RMetS
Royal Meteorological Society

A new equitable score suitable for verifying precipitation in numerical weather prediction

Mark J. Rodwell,* David S. Richardson, Tim D. Hewson and Thomas Haiden
European Centre for Medium-Range Weather Forecasts, Reading, UK
*Correspondence to: Mark J. Rodwell, ECMWF, Reading, Berkshire, RG1 9AX, UK. E-mail: mark.rodwell@ecmwf.int

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 the error in 'probability space' through use of the climatological cumulative distribution function. For sufficiently skilful forecasting systems, the new score is less sensitive to sampling uncertainty than other established scores. It is therefore called here the '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 discouraging 'hedging'. To permit continuous monitoring of a system with resolution increasing in 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 relatively little impact. SEEPS can identify key forecasting errors including the overprediction of drizzle, failure to predict heavy large-scale precipitation and incorrectly locating convective cells. Area averages are calculated taking into account the observation density. A gain of ∼2 days, at lead times of 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.

Key Words: equitability; probability space; sampling uncertainty; refinement; hedging

Received 14 August 2009; Revised 23 April 2010; Accepted 10 May 2010; Published online in Wiley InterScience 23 July 2010

Citation: Rodwell MJ, Richardson DS, Hewson TD, Haiden T. 2010. A new equitable score suitable for verifying precipitation in numerical weather prediction. *Q. J. R. Meteorol. Soc.* **136**: 1344–1363. DOI:10.1002/qj.656

1. Introduction

Routine verifying is crucial in numerical weather prediction (NWP) for monitoring progress, setting targets, comparing forecasts by different centres and 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 eight-day ($D + 8$) European Centre for Medium-Range Weather Forecasts (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 Figure 1 show (a) observed (i.e. analyzed) and (b) $D + 4$ forecast $Z500$ verifying at 1200 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 characterize the entire weather pattern. Precipitation (shaded), for example, was rather poorly predicted over Europe on this date. This emphasizes 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,

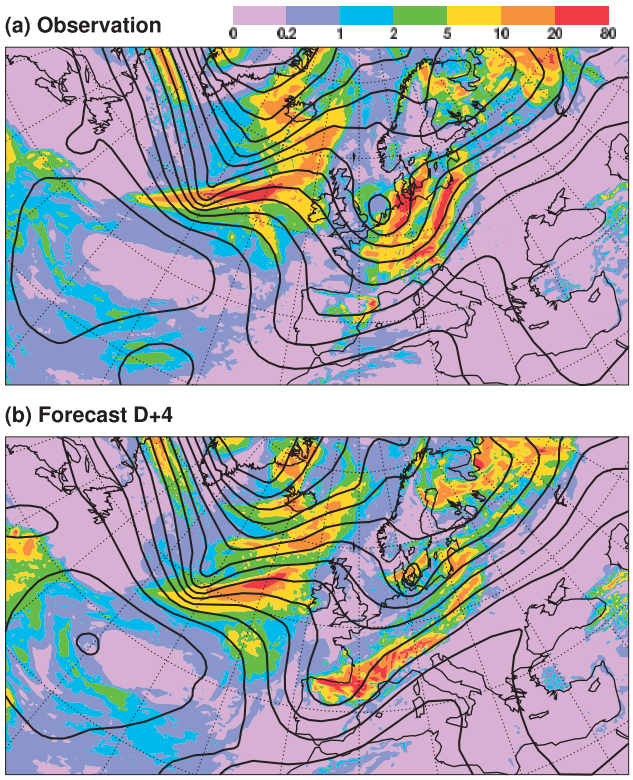

Figure 1. 500 hPa geopotential height field ($Z500$, contoured with interval 50 m) and 24 hour accumulated precipitation (shaded, mm). (a) 'Observations': analyzed $Z500$ and short-range ($D+0-D+1$) forecast precipitation centred at time 1200 UTC on 23 August 2008. (b) Forecast: $D+4$ forecast $Z500$ and $D+3\frac{1}{2}-D+4\frac{1}{2}$ forecast precipitation verified at the same time.

to make development decisions based on many scores. Ideally decisions should be based on some minimal set of scores that concisely summarizes 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, often with a large number of dry days and occasional very extreme events (notice the nonlinear colour scale in Figure 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. Using a scale-selective score, Roberts and Lean (2008) showed persuasively the scales at which high-resolution forecasts possess useful information about convection. Such decompositions are essential to truly understand the nature of forecast error but are not ideally suited for the present objectives of routine monitoring and high-level decision-making. For example, re-calibration within the methodology of Casati *et al.* (2004) renders their score insensitive to multiplicative changes in precipitation, an issue that could be important in the context of the present study. Other research has centred on the verification of extreme precipitation (Stephenson *et al.*, 2008). This is highly desirable from the user's 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 summarized as follows.

(a) Monitoring Progress.

- A single score should be sought that assesses forecast skill for dry weather and precipitation quantity.
- Verification against point observations is required in order to permit continuous monitoring of a system with resolution 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 minimized, while maintaining the ability to differentiate between 'good' and 'bad' forecasts.
- For area and temporal averages to be meaningful, it should be possible to aggregate scores from different climate regions and different times of the year.

(b) Aiding decision-making.

- To facilitate the identification of model error, it should be possible to plot a map of scores for a single 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 the 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 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 unskilful 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, which take account of observation density, are

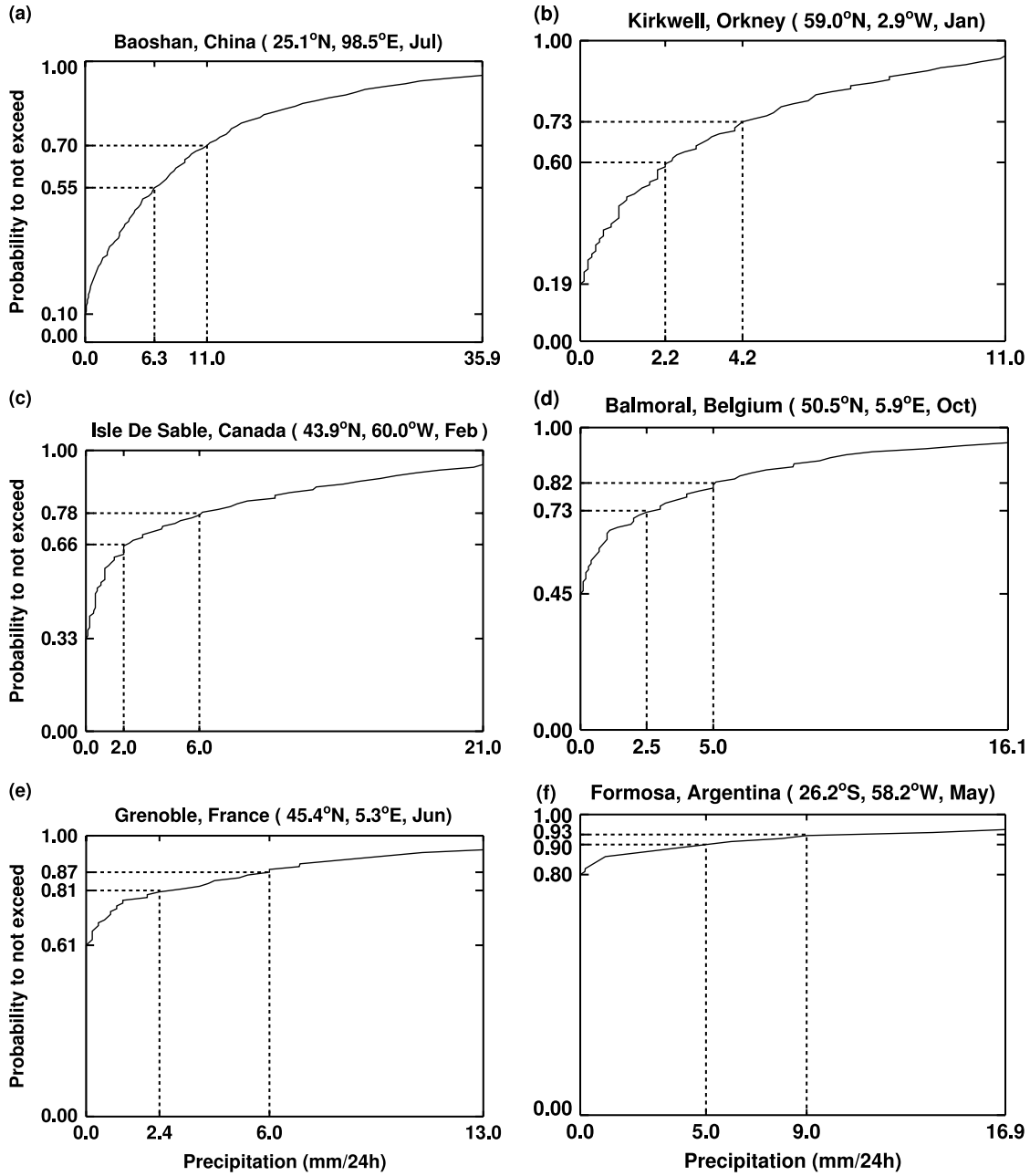

Figure 2. Cumulative distributions for selected SYNOP stations and months based on 1200–1200 UTC 24 hour precipitation accumulations for 1980–2008. The extreme right of each graph corresponds to the 95th percentile of the distribution. Dotted lines indicate the subdivision of the wet days in the ratios 1:1 and 2:1.

