEPS scores for all stations and all months of the year lie between 0 and 1 and this makes the aggregation of all the stations within an area a meaningful and useful concept (despite sub-regions having very different climates).

Fig. 10(b) shows (light green) daily SEEPS scores at D+4 for the same operational forecasts. The general improvement over the years is clearly apparent when a 365-day running mean is applied (black). The 31-day running mean (dark green) highlights a seasonal cycle in SEEPS scores. This feature is common to many precipitation scores and reflects the fact that large-scale precipitation (in winter) is generally easier to predict than convective precipitation (in summer). (Note that the vast majority of the stations used each day are in the Northern Hemisphere and weighting is not sufficient to accord equal influence to the southern extratropical observations).

Fig. 10(c) shows the annual-mean of the leadtime at which the SEEPS score for each daily forecast first reaches a value of 0.6. The value of 0.6 was chosen because it corresponds approximately to the present score at D+4. The red curve relates to the operational forecast data shown in Fig. 10(a) and (b). The gains in leadtime amount to ∼2 days over the 14-year period. The graph is annotated to show when the model's (spectral) resolution was changed during this period and also to show when one key model cycle (25R4) was introduced. This model cycle had many updates that could have directly affected the forecast of precipitation. However, there were 40 packages of updates applied to the operational data assimilation and forecasting system over this period and many of these will have contributed to the improvement.

The blue curve in Fig. 10(c) shows comparable results for forecasts made within the ERA-Interim re-analysis project. ERA-Interim is based on a single model cycle (31R2) and a single model resolution (T255). The date that this cycle was first used in the operational forecast system (12 December 2006) is also indicated on the graph. The differences between the red and blue curves at this date highlight the impact of resolution. The flatness of the ERA-Interim SEEPS curve is striking. It indicates that inevitable changes over the years to the network of SYNOP stations has not had a major impact on scores. More controversially, it also indicates that the

increase in available sources and volume of data used to initialise the forecast (a 100× increase over this period) has had almost no lasting impact on the prediction of precipitation. Instead, the lasting improvements in the extratropical operational scores must be due to improvements to model physics, increases in model resolution, and to the way the data assimilation system has improved to better use the available observations. New data sources will target more directly the hydrological cycle so the conclusions from the 1995–2008 period may not hold in future.

9.3 Europe

The SEEPS timeseries for Europe [12.5^{o}W–42.5^{o}E, 35^{o}N–75^{o}N] at D+4 (Figure 11a) show a similar improvement to that of the Extratropics, but with more variability (for comparison, the plot has the same axes as Fig. 10b and thus daily scores often extend outside the region shown). There is an oscillation in the 1-year running-mean score around 2003. This is also apparent, but less prominent, in the extratropical timeseries (Fig. 10b). Since ERA-Interim results display also this oscillation (not shown), it is not associated with changes in model cycle or resolution. Instead it is an artifact of the flow itself. From close inspection of Fig. 11(a), it would appear that the dry weather during European summer heatwave of 2003 was anomalously easy to predict and that the precipitation in the preceeding year was anomalously hard to predict.

9.4 South America

The SEEPS scores for the South American region [70^{o}W–35^{o}W, 40^{o}S–10^{o}N] at D+4 (Fig. 11b) show an improving trend although with a lot of variability. Close inspection of the data reveals an alarming seasonal cycle in the number of precipitation observations used in the score. Up to 200 observations are used during the wet season but as few as 50 are used during the dry season. It is possible that this is due to non-reporting of zero rain. The small sample size leads to more uncertainty and this should be taken into account when making development decisions.

10 Detecting improvements

10.1 Trends in operational forecasts: Sensitivity to SEEPS parameter settings

The confidence intervals in Fig. 10(a) indicate that a few years are required before improvements are detectable above the level of sampling uncertainty. Here the choice of bounds for p_1 and the value of p_2/p_3 are assessed in relation to SEEPS' ability to detect improvements. Since the improving trends in Figs. 10(b) and 11(a) appear to be quite linear, this ability-to-detect is estimated by dividing the linear trend by the standard deviation of departures (of the 1-year mean curve) from it. Table 13 shows 'Trend/StDev' at D+4 for the Extratropics and Europe. The smaller sampling uncertainty associated with the larger, extratropical, region makes trends easier to detect.