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 Figure 1(a). The utility 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 six-hour 0000–0600 UTC accumulation is derived by subtracting the previous six-hour 1800–0000 UTC accumulation from the 12 hour 1800–0600 UTC accumulation. European 24 hour 0000–0000 UTC accumulations are then deduced by combining this derived six-hour 0000–0600 UTC accumulation with the subsequent reported 12 hour 0600–1800 UTC and the six-hour 1800–0000 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 worldwide is increased from ∼500 to ∼4000 for periods ending at 0000, 0600, 1200 and 1800 UTC. The results presented here mainly focus on accumulations ending at 1200 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 <1 m.

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 ∼5 years of observations ($5 \times 30 = 150$).

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 1200–1200 UTC accumulations. These cumulative distribution functions have a different structure from 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 (Figure 2(a)) is frequently wet in July, with only 10% of days being 'dry'. Formosa, Argentina (Figure 2(f)) is dry 80% of the time in 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 0600–0600 UTC 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 1200 UTC high-resolution ('deterministic') forecast is used to obtain 24 hour 1200–1200 UTC accumulated precipitation forecasts for lead times of 1–10 days, for the period 1995–2008. These data are matched to all 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 (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–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, even where reporting practices vary, and should allow a consistent comparison with forecast data. The solution has been to base the definition on the World Meteorological Organization (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 the following.

- '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 a '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 are 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 (as opposed to categorical) scores of precipitation have previously been considered. For example, the spatial correlation of normalized 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

3.2.1. Equitability

The 'Hit-Rate', or 'Probability of Detection' is a two-category score defined as $H/(H+M)$ where H is the number of correctly forecast events (hits) and M is the number of observed events that were not predicted (misses). Note that $H+M$ is the total number of events that actually occurred. For a perfect forecasting system, Hit-Rate$=1$. However, the converse is not true. A trivial forecast that always predicted the event would have Hit-Rate$=1$ but is clearly not a perfect forecasting system. Development 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) or a bonus for correctly predicting that the event would not occur (a correct negative).

Before discussing methods of constructing scores that can ensure that they do not suffer in the way that the Hit-Rate does, it is useful to convert to a more useful notation. The sample can be thought of as consisting of a set of observation/forecast pairs (v,f), where f is the forecast category and v is the verifying observation category. Using tilde ($\tilde{}$) to denote sample-mean values, as opposed to expected (i.e. population-mean or climatological-mean) values or constants, one can write the joint sample distribution as $\{\tilde{p}_{vf}\}$ and the observed sample distribution as $\{\tilde{p}_v\}$. With this notation, the sample-mean Hit-Rate, for hits of category 1, can be written as

$$\tilde{S}^{\mathrm{HR}} = \frac{\tilde{p}_{11}}{\tilde{p}_1}\,. \tag{1}$$

Following previous attempts (Gringorten, 1967; Jolliffe and Foord, 1975), Gandin and Murphy (1992) formalized some 'equitability' constraints that can be used to construct more useful scores. Firstly, they highlighted the desirability of separating the forecasting and scoring tasks by writing a score (with n categories) in the form

$$\tilde{S} = \sum_{v,f} \tilde{p}_{vf} s_{vf}, \tag{2}$$

Table I. Scoring matrix for the Hit-Rate score (for hits of category 1). 'FC' refers to the forecast, 'Obs' refers to the verifying observations and the values $\{\tilde{p}_v\}$ refer to the observed sample frequencies of the categories.

			Obs	
	Freq		$\tilde{p}_1$	$\tilde{p}_2$
	Cat		v	
			1	2
FC	f	1	$\frac{1}{\tilde{p}_1}$	0
		2	0	0

where $\{s_{vf}\}$ is an $n \times n$ scoring matrix independent of $\{\tilde{p}_{vf}\}$, but possibly dependent on the observed sample distribution $\{\tilde{p}_v\}$. Although considered unnecessarily restrictive by Hogan *et al.*. (2010), such a separation would appear to be particularly useful when comparing two forecast systems and investigating why one system scores better than the other on any particular day. For the Hit-Rate, an obvious way to represent (s_{vf}^{HR}) is given in Table I.

Secondly, Gandin and Murphy (1992) prescribed constraints to ensure that all systems that predict a constant category (such as the climatologically most likely category for example) are awarded the same score, S_{c} say, and a perfect forecasting system is awarded a score S_{p} ($\neq S_{\mathrm{c}}$):

$$\left.\begin{array}{ll} \text{Perfect FC:} & \sum_v \tilde{p}_v s_{vv} = S_{\mathrm{p}}, \\ \text{Constant FC:} & \sum_v \tilde{p}_v s_{vf} = S_{\mathrm{c}} \quad \forall f, \end{array}\right\} \tag{3}$$

where '$\forall$' means 'for all'. In (3), the perfect forecast constraint is a sum along the diagonal of $\{s_{vf}\}$, and the constant forecast constraints are sums along the rows, all weighted by the observed sample distribution. A score that is separated from the forecasting task in the manner described by (2) and which satisfies the constraints (3) is known here as 'equitable'.

As Gandin and Murphy (1992) point out, the constant forecast constraints imply that the expected score for any random forecasting system, with climatological distribution $\{q_f\}$ (where $\sum_f q_f = 1$), is also S_{c} since

$$\sum_{v,f} q_f \tilde{p}_v s_{vf} = \sum_f q_f \left(\sum_v \tilde{p}_v s_{vf} \right) = S_{\mathrm{c}}. \tag{4}$$

It will be assumed from now on that, for a skill score, $S_{\mathrm{p}} = 1$ and $S_{\mathrm{c}} = 0$.

The implications of equitability, as discussed by Hogan *et al.* (2010), can be summarized as follows. If a score is *in*equitable, and it accords different scores to two unskilful (e.g. random) forecast systems, then adding some skill to the system with the poorer score could still leave it apparently worse than the other unskilful system. Equitability removes this undesirable possibility. It is also noted here that, by heavily penalizing 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

observed 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 I) does not satisfy the constant constraint for $f = 1$ in (3) and, thus, is not equitable. For a two-category equitable score, the three equations in (3) (one perfect and two constant) are not enough to constrain the four elements of the 2×2 scoring matrix $\{s_{vf}\}$. Nevertheless, it is interesting to note that the score value calculated using (2) is uniquely determined. It can be written as

$$\tilde{S}^{\mathrm{P}} = \frac{\tilde{p}_{11}}{\tilde{p}_1} - \frac{\tilde{p}_{21}}{\tilde{p}_2}. \tag{5}$$

To show this, write each s_{vf} in terms of s_{11} and use the relations $\tilde{p}_{11} + \tilde{p}_{12} = \tilde{p}_1$, $\tilde{p}_{21} + \tilde{p}_{22} = \tilde{p}_2$ and $\tilde{p}_1 + \tilde{p}_2 = 1$. This score is the Hit-Rate minus the False-Alarm-Rate, first defined by Peirce (1884). It is called here the Peirce skill score, although it has been named differently over the years (for example the 'Hanssen–Kuipers discriminant', 'Kuipers' performance index' and 'true skill statistic'). It is thus the only equitable two-category score (to within linear transformations) according to the definition of Gandin and Murphy (1992). Unlike the Hit-Rate alone, the Peirce skill score does include a penalty for false alarms and is less easily increased by overpredicting the event.

For 'small' but realistic sample sizes, there is a chance that $\tilde{p}_1 = 0$ or $\tilde{p}_2 = 0$ in (5). For example if the climatological probability $p_1 = 0.05$, the chance that $\tilde{p}_1 = 0$ for a sample size of 30 is $(1 - 0.05)^{30} = 0.21$. Hence there is a strong possibility that a score such as the Peirce skill score in (5) will not be defined for realistic sample sizes. In addition, it is unclear how a coherent set of daily scores could be produced and augmented if the scoring matrix were based on the observed sample distribution (as, e.g., in Table I). One approach to solving this issue, envisaged by Gringorten (1967), is to base the scoring matrix on a climatological observed distribution, $\{p_v\}$ rather than the sample observed distribution, $\{\tilde{p}_v\}$. The 29 year climatology developed here makes this feasible for global scores of precipitation. One valid scoring matrix for the Peirce skill score (based on the climatological distribution, and after imposing symmetry as the fourth constraint) is given in Table II.

Table II. Symmetric scoring matrix for the Peirce skill score based on the climatological distribution.

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

Unlike the case for the sample-based equitable two-category scores, finite sample means of the climatology-based Peirce skill score do depend on the choice of $s_{vf}(\{p_v\})$ and the equitability constraints (3) only strictly apply in the limit as the sample size tends to infinity (i.e. in terms of expectation rather than realized scores):

$$\begin{aligned} &\text{Perfect FC:} && \textstyle\sum_v p_v s_{vv} = 1, \\ &\text{Constant FC:} && \textstyle\sum_v p_v s_{vf} = 0 \quad \forall f. \end{aligned} \tag{6}$$

Note that if the two observed categories have equal climatological probabilities ($p_1 = p_2$), then the diagonal elements of the scoring matrix in Table II satisfy what will be called here the 'strong perfect forecast constraints':

$$\text{Strong Perfect FC:} \quad s_{vv} = 1 \quad \forall v, \tag{7}$$

and the sample-mean score of 1 for a perfect forecast system is effectively re-imposed, not just in an expected sense but for any finite sample. Satisfying (7), even in situations of unequal $\{p_v\}$, is found here to be a desirable attribute and will be discussed further.

3.2.2. *Linear error in probability space*

For a categorical score that assesses both the prediction of dry weather and precipitation quantity, more than two categories are required. Below, the attributes of some established equitable n-category scores are discussed. In all cases, scores will be considered to be defined by the climatology, with equitability defined in terms of expectation (6) rather than sample-mean scores (3).

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 in any category equally and penalizes all misses equally, regardless of the class of category error. The Heidke skill-scoring matrix for the three-category score is shown in the form $(3I_3 - 1)/2$ in Table III.

The Heidke skill-scoring matrix satisfies the strong perfect forecast constraints (7) and, for equiprobable categories ($p_v = \frac{1}{3}\ \forall v$ in the case of three categories), it also satisfies the equitability constraints (6).

Barnston (1992) modified the Heidke skill score for equiprobable 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 three-category scoring matrix is given in Table IV. Its dependence on the class of

Table III. Scoring matrix for a three-category Heidke skill score.

			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 IV. Scoring matrix for a three-category Barnston skill score with equiprobable climatological categories.

			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	$\frac{9}{8}$

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 constraints (7).

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

$$s^L_{vf} = |f - v|, \tag{8}$$

where v and f are the observed and forecast categories, respectively (defined by terciles, quintiles, etc) and L refers to 'LEPS'. After subsequent adjustments including those for equitability, the scoring matrix for the three-category LEPS skill score (Potts *et al.*, 1996) with equiprobable climatological categories is given in Table V. Notice that the final scoring matrix is not entirely linear.

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

$$\text{Constant Obs:} \quad \textstyle\sum_f p_f s_{vf} = 0 \quad \forall v \tag{9}$$

is also satisfied. This means that the expected skill score for a constant observation is also 0. However, this is only realized 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 constraints (7). Note that, for daily precipitation, it is not possible to make categories equiprobable if 'dry' weather is to be defined as a category in itself.

Gerrity (1992) demonstrated how, for unequal probabilities, an equitable n-category score with a symmetric scoring matrix could be constructed as the mean of $n - 1$ Peirce skill scores with 2×2 symmetric scoring matrices of the form given in Table II. The three-category Gerrity skill-scoring matrix, $\{s^G_{vf}\}$, for any $\{p_v\}$ is given by

$$\{s^G_{vf}\} = \frac{1}{2}\begin{Bmatrix} \frac{1-p_1}{p_1}+\frac{p_3}{1-p_3} & \frac{p_3}{1-p_3}-1 & -2 \\ \frac{p_3}{1-p_3}-1 & \frac{p_1}{1-p_1}+\frac{p_3}{1-p_3} & \frac{p_1}{1-p_1}-1 \\ -2 & \frac{p_1}{1-p_1}-1 & \frac{p_1}{1-p_1}+\frac{1-p_3}{p_3} \end{Bmatrix}. \tag{10}$$

This score would allow 'dry' days to be defined as a single category. The scoring matrix in (10) for variable $\{p_v\}$ will be discussed later. Substituting $p_1 = p_2 = p_3 = \frac{1}{3}$ into (10) gives the three-category scoring matrix for equiprobable climatological categories (Table VI). As with the Barnston and LEPS skill scores, the Gerrity scoring matrix does not satisfy the strong perfect forecast constraints (7).

4. Stable equitable error in probability space

The aim here is to construct a score that possesses the desirable attributes listed in section 1. Measuring errors in probability space and ensuring equitability (in terms

Table V. Scoring matrix for a three-category LEPS skill score with equiprobable climatological categories.

			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 VI. Scoring matrix for a three-category Gerrity skill score with equiprobable climatological categories.

			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	$-\frac{1}{4}$	$\frac{5}{4}$

of expectation) should help in this regard. To allow 'dry' weather to be a category in itself, the score must accommodate categories with variable probabilities. To aid in the detection of performance trends, a score's sensitivity to sampling uncertainty needs to be minimized, while maintaining its ability to differentiate between good and bad forecasts. With the aim of minimizing 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 the error in probability space, it is formulated as an 'error score' rather than a 'skill score'.

The lack of perfect linearity (of the error in probability space) in the LEPS scoring matrix indicates that such linearity is not possible for an equitable score. Hence a less constrained structure than (8) is initially proposed here for an n-category error score. The first category represents 'dry' weather and has climatological probability p_1. The remaining categories represent bins with successively heavier precipitation and have equal climatological probabilities $p_i = (1 - p_1)/(n - 1)\ \forall i > 1$. The proposed structure is given by

$$s_{vf} = \left\{ \begin{array}{lll} |f - v|\, a & +\delta_{1f}(c - a) & \text{if } v > f, \\ |f - v|\, b & +\delta_{v1}(d - b) & \text{if } v < f, \\ 0 & & \text{if } v = f, \end{array} \right\} \qquad (11)$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. If $v > f$ ($v < f$), the error increases linearly by a value $a > 0$ ($b > 0$) for each extra category separating the forecast, f, and verifying observation, v. 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. Note that (11) is consistent with the strong perfect forecast constraints for an error score:

$$\text{Strong Perfect FC (error):} \quad s_{vv} = 0 \quad \forall v. \qquad (12)$$

The equitability constraints for an error score can be written as

$$\begin{array}{ll} \text{Perfect FC (error):} & \sum_v p_v s_{vv} = 0, \\ \text{Constant FC (error):} & \sum_v p_v s_{vf} = 1 \quad \forall f. \end{array} \qquad (13)$$

The perfect forecast constraint in (13) is automatically satisfied because (12) is. However, it is not possible to satisfy the constant forecast constraints in (13) if $n > 3$. This is because the combination of constant forecast constraints $\sum_v p_v s_{v2} - 2\sum_v p_v s_{v3} + \sum_v p_v s_{v4}$ implies that $(a + b)p_3 = 0$, and this is not possible since $a, b, p_3 > 0$.

A three-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 (11), is given in Table VII. 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).

Table VII. Error-matrix structure for a new three-category score. Here p_1, p_2, p_3 represent the climatological probabilities of 'dry weather', 'light precipitation' and 'heavy precipitation', respectively.

			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

The three constant forecast constraints in (13) are used to write b, c and d in terms of a:

$$\begin{array}{rcl} b & = & \dfrac{p_3 a}{1 - p_3}, \\ c & = & \dfrac{1 - p_3 a}{1 - p_1}, \\ d & = & \dfrac{1 - p_3 a}{p_1}. \end{array} \qquad (14)$$

Notice that, in general, $b \neq a$ and $c \neq d$. 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 VII, $(s_{31} - s_{21}) = (s_{32} - s_{22}) = a$, both of which relate to the same difference in probability space between observed categories 3 and 2. Similarly, $(s_{12} - s_{22}) = (s_{13} - s_{23}) = d$, 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 three 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 two-category score rather than a three-category score.

Since b, c and d must all be greater than 0, (14) 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 were allowed to take its extreme values. For $a = 0$ (upper error matrix in Table VIII), 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 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 VIII), 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

Table VIII. Error matrices for the two sets of extreme values of a, b, c and d. See the main text for more details.

			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

by defining a 'refinement constraint' that maximizes the lower bound on the expected error for any forecast system that never predicts category 1 or category 3. Before deducing this value of a, it is worth noting that it is less important to penalize 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 penalized by equitability (13).

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 for the fraction p_3 (p_1) of times that category 3 (1) occurs. This leads to an expected score of $p_3 s_{32} = p_3 a$ ($p_1 s_{12} = p_1 d = 1 - p_3 a$, from (14)). The lower bound for the expected error for a two-category system is therefore $\min(p_3 a, 1 - p_3 a)$, which is maximized at $\frac{1}{2}$ when $a = 1/(2p_3)$. Choosing this value of a should reward 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, $\{s^{\mathrm{S}}_{vf}\}$, is given by

$$\{s^{\mathrm{S}}_{vf}\} = \frac{1}{2}\begin{Bmatrix} 0 & \frac{1}{1-p_1} & \frac{1}{p_3}+\frac{1}{1-p_1} \\ \frac{1}{p_1} & 0 & \frac{1}{p_3} \\ \frac{1}{p_1}+\frac{1}{1-p_3} & \frac{1}{1-p_3} & 0 \end{Bmatrix}. \quad (15)$$

In anticipation of its reduced sensitivity to sampling error, this score will be called here the 'stable equitable error in probability space' (SEEPS).

Table IX. Scoring matrix for a three-category SEEPS skill score with equiprobable climatological categories.

			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

5. Comparison with other scores

Here a comparison is made with the skill scores reviewed in section 3.2.2 (all assumed to be defined in terms of the climatological observed distribution $\{p_v\}$ rather than the sample observed distribution $\{\tilde{p}_v\}$). The comparison requires a SEEPS 'skill score'. This can readily be produced by calculating 1−SEEPS, which clearly satisfies the equitability constraints (6). For three equiprobable categories (as is the case for Isle De Sable, Figure 2(c), with $p_2/p_3 = 1$), the scoring matrix for the SEEPS skill score is given in Table IX.

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 (equiprobable category) scoring matrices in Tables III, IV, V, VI and IX. Interestingly, this maximum is the same ($\frac{1}{2}$) for all scores. As with SEEPS, the three-category Gerrity skill score has this maximum value for all $\{p_v\}$.