The results tend to confirm the choices made in section 6 (shown in bold in Table 13). The higher threshold between 'light' and 'heavy' precipitation usefully sets a harder forecasting challenge with only a slight deterioration in ability to detect extratropical trends. Additionally increasing the upper-bound on p_1 to 0.90 permits the use of very few extra stations in arid climates (for Europe, those coloured orange in Fig. 6a and b) with a more marked deterioration in ability-to-detect extratropical trends.

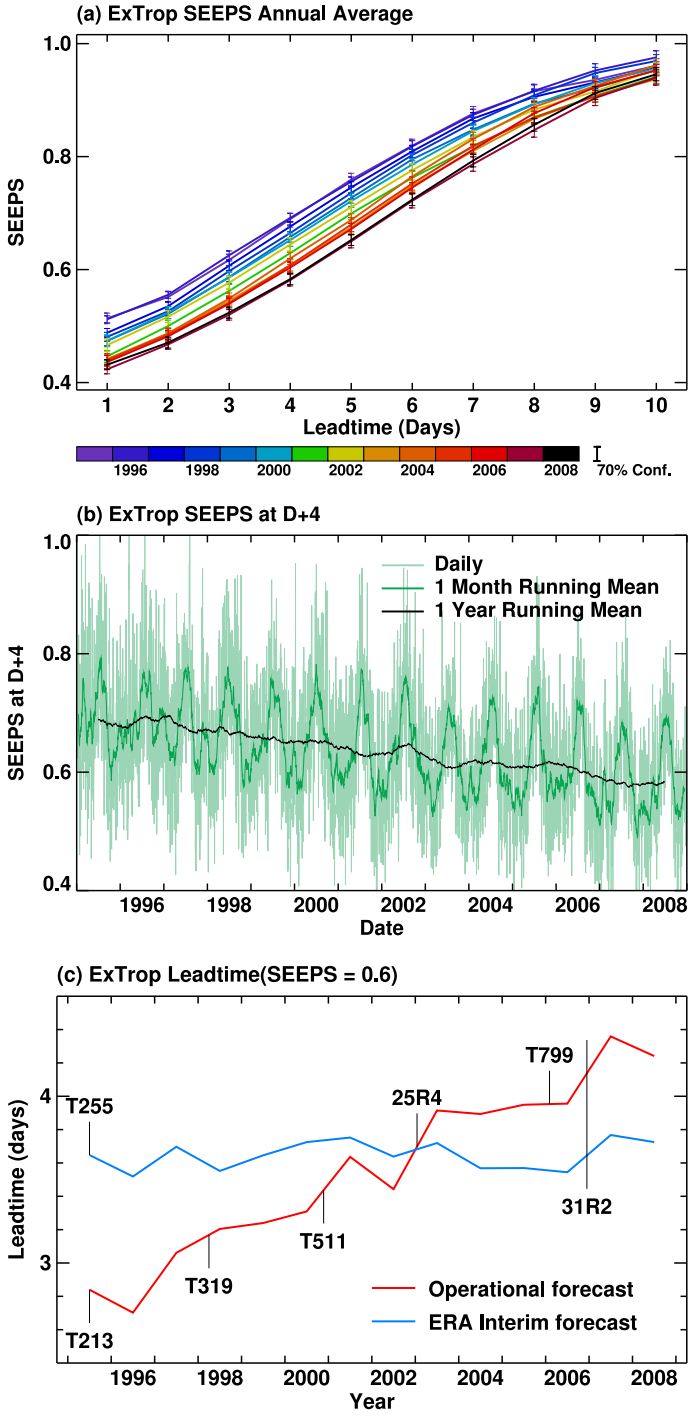


Figure 10: Extratropical-mean SEEPS results. (a) Annual-mean of daily operational scores as a function of lead-time. 70% confidence intervals for these annual means are indicated. (b) Timeseries of operational scores at D+4 with running means as indicated. (c) Annual-mean lead-time at which the score rises to 0.6 based on the operational forecasts and on the forecasts made during the production of the ERA-Interim re-analysis, as indicated. The extratropical average is over the combined region north of 30^{o}N and south of 30^{o}S, taking account of observation density.