5.2. *Sensitivity to sampling 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 (from finite samples of forecasts). Since SEEPS satisfies the strong perfect forecast constraints (12), 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 constraints make 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 deterministic 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

$$p_{v|f} = (1-\gamma)p_v + \gamma\delta_{vf}, \quad (16)$$

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 (16) is consistent with this assumption, since

$$\begin{aligned}\sum_f p_{v|f}p_f &= \sum_f \left[(1-\gamma)p_v + \gamma\delta_{vf}\right]p_f \\ &= \left[(1-\gamma)p_v \sum_f p_f\right] + \gamma p_v \\ &= (1-\gamma)p_v + \gamma p_v \\ &= p_v.\end{aligned} \tag{17}$$

The range of γ in (16) 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 unskilful (Murphy and Winkler, 1987). For $\gamma = 1$, it can be seen that $p_{v|f} = 1$ 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 three equiprobable categories (Tables III, IV, V, VI and IX). With the conditional distribution defined in (16), the expected skill for all these scores (indeed any equitable skill score satisfying (6)) is simply γ:

$$\begin{aligned}S &= \sum_{v,f} p_{vf}s_{vf} \\ &= \sum_{v,f} p_{v|f}p_f s_{vf} \\ &= \sum_{v,f} \left[(1-\gamma)p_v + \gamma\delta_{vf}\right]p_f s_{vf} \\ &= (1-\gamma)\sum_f \left(p_f \sum_v p_v s_{vf}\right) + \gamma \sum_v p_v s_{vv} \\ &= (1-\gamma)\sum_f \left(p_f \times 0\right) + (\gamma \times 1) \\ &= \gamma,\end{aligned} \tag{18}$$

where the equitability constraints (6) have been invoked in the penultimate line. The expected standard deviation of an equitable score with scoring matrix $\{s_{vf}\}$ can thus be written as

$$\sigma(\gamma) = \sqrt{\sum_{v,f} \left(\{s_{vf}\} - \gamma\right)^2 p_{v|f}p_f}. \tag{19}$$

Figure 3 shows $\sigma(\gamma)$ for each score. (To obtain the standard deviation, and thus confidence intervals, of a sample mean with finite 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 the present mean-forecast skill is already better than $\frac{1}{2}$ at short lead times (see later), SEEPS would appear to be preferable. The higher standard deviation of SEEPS for $\gamma < \frac{1}{2}$ is less relevant and will become even more so in future.

5.2.1. *Relationship between SEEPS and Gerrity scores*

It can be seen that in each of the columns of the Gerrity and SEEPS skill-scoring matrices for equiprobable categories (Tables VI and IX) the values differ only by a constant (dependent only on v). By comparing (10) and (15) it can

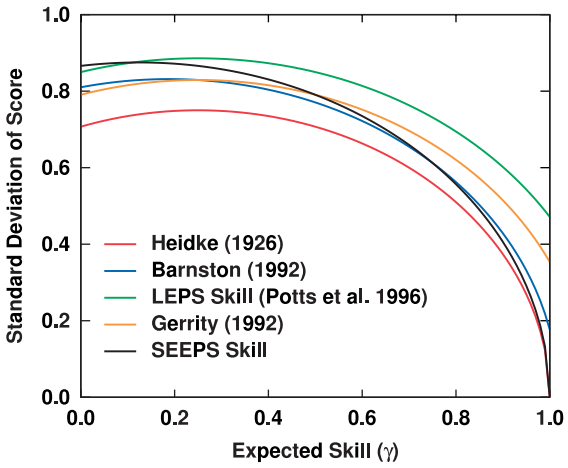

Figure 3. Expected standard deviation of a range of three-category forecast scores as a function of expected skill, γ. Equiprobable categories are used for each score indicated in the key.

Table X. The two-category equitable error matrix for a score that SEEPS can be built from.

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

readily be shown that this is true in general, for any given $\{p_v\}$, so that

$$\begin{aligned} s^{\mathrm{G}}_{vf} &= (1 - s^{\mathrm{S}}_{vf}) + \lambda(v) \qquad \forall v,f, \\ \tilde{S}^{\mathrm{G}} &= \tilde{S}^{\mathrm{S}} + \textstyle\sum_v \tilde{p}_v\lambda(v) \\ &\equiv \tilde{S}^{\mathrm{S}} + \Lambda(\{\tilde{p}_v\}), \end{aligned} \tag{20}$$

where $\tilde{S}^{\mathrm{G}}$ and $\tilde{S}^{\mathrm{S}}$ are sample-mean Gerrity and SEEPS *skill* scores calculated as in (2). Equation (20) 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 $\{\tilde{p}_v\}$.

It can be shown that $\Lambda(\{p_v\}) = 0$ and so, for an *infinite* sample size, the two skill scores have identical expected values: $S^{\mathrm{G}} = S^{\mathrm{S}}$. The similarity between the two scores extends further, since SEEPS can also be written as the mean of two two-category error scores, each with an error matrix of the form shown in Table X. The first of these has its two categories defined by the dry/light threshold, the other by the light/heavy threshold. As with the two-category scoring matrix used to generate the Gerrity skill score sequence, (1−Table X) is also a valid choice for representing a scoring matrix for the Peirce skill score (based on the climatology).

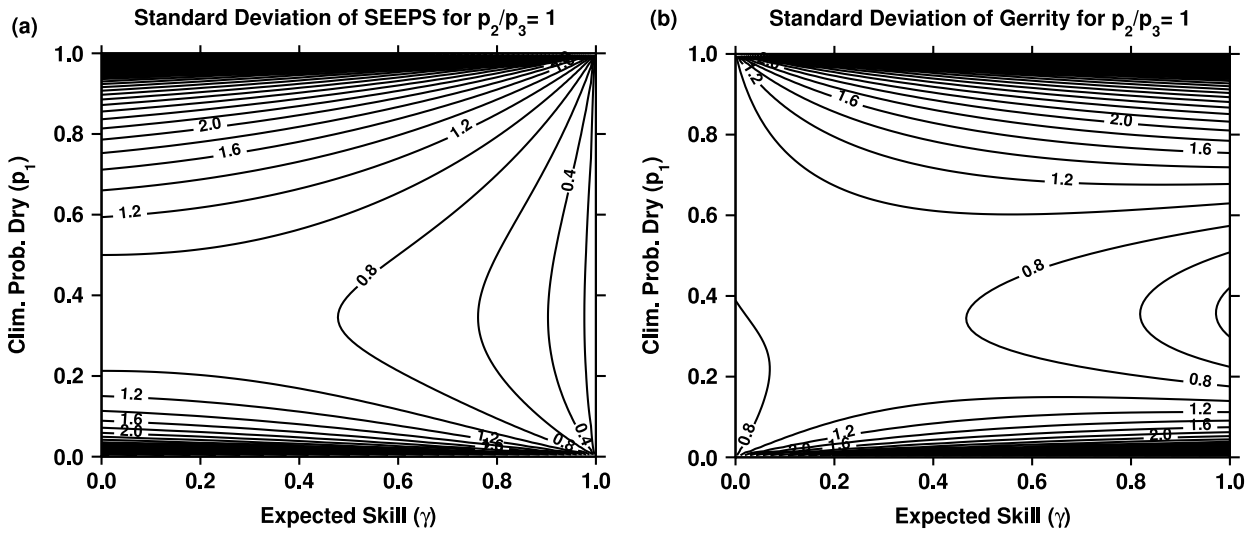

Figure 4. Standard deviation of scores as a function of (γ, p_1) for $p_2/p_3 = 1$ when the forecast system is defined by (16). (a) SEEPS. (b) Gerrity skill score.

The difference between the Gerrity and SEEPS skill scores lies in their sensitivity to sampling uncertainty for *finite* sample sizes. Figure 4(a) and (b) shows $\sigma(\gamma, p_1)$ for the SEEPS and Gerrity skill scores respectively (when the forecast system is defined by (16) and $p_2/p_3 = 1$). As the system's performance improves, the Gerrity skill score becomes even more sensitive to sampling uncertainty when p_1 diverges from $\frac{1}{3}$, while SEEPS' uncertainty smoothly converges to zero for all p_1. It can be shown that, for any $\{p_v\}$ (not just when $p_2/p_3 = 1$), the Gerrity skill score is more sensitive to sampling uncertainty than SEEPS when assessing systems that have expected skill $> \frac{1}{2}$. SEEPS is more sensitive when the expected skill is $< \frac{1}{2}$, but this is less relevant for the reason discussed above. Numerical investigation shows that these results are valid (approximately) for all realistic forecasting systems, not just those defined by (16).

It is interesting to discover that the derivation of SEEPS here, which addresses key requirements for the monitoring of precipitation forecasts, produces a score very similar to the rather elegantly constructed Gerrity skill score. The only difference is in the choice of climatological scoring matrix for the Peirce skill score on which they are based. Gerrity (1992) constrained the last degree of freedom of this two-category equitable score by imposing symmetry. The results presented here demonstrate that symmetry is not a useful attribute (in this situation) and the last degree of freedom is, instead, constrained by requiring all perfect forecasts to have zero error. This difference is considered important and should render SEEPS more stable for assessing forecast systems with sufficient expected skill.

It is possible that Table X 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 (11) for $n > 3$.

The comparison of uncertainty in section 5.2 (and 5.2.1) 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. For example, taking a difference will eliminate the $\Lambda(\{\tilde{p}_v\})$ term in (20).

5.3. *Hedging*

While the concept of 'hedging' has been investigated by previous authors (Stephenson, 2000; Hogan *et al.*, 2009), a somewhat different approach is attempted here. 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 whether 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 here 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 (15), 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

$$\begin{aligned}
\delta\text{SEEPS} &= \delta q \textstyle\sum_v p_{v|f=1}(s_{v2} - s_{v1}) \\
&= \frac{\delta q}{2}\left(\frac{p_{v=1|f=1}}{p_1} - \frac{p_{v=2|f=1} + p_{v=3|f=1}}{1-p_1}\right) \\
&= \frac{\delta q}{2}\left(\frac{p_{v=1|f=1}}{p_1} - \frac{1-p_{v=1|f=1}}{1-p_1}\right) \\
&= \frac{\delta q}{2p_1(1-p_1)}(p_{v=1|f=1} - p_1).
\end{aligned} \tag{21}$$

Hence SEEPS is reduced only if $p_{v=1|f=1} < p_1$, and thus only if the original forecasting system is worse in its prediction of dry weather than a climatological forecast. Using similar mathematics, the change 2→1 only decreases SEEPS if $p_{v=1|f=2} > p_1$, and thus again only if the original forecasting

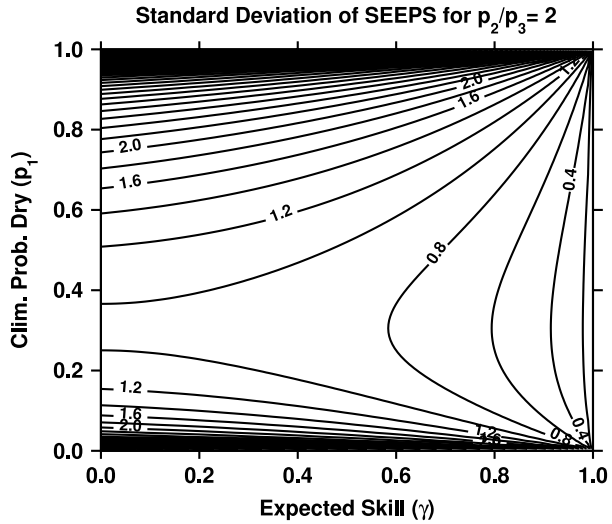

Figure 5. As Figure 4(a) but for $p_2/p_3 = 2$.

system is unrealistically poor. Similarly, 3→2 only reduces SEEPS if $p_{v=3|f=3} < p_3$ and 2→3 only reduces 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 $\{p_v\}$ and is not dependent on the refinement constraint (similar mathematics holds for any a with $0 < a < 1/p_3$).

Similarly, the three-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. 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 $p_{v|f=v} \geqslant p_v$ and $p_{v|f\neq v} \leqslant p_v$ $\forall v$. However, LEPS can be hedged even under these constraints.

Other approaches to hedging, particularly associated with dynamically unconstrained post-processing of model output, may allow selection from within a forecast category and are not precluded by the above analysis. While post-processing is not relevant to the present study, further investigation of the susceptibility of scores to hedging is warranted.

6. SEEPS parameter settings

Equation (15) 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 (Figure 2(e)), for example, the threshold increases from 2.4 to 6.0 mm. These higher thresholds set a more challenging task for a forecasting system.

Using the conditional distribution (16), Figure 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 ($p_1 > 0.85$) is effectively considered as 'extreme weather' and will be neglected here 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.* Figure 4(a) and Figure 5). Unless otherwise specified, these are the settings used from now on. Section 10.1 tabulates real forecast results that tend to confirm these choices.

Table XI. 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'.

			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

7. SEEPS: summary of the score

Table XI 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 ($p_2/p_3 = 2$). Although Table XI shows individual SEEPS scores as large as 13.33, equitability ensures that time-mean scores (averaged over a sufficient number of forecasts) should lie within $[0, 1]$.

Notice that a prediction for wet conditions when it turns out to be dry is more heavily penalized in climatologically wet regions than in climatologically dry regions (c.f. top and bottom panels in Table XI). In general, a forecast for a climatologically likely category that turns out to be incorrect is penalized more heavily than a forecast for an unlikely category that turns out to be incorrect. This is a desirable

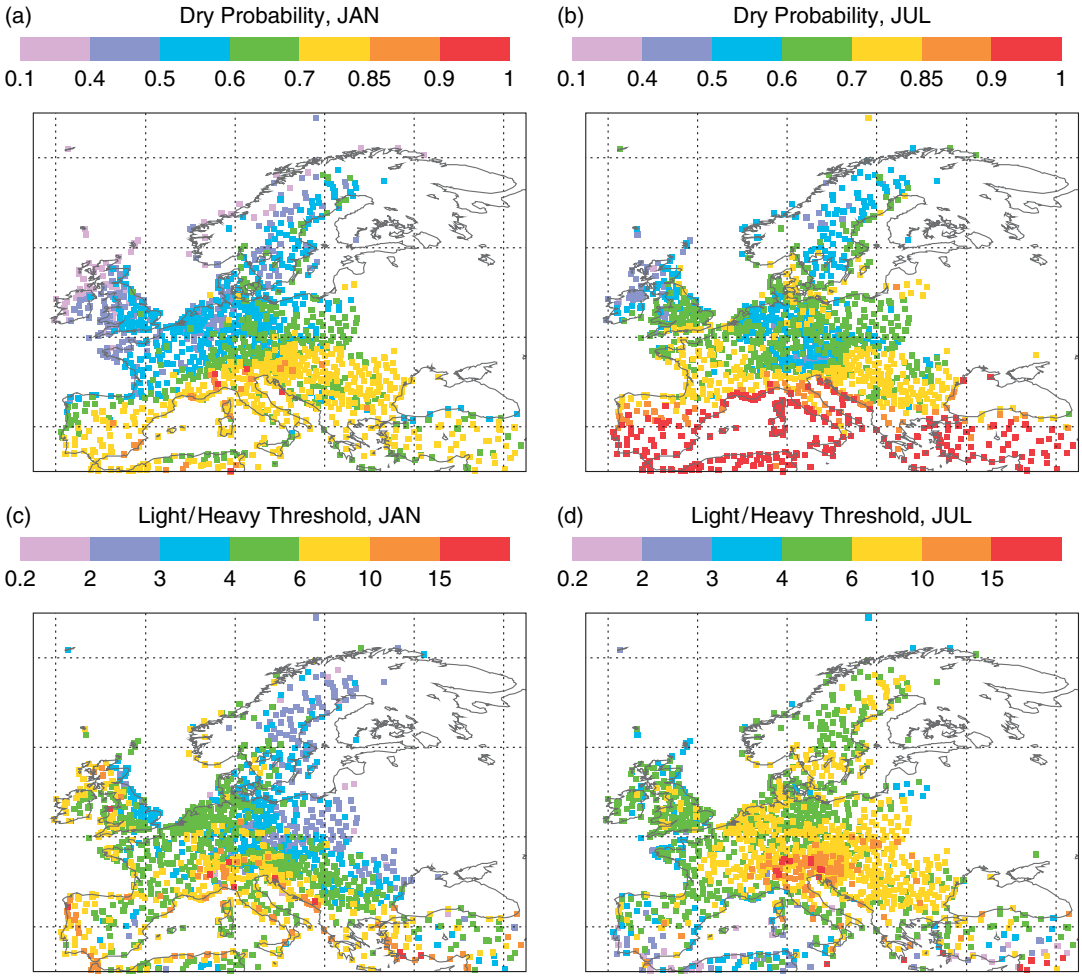

Figure 6. (a) Probability of a 'dry' day for January. (b) As (a) but for July. (c) Precipitation amount (in mm) marking the threshold between 'light' and 'heavy' precipitation for January. (d) As (c) but for July. By definition, 'light precipitation' occurs twice as often as 'heavy precipitation'. Results are based on 24 hour precipitation accumulations (1200 UTC–1200 UTC) from the 1980–2008 climatology.

attribute, since it should encourage developments that allow the model (physics) to represent all categories, whatever their climatological frequency.

For European stations, p_1 is shown in Figure 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 Figure 6(c) and (d) for January and July, respectively. For Europe, this threshold is generally between 3 and 10 mm, 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. It is *stable* in the face of sampling uncertainty (for fairly skilful forecasts) because it satisfies a strong perfect forecast constraint. It is *equitable* and, because it measures *error* in *probability space*, it is robust with respect to the skewed distribution of precipitation. SEEPS rewards systems that predict all categories and it also inhibits hedging. SEEPS should, therefore, be useful for monitoring performance and for guiding development decisions.

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.

Figure 7(a) shows observed 24 hour accumulated precipitation (in mm) on 16 December 2008, and Figure 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 ahead of a frontal system but were actually 'dry' (pink). In this case, recorded values were 0.0 mm rather than 0.1 or 0.2 mm. Since this region is generally wet in December (Figure 7(c)) and an incorrect forecast for a likely category is strongly penalized, the differences in precipitation categories (c.f. Figure 7(d) and (e)) lead to relatively large SEEPS scores (Figure 7(f)). Large SEEPS scores along the southern coast of France (Figure 7(f)) reflect unpredicted heavy precipitation (c.f. Figure 7(d) and (e)) associated with a Mediterranean low-pressure system in this relatively dry climate region (Figure 7(c)). These issues explain why the mean European score for this forecast was one of the worst in 2008.