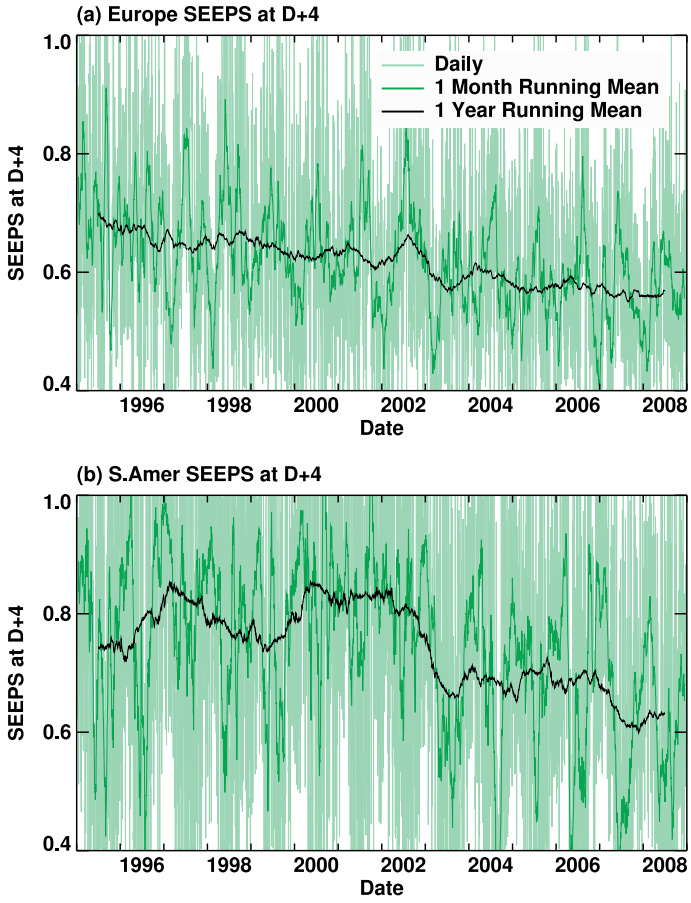


Figure 11: As Fig. 10(b) but for (a) Europe, [12.5°W–42.5°E, 35°N–75°N] and (b) South America, [70°W–35°W, 40°S–10°N].

Probabilities		Trend/StDev (yr^{-1})	
Dry	Light / Heavy	ExTrop	Europe
[0.10,0.85]	1	-1.31	-0.88
[0.10,0.85]	**2**	-1.25	-0.70
[0.10,0.90]	1	-1.23	-0.80
[0.10,0.90]	2	-1.10	-0.65
[0.10,0.95]	1	-1.13	-0.63
[0.10,0.95]	2	-0.95	-0.52

Table 13: Ability to detect trends in operational performance, and its sensitivity to SEEPS parameter settings. Values are based on daily forecasts for the years 1995–2008.