Note that the station scores in Figure 7(f) are plotted with variable sizes to indicate their relative weight within an

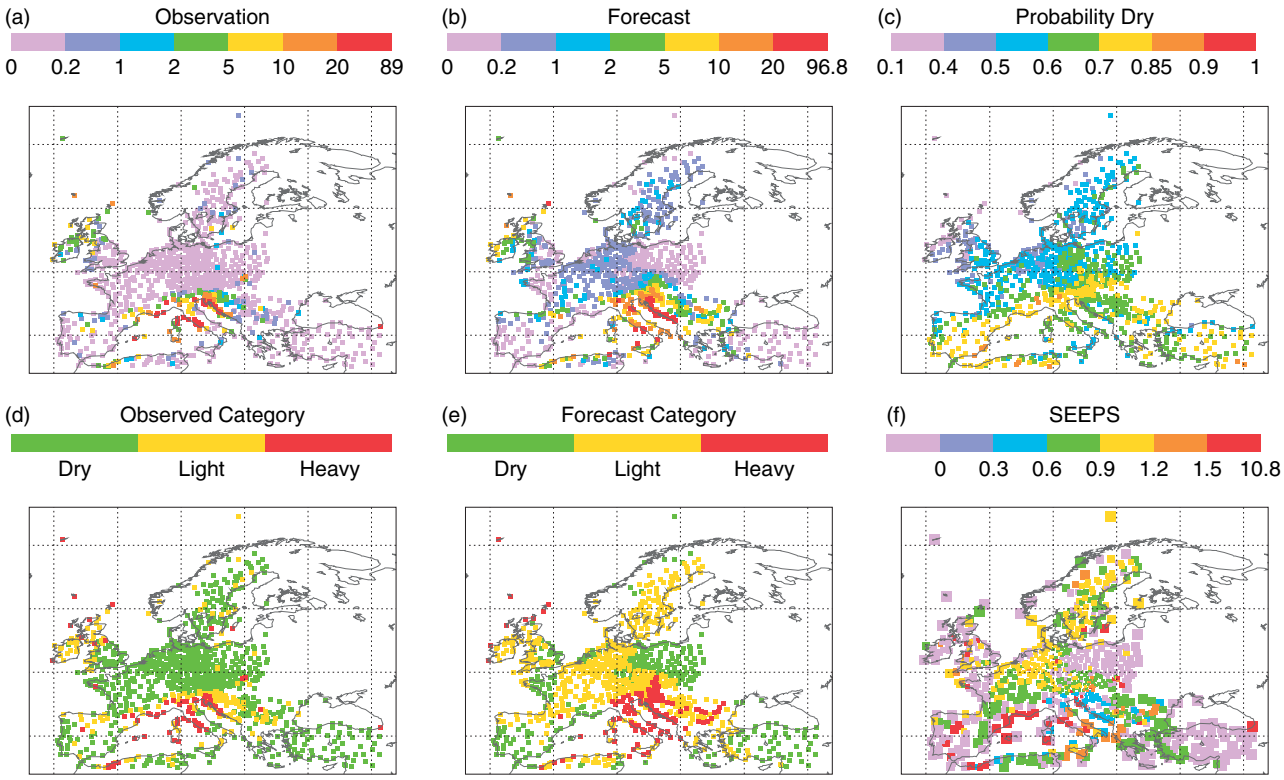

Figure 7. (a) Observed precipitation accumulated over 24 hours for 15 December 2008 at 1200 UTC to 16 December 2008 at 1200 UTC. (b) Forecast precipitation accumulated over lead times of 72–96 h and valid for the same period as the observations. (c) Probability of a 'dry' day in December, based on the 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.

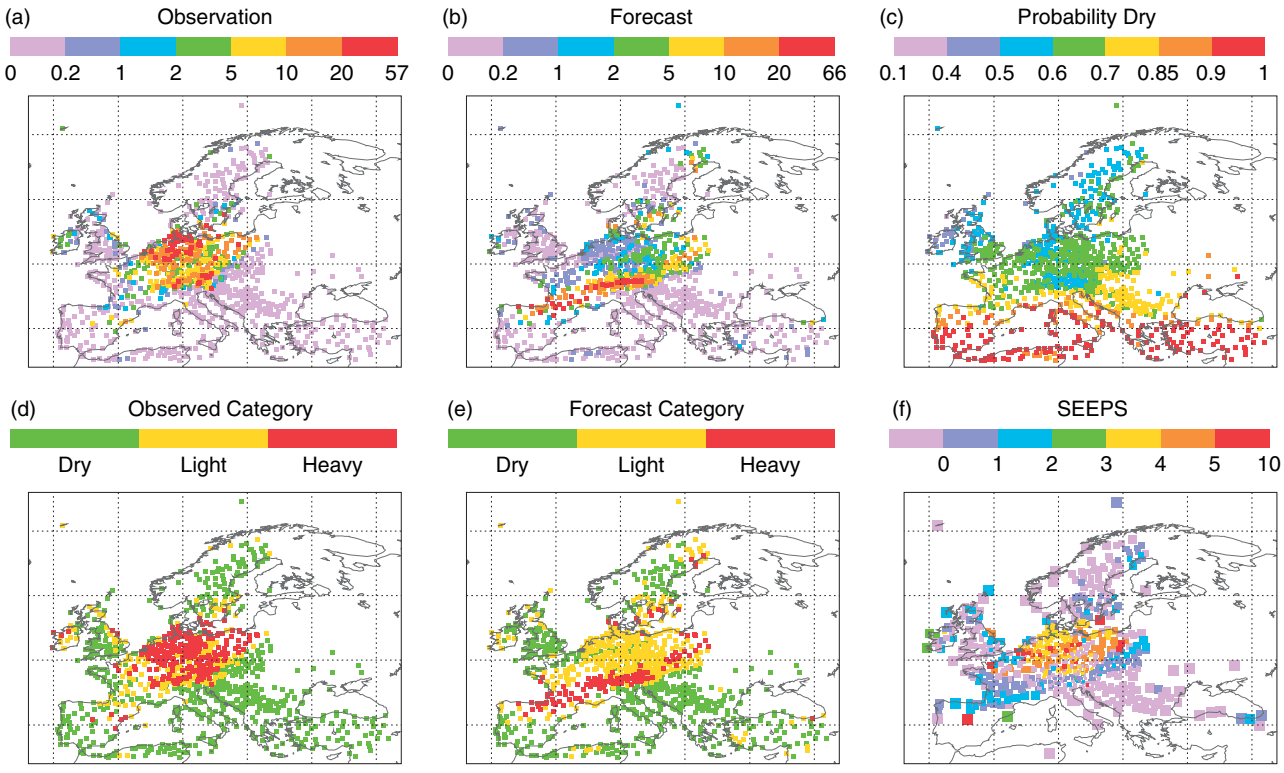

Figure 8. As Figure 7 but for the prediction of precipitation accumulated over the 24 hour period from 22 August 2008 at 1200 UTC to 23 August 2008 at 1200 UTC.

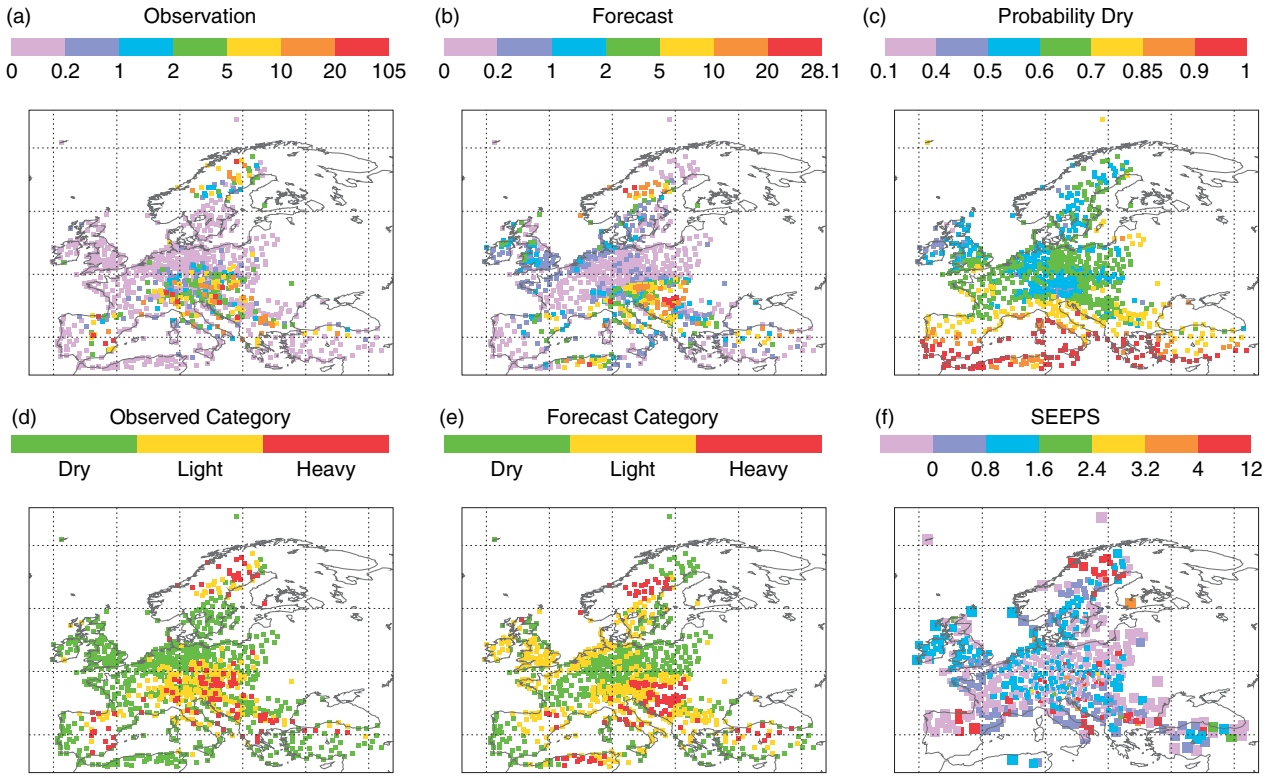

Figure 9. As Figure 7 but for the prediction of precipitation accumulated over the 24 hour period from 8 June 2008 at 1200 UTC to 9 June 2008 at 1200 UTC.