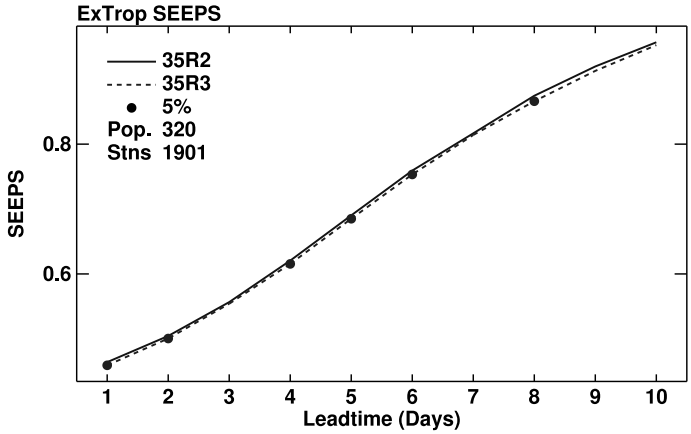


Figure 12: Mean extratropical SEEPS scores for two cycles of the ECMWF forecasting system as a function of lead-time: 35R2 (solid) and 35R3 (dashed). A filled circle on a given curve indicates that the mean score for that model cycle is statistically significantly better than that of the other cycle, at the 5% level using a 2-sided, paired Student's t-test, taking autocorrelation into account. Results for both cycles are based on all 320 forecasts initiated at 0 and 12UTC between 2009/04/01 at 12UTC and 2009/09/08 at 0UTC. On average, 1901 extratropical station observations are used in the score on any given day. The Extratropics are defined as everywhere north of 30°N combined with everywhere south of 30°S.

10.2 Differences between forecast system cycles

When an experimental forecast suite (or 'cycle') is being assessed, a set of forecasts are compared with those of the operational system, using the same set of start dates. Sampling uncertainty is greatly reduced by using the same start dates but it is not completely eliminated. Hence the optimisation of SEEPS parameters is still relevant. Figure 12 shows a comparison of extratropical SEEPS scores for two consecutive ECMWF forecast cycles (35R2 and 35R3) based on 320 start dates. The newer cycle (dashed) is better than the older cycle (solid) at all lead-times. It is statistically significantly better at the 5% level (indicated by the filled circles) for six of these lead-times. Clearly it is much easier to detect incremental improvements to the forecast system using these parallel experimental suite tests than from the operational forecasts alone. With these tests, SEEPS should provide useful information on which to make developmental decisions.

11 Observation error and representativeness

The SYNOP precipitation observations contain errors and they are also not necessarily representative of any grid-box average produced by a forecast model. These issues impose a non-zero lower limit on the SEEPS score, that even a perfect forecasting system can never surpass. The impact is likely to be ameliorated by verifying 24-hour accumulations, using the nearest grid-point for matching model data to observations, and measuring forecast error in probability space. However, the magnitude of the remaining problem remains to be determined. An achievable lower limit for SEEPS is estimated here by adapting the method of Göber et al. (2008). Gridded (6–6UTC) accumulations from the European high-density observation network (Ghelli and Lalaurette, 2000) are used as truth (to represent the output from a perfect forecasting system) and scored against the corresponding SYNOP observations. Scores are produced for a range of 'model' resolutions.

For the high-density data to represent the truth at a given resolution, there need to be sufficient observations in each grid-box. For the grid resolutions of ∼80, 40 and 25km assessed here, minima of 40, 18 and 6, respectively, are specified (see Table 14). Groisman and Legates (1994) point-out that area-mean precipitation can be biased in mountainous regions as most (U.S.) stations are located at low elevations. This possibility is not addressed here, although averages of scores over the whole of Europe should reduce any impact.

The implied lower bound for SEEPS is remarkably small for all resolutions. It is the value at the highest resolution (T799), converted to a skill score of $1 - 0.042 = 0.958$, which is indicated by the vertical dotted line in Fig. 3. Hence observation error and lack of representativity of grid-box averages does not impose any strong limit on the upper bound of SEEPS.

Resolution		Min High Density	Mean SYNOP used	SEEPS	
Spec.	Grid			Mean	70% Conf.
T255	80km	40	228	0.068	0.004
T511	40km	18	220	0.058	0.003
T799	25km	6	189	0.042	0.002

Table 14: Mean SEEPS scores and their 70% confidence intervals for a 'perfect model'. Results are based on the daily verification of gridded high-density observations against SYNOP observations. The gridded data are considered to represent a perfect model forecast. Results are shown for a range of 'model' resolutions.