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 synoptic situation presented in Figure 1. The SYNOP observations (Figure 8(a)) show northeast Europe received over 10 mm, and up to 57 mm, of precipitation associated with a low-pressure system centred over Germany. The $D+4$ forecast (Figure 8(b)) had less than 5 mm (often less than 1 mm) in this region and instead predicted convective outbreaks along a front to the south. The category maps (Figure 8(d) and (e)) clearly show these issues. SEEPS also highlights both the errors (Figure 8(f)) but, since even northern Europe is generally dry in August (Figure 8(c)), it is the category difference indicating underprediction that leads to the largest scores.

Note that no SEEPS scores are plotted in Figure 8(f) for the southern Iberian peninsula, 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 (Figure 9). This case demonstrates that SEEPS can highlight the mislocation 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 initialization, 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 subregions 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:

$$\rho_k = \sum_l \mathrm{e}^{-(\alpha_{kl}/\alpha_0)^2}, \tag{22}$$

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. 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 $\alpha_0 = 0.75^\circ$ (83 km) is chosen because it is the smallest possible that ensures approximately equal representation of all subregions 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

$$\begin{aligned} w_k &= \frac{1}{\rho_k}, \\ S &= \frac{\sum_k w_k S_k}{\sum_k w_k}. \end{aligned} \tag{23}$$

As mentioned in section 8, the areas of the squares in Figures 7(f), 8(f) and 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 subregion is favoured over any other. Density weighting also ensures that Europe will not dominate a score of the extratropics so heavily in general. The methodology is currently being developed to utilize

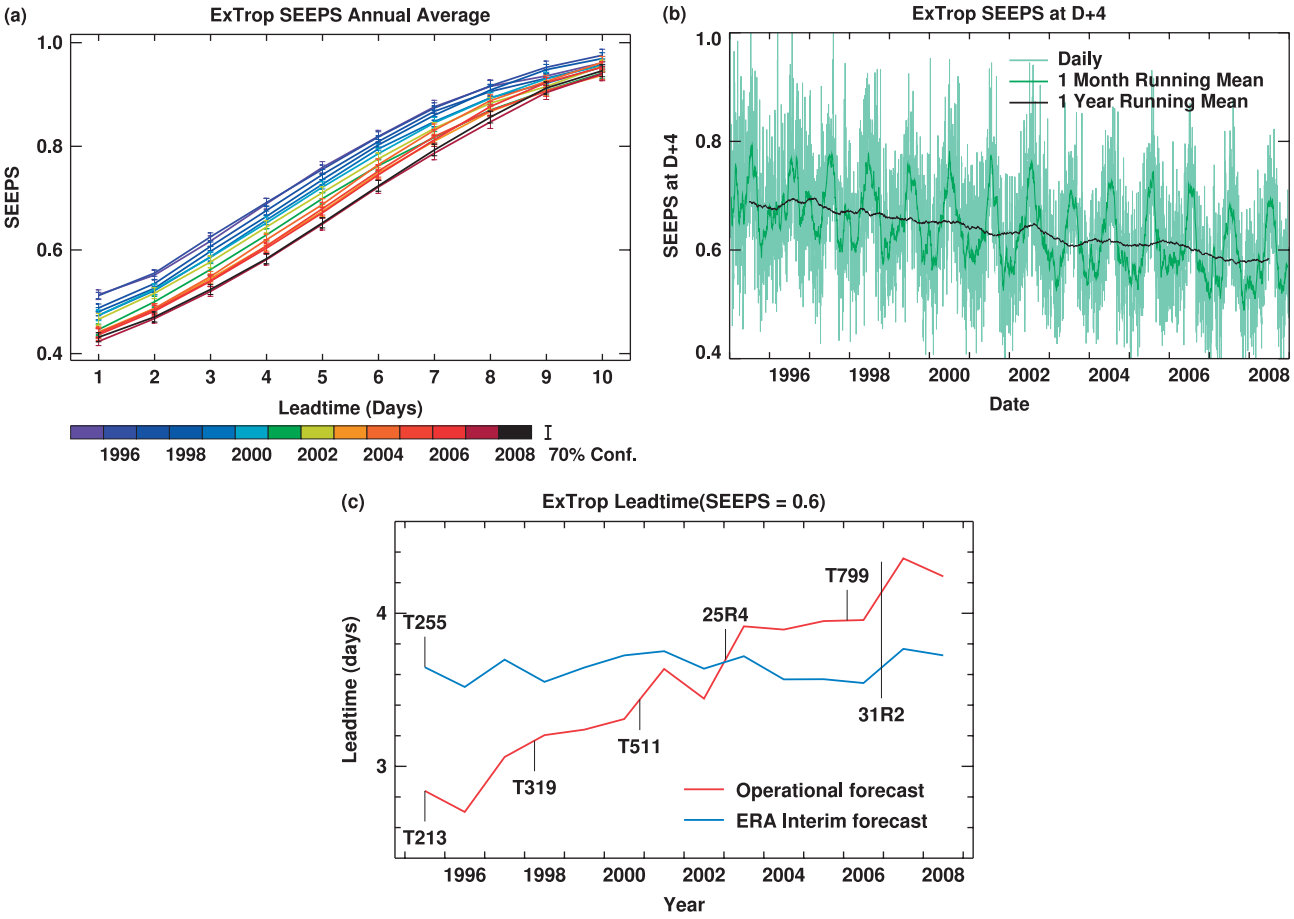

Figure 10. Extratropical mean SEEPS results. (a) Annual mean of daily operational scores as a function of lead time. 70% confidence intervals for these annual means are indicated. (b) Time series of operational scores at $D + 4$ with running means as indicated. (c) Annual mean lead time at which the 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°N and south of 30°S, taking account of observation density.

observations at all times of the day within the verification. This will mean that, for example, eastern Europe will be much better represented in area-mean scores than indicated in Figure 7(f), for example.

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°N *and* south of 30°S), based on ∼2000 stations per day, are shown in Figure 10. Figure 10(a) shows the annual mean scores based on the 1200 UTC operational forecasts as a function of lead time. 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 Figure 10(a) show the degree of uncertainty in the annual means. They are deduced from the daily scores taking autocorrelation into account following the Student's t-test 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 Figure 1(a)) cannot be considered as reliable daily observations at present. Nevertheless, current mean 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 Figures 3 and 4) and for the refinement constraint (section 4) to benefit development decisions.

It can be seen that by $D + 10$ the SEEPS score is tending towards 1. Evidently, imposing equitability in terms of expectation (13) is sufficient to ensure that annual-mean extratropical-mean error scores converge to 1 in the situation of no skill. Equitability makes the aggregation of all the stations within an area a meaningful and useful concept (despite subregions having very different climates).

Figure 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 an annual 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).

Figure 10(c) shows the annual mean of the lead time 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 annual-mean score at $D+4$. The red curve relates to the operational forecast data shown in Figure 10(a) and (b). The gains in lead time amount to $\sim$2 days over the 14 year period. The graph is annotated to show when the model's 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 Figure 10(c) shows comparable results for re-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 have not had a major impact on scores. More controversially, it also indicates that the increase in available sources and volume of data used to initialize the forecast (a 100$\times$ 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 the way the data assimilation system has improved to make better use of the available observations. New data sources will target the hydrological cycle more directly, so the conclusions from the 1995–2008 period may not hold in future.

9.3. *Europe*

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

Table XII. 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.