12 Discussion and Conclusions

The aim of this study has been to develop a tailor-made precipitation score for monitoring progress in NWP and accurately comparing one model (cycle) with another. The outcome is an error score called here 'Stable Equitable Error in Probability Space' (SEEPS). It is a three-category error score that incorporates four key principles:

1. Error measured in 'probability space' (Ward and Folland, 1991). The climatological cumulative distribution function (Fig. 2) is used to transform errors into probability space. This allows the difficult distribution of precipitation to be accommodated in a natural way and reduces sampling uncertainty associated with extreme (possibly erroneous) data.

2. Equitability (Gandin and Murphy, 1992). By applying the equitability constraints (10), a forecast system

with skill will have a better expected score than a random or constant forecast system. In addition, scores from different climate regions can be readily combined.

3. Refinement (Murphy and Winkler, 1987). A constraint is devised to encourage a forecast system to predict all possible outcomes; thereby promoting a better distribution of forecast categories.

4. Reduction of sensitivity to sampling uncertainty by applying a 'strong perfect forecast' constraint (9). This constraint differentiates SEEPS from the skill score of (Gerrity, 1992, Fig. 4) - rendering scores more stable for forecasts (such as current ECMWF D+1 and D+2 forecasts) that have SEEPS error $< \frac{1}{2}$.

The categorical approach permits a strong link between the score and model error. The first category represents 'dry weather'. Here, 'dry' is defined with reference to WMO guidelines in order to be as compatible as possible with the varying reporting practices over the World and with model output. The other two categories, representing 'light' and 'heavy' precipitation, are defined in terms of climatological probabilities and are therefore dependent on the location and time of the year. Here, it is suggested that 'light' precipitation should be defined to occur twice as often as 'heavy' precipitation (Fig. 6).

The SEEPS error matrix naturally adapts to the climate of the location in question so that it can assess the salient aspects of the local weather. Asymmetries in the matrix penalise most heavily forecasts for a climatologically likely category that turn-out to be incorrect, and thus act to promote 'discrimination' in forecast systems. The hope is that this should accelerate forecast system improvement by encouraging developments that permit the model to represent all categories of local weather.

Except for very poor forecast systems, some physical understanding of forecast error is required to improve the SEEPS. Randomly changing a forecast category can only deteriorate the score. In this sense, SEEPS cannot be 'hedged'.

Verification is against point data (here 'SYNOP' data is used) so that it is possible to continuously monitor a system whose resolution is changing with time. With this point verification, the last remaining requirement in the list of desirable attributes (section 1) is satisfied.

Case-studies demonstrate that SEEPS is sensitive to key forecasting errors including the over-prediction of drizzle (Fig. 7), failure to predict heavy large-scale precipitation (Fig. 8) and incorrectly locating convective cells (Fig. 7).

The density of the observation network is taken into account when calculating area-mean scores. This implies, for example, that each sub-region of Europe will contribute approximately equally to the European-mean score. Area-mean results show an improving trend over the last 14 years (Figs. 10, Figs. 11). For the Extratropics, this amounts to $\sim$2 days gain in forecast skill at lead-times of 3–9 days. If this long-term trend is maintained, SEEPS will have a good chance of detecting improvements in new forecast cycles when compared over the same observational periods (Figs. 12).

By using gridded high-density observations for Europe to represent a 'perfect forecast', it has been shown that SYNOP observation error and lack of representativity of a grid-box average have minimal impact on the score. This is probably because 24-hour accumulations are being verified, the nearest grid-point is used when matching model output to point observations (rather than bilinear interpolation), and because forecast error is measured in probability space.

Experiments are underway to investigate if 6-hour accumulations can be verified for higher-resolution, limited-area model output. If feasible, this would partially resolve the important diurnal cycle in precipitation. It is possible that limited area model scores could be used to set realistic targets for global NWP. Separate experiments will apply SEEPS to ECMWF's probabilistic (ensemble) prediction system.