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

9.4. *South America*

The SEEPS scores for the South American region [70°W–35°W, 40°S–10°N] at $D+4$ (Figure 11(b)) show an improving trend although with a lot of variability. Close inspection of the data reveals an alarming annual 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 Figure 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 the ability of SEEPS to detect improvements. Since the improving trends in Figures 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 one-year mean curve) from it. Table XII 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 XII). 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 Figure 6(a) and (b)) with a more

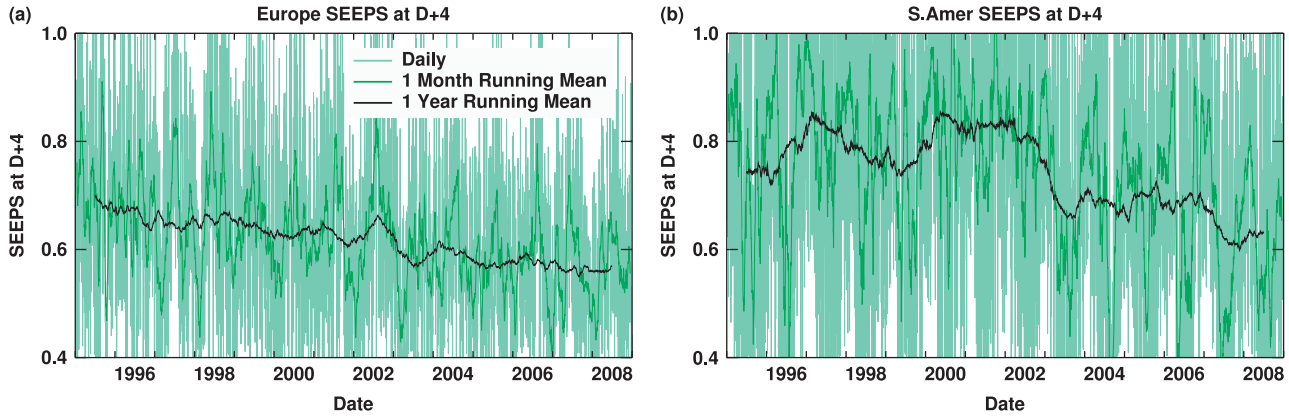

Figure 11. As Figure 10(b) but for (a) Europe [12.5°W–42.5°E, 35°N–75°N] and (b) South America [70°W–35°W, 40°S–10°N].

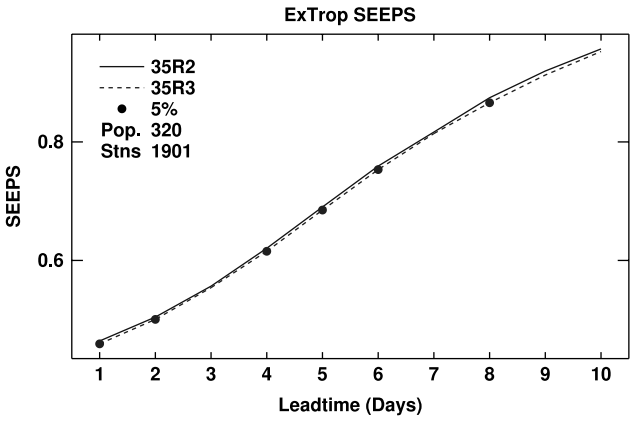

Figure 12. Mean extratropical SEEPS scores for two cycles of the ECMWF forecasting system as a function of lead time: 35R2 (solid) and 35R3 (dashed). A filled circle on a given curve indicates that the mean score for that model cycle is statistically significantly better than that of the other cycle at the 5% level using a two-sided, paired Student's *t*-test, taking autocorrelation into account. Results for both cycles are based on all 320 forecasts initiated at 0000 and 1200 UTC between 1 April 2009 at 1200 UTC and 8 September 2009 at 0000 UTC. 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.

marked deterioration in ability to detect extratropical trends.

10.2. Differences between forecast system cycles

When an experimental forecast suite (or 'cycle') is being assessed, a set of forecasts is 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 optimization 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 with 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, which 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 (rather than bilinear interpolation) and measuring forecast error in probability space. However, the severity 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 (0600–0600 UTC) 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 needs to be sufficient observations in each grid box. For the grid resolutions of ∼80, 40 and 25 km assessed here, minima of 40, 18 and 6, respectively, are specified. Groisman and Legates (1994) point out that area-mean precipitation can be biased in mountainous regions as most (US) 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.

Table XIII shows that the lower bound for SEEPS is reasonably small for all resolutions and gets smaller with increasing resolution. The implication is that the present operational forecast score is not limited by observation error or lack or representativity of a grid-box average.

12. 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 the '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 (Figure 2) is used to transform

Table XIII. 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.

Resolution		Min	Mean	SEEPS	
Spec.	Grid	high density	SYNOP used	Mean	70% Conf.
*T*255	80 km	40	239	0.278	0.014
*T*511	40 km	18	242	0.242	0.011
*T*799	25 km	6	198	0.204	0.009

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 (13), 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 'strong perfect forecast' constraints (12). These constraints differentiate SEEPS from the skill score of Gerrity (1992), figure 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 (Figure 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. The score penalizes most heavily forecasts for a climatologically likely category that turn out to be incorrect. This should encourage system developments that permit the model to represent all categories of local weather, whatever their climatological frequency.

Except for very poor forecast systems, some physical understanding of forecast error is required to improve 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 are used) so that it is possible to continuously monitor a system with resolution changing in 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 overprediction of drizzle (Figure 7), failure to predict heavy large-scale precipitation (Figure 8) and incorrectly locating convective cells (Figure 9).

The density of the observation network is taken into account when calculating area-mean scores. This implies, for example, that each subregion of Europe will contribute approximately equally to the European mean score. Area-mean results show an improving trend over the last 14 years (Figures 10 and 11). For the extratropics, this amounts to ∼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 (Figure 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 relatively little 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 forecast error is measured in probability space.

Experiments are under way to investigate whether SEEPS can be used to verify six-hour accumulations from 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 (Figure 10(c)). This indicates, strikingly, that the ∼hundredfold 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), a more prognostic treatment of precipitation and increasing model resolution.

Detailed and multifaceted 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 would be especially useful whenever the verification parameter has a difficult spatial or temporal distribution.

Acknowledgements

The authors thank Ian Jolliffe and two anonymous reviewers for their insightful comments, Anna Ghelli for the gridded station data, Martin Göber, Thomas Jung and Cristina Primo for valuable discussions and ECMWF's Advisory Committee on Verification for helpful guidance.

References

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

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

Casati B, Wilson LJ, Stephenson DB, Nurmi P, Ghelli A, Pocernich M, Damrath U, Ebert EE, Brown BG, Mason S. 2008. Forecast verification: current status and future directions. *Meteorol. Appl.* **15**: 3–18.

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

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

Gandin LS, Murphy AH. 1992. Equitable skill scores for categorical forecasts. *Mon. Weather Rev.* **120**: 361–370.

Gerrity JP. 1992. A note on Gandin and Murphy's equitable skill score. *Mon. Weather Rev.* **120**: 2709–2712.

Ghelli A, Lalaurette F. 2000. '*Verifying precipitation forecasts using upscaled observations*', ECMWF Newsletter 87. ECMWF: Reading, UK. Available at http://www.ecmwf.int/publications/.

Göber M, Zsóster E, Richardson DS. 2008. Could a perfect model ever satisfy a naïve forecaster? On grid box mean versus point verification. *Meteorol. Appl.* **15**: 359–365.

Gringorten II. 1967. Verification to determine and measure forecasting skill. *J. Appl. Meteorol.* **6**: 742–747.

Groisman PY, Legates DR. 1994. The accuracy of United States precipitation data. *Bull. Am. Meteorol. Soc.* **75**: 215–227.

Heidke P. 1926. Berechnung des Erfolges und der Güte der Windstärkevorhersagen in Sturmwarnungsdienst. *Geografiska Annaler* **8**: 301–349.

Hoffman RN, Liu Z, Louis JF, Grassoti C. 1995. Distortion representation of forecast errors. *Mon. Weather Rev.* **123**: 2758–2770.

Hogan RJ, O'Connor EJ, Illingworth AJ. 2009. Verification of cloud-fraction forecasts. *Q. J. R. Meteorol. Soc.* **135**: 1494–1511.

Hogan RJ, Ferro CAT, Jolliffe IT, Stephenson DB. 2010. Equitability revisited: Why the 'equitable threat score' is not equitable. *Weather and Forecasting*. **19**: 710–726.

Jolliffe IT, Foord JF. 1975. Assessment of long-range forecasts. *Weather* **30**: 172–181.

Murphy AH. 1978. Hedging and the mode of expression of weather forecasts. *Bull. Am. Meteorol. Soc.* **59**: 371–373.

Murphy AH, Winkler RL. 1987. A general framework for forecast verification. *Mon. Weather Rev.* **115**: 1330–1338.

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

Potts JM, Folland CK, Jolliffe IT, Sexton D. 1996. Revised 'LEPS' scores for assessing climate model simulations and long-range forecasts. *J. Climate* **9**: 34–53.

Roberts NM, Lean HW. 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Weather Rev.* **136**: 78–97.

Rodwell MJ. 2005. '*Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurrence better*', ECMWF Newsletter 106. ECMWF: Reading, UK. Available at http://www.ecmwf.int/publications/.

Simmons AJ, Uppala S, Dee D, Kobayashi S. 2007.. '*ERA-Interim: New ECMWF reanalysis products from 1989 onwards*', ECMWF Newsletter 110. ECMWF: Reading, UK.

Stephenson DB. 2000. Use of the 'Odds Ratio' for diagnosing forecast skill. *Weather and Forecasting* **15**: 221–232.

Stephenson DB, Casati B, Ferro CAT, Wilson CA. 2008. The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorol. Appl.* **15**: 41–50.

von Storch H, Zwiers FW. 2001.. *Statistical Analysis in Climate Research*. Cambridge University Press: Cambridge, UK; 484 pp.

Ward MN, Folland CK. 1991. Prediction of seasonal rainfall in the north Nordest of Brazil using eigenvectors of sea-surface temperature. *Int. J. Climatol.* **11**: 711–743.