SEEPS scores for forecasts made within 'ERA-Interim' (which, unlike the operational system, uses a fixed model cycle and resolution) show almost no trend over the last 14 years (Fig. 10c). This indicates, strikingly, that the ∼100-fold increase in observations assimilated over this period has had no lasting impact on the operational forecast scores for precipitation. However, new observations that directly target the hydrological cycle may have more success. Future forecast system improvements could also arise from the better assimilation of existing observations (*e.g.* 'cloud-affected' radiances), with a more prognostic treatment of precipitation and with increasing model resolution.

Detailed and multi-facetted precipitation verification, beyond the abilities of SEEPS, will continue to be required but it is hoped that SEEPS can play a useful role in monitoring overall progress and in guiding developments in the right direction. Further, it is possible that SEEPS could be more widely applicable, and especially useful whenever the verification parameter has a difficult spatio or temporal distribution.

Acknowledgements

The authors would like to thank Ian Jolliffe and two un-named reviewers who provided numerous very useful comments, Anna Ghelli for the gridded high-resolution station data, and Thomas Jung, Cristina Primo and Martin Göber for additional valuable discussions.

References

Barnston, A. G., 1992: Correspondence among the correlation, RMSE, and the Heidkeforecast verification measures; refinement of the Heidke Score. *Weather and Forecasting*, **7**, 699–709.

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl.*, **11**, 141–154.

Casati, B., L. J. Wilson, D. B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E. Ebert, B. G. Brown, and S. Mason, 2008: Forecast verification: current status and future directions. *Meteor. Appl.*, **15**, 3–18.

Cherubini, T., A. Ghelli, and F. Lalaurette, 2002: Verification of precipitation forecasts over the alpine region using a high-density observing network. *Wea. Forecasting*, **17**, 238–249.

Du, J., S. L. Mullen, and F. Sanders, 2000: Removal of distortion error from an ensemble forecast. *Mon. Wea. Rev.*, **128**, 3347–3351.

Gandin, L. S. and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361–370.

Gerrity, J. P., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.*, **120**, 2709–2712.

Ghelli, A. and F. Lalaurette, 2000: Verifying precipitation forecasts using upscaled observations. (87), 9–17. Available at http://www.ecmwf.int/publications/.

Göber, M., E. Zsóster, and D. S. Richardson, 2008: Could a perfect model ever satisfy a naïve forecaster? on grid box mean versus point verification.. *Meteor. Appl.*, **15**, 359–365.

Groisman, P. Y. and D. R. Legates, 1994: The accuracy of United States precipitation data. *Bull. Amer. Meteor. Soc.*, **75**, 215–227.

Heidke, P., 1926: Berechnung des Erfolges und der Güte der Wind-stärkevorhersagen in Sturmwarnungsdienst.

Hoffman, R. N., Z. Liu, J. F. Louis, and C. Grassoti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.*, **123**, 2758–2770.

Murphy, A. H., 1978: Hedging and the mode of expression of weather forecasts. *Bull. Amer. Meteor. Soc.*, **59**, 371–373.

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

Murphy, A. H. and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454.

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.

Rodwell, M. J., 2005: Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurence better. ECMWF Newsletter 106, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

Simmons, A. J., S. Uppala, D. Dee, and S. Kobayashi, 2007: ERA-Interim: New ECMWF reanalysis products from 1989 onwards. ECMWF Newsletter 110, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

Stephenson, D. B., B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008: The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteor. Appl.*, **15**, 41–50.

von Storch, H. and F. W. Zwiers, 2001: *Statistical Analysis in Climate Research*. Cambridge University Press. 484 pp.

Ward, M. N. and C. K. Folland, 1991: Prediction of seasonal rainfall in the north Nordest of Brazil using eignenvectors of sea-surface temperature. *Int. J. Climatol.*, **11**, 711–743.