ECMWF Feature article

from Newsletter Number 128 – Summer 2011

METEOROLOGY

Developments in precipitation verification

www.ecmwf.int/en/about/news-centre/media-resources

doi:10.21957/x49qpn1o

This article appeared in the Meteorology section of ECMWF Newsletter No. 128 – Summer 2011, pp. 12–16.

Developments in precipitation verification

Mark J. Rodwell, Thomas Haiden, David S. Richardson

ECMWF's new strategy places more emphasis on the verification of weather parameters such as precipitation and near-surface wind. This change in emphasis is a result of user requirements and scientific developments. It led to the establishment of an ECMWF Technical Advisory Committee Sub-group on Verification Measures. The Sub-group recommended that some new headline scores be adopted to supplement our established primary headline scores anomaly correlation of 500 hPa geopotential, and continuous ranked probability score of 850 hPa temperature, see e.g. *Richardson et al.*, 2010). Among these supplementary scores is the newly developed 'SEEPS' score (*Rodwell et al.*, 2010) used for the verification of deterministic precipitation forecasts.

Here we explain the SEEPS score, and present examples of how it is being used to monitor and compare deterministic forecast performance, guide development decisions, and assess the spread–error relationship within the Ensemble Prediction System. Finally, we discuss potential future developments.

The SEEPS score

The task of forecasting precipitation beyond a day-or-two in advance is very much a probabilistic one, which must take account of a range of uncertainties. The ECMWF Ensemble Prediction System (EPS) takes account of uncertainties in initial conditions and sub-grid scale processes. Appropriate scores to assess the overall performance of probabilistic forecasts are 'proper' scores for which there is no benefit in hedging. Examples of such scores are those derived from the Brier and Ignorance Scores (e.g. *Gneiting & Raftery*, 2007).

As well as making probability forecasts, there is also a need to make high-resolution deterministic precipitation forecasts. High resolution is beneficial, for example, within the data assimilation process in order to produce the best initial conditions for subsequent forecasts. At short ranges, high-resolution precipitation forecasts provide complementary information to that provided by the lower-resolution EPS (*Rodwell*, 2006). In addition, the diagnosis and improvement of high-resolution deterministic forecast error prepares the model for future use at a higher-resolution within the EPS (on next-generation computers).

A score is required that can be used to monitor the performance of deterministic precipitation forecasts. Although probabilistic scores can sometimes be applied to deterministic forecasts, they are generally not appropriate. For example, the Brier Score and Ranked Probability Score unduly reward deterministic forecasts for always predicting the category containing the median. Instead it is more appropriate, for deterministic forecasts, to use 'equitable' scores which heavily penalise constant and purely random forecasts (*Gandin & Murphy*, 1992).

A number of equitable scores have been used in the verification of deterministic precipitation forecasts. Amongst the most common is the True Skill Score (TSS), also known as the Peirce Skill Score (PSS). This is based on a 2-category contingency table (for the occurrence of a given event) of the form:

		Observed	
		Yes	No
Forecast	Yes	Hits	False-alarms
	No	Misses	Correct-nulls

1–PSS can be written as:

$$1-\text{PSS} = \text{Miss rate} + \text{False alarm rate}$$

$$= \frac{\text{Misses}}{\text{Total events}} + \frac{\text{False alarms}}{\text{Total non-events}}$$

However, this score, along with others that are commonly used, does not appear to possess all the attributes desirable for routine monitoring of the performance of deterministic precipitation forecasts. A simple example is that it is impossible to assess the prediction of dry weather and precipitation-amount with only two categories. Because of this, a new equitable score ('SEEPS') has recently been developed by *Rodwell et al.* (2010).

SEEPS (Stable Equitable Error in Probability Space) uses three categories: 'dry', 'light precipitation' and 'heavy precipitation'. Here 'dry' is defined, with reference to WMO guidelines for observation reporting, to be any accumulation (rounded to the nearest 0.1 mm) that is less than or equal to 0.2 mm. To ensure that the score is applicable for any climatic region, the 'light' and 'heavy' categories are defined by the local climatology so that 'light' precipitation occurs twice as often as 'heavy' precipitation. Here a global 30-year climatology of SYNOP station observations is used, and the resulting threshold between the 'light' and 'heavy' categories ($T_{L/H}$ in Figure 1) is generally between 3 and 15 mm for Europe, depending on location and month. This approach to defining categories was motivated by the 'Linear Error in Probability Space' methodology of *Ward & Folland* (1991).

SEEPS can be written as the mean of two 2-category scores that individually assess the dry/light and light/heavy thresholds. Each of these 2-category scores is rather like the 1–PSS but written as:

$$\frac{\text{Misses}}{\text{Expected events}} + \frac{\text{False alarms}}{\text{Expected non-events}}$$

where the word 'expected' implies a climatological-mean rather than a sample-mean. The result is that SEEPS permits the construction of daily error time series that can be augmented as new data become available. A summary of the main attributes of SEEPS is given in Box A. All these attributes are important for monitoring purposes.

Here, SEEPS is used to compare 24-hour accumulations derived from global SYNOP observations (exchanged over the Global Telecommunication System; GTS) with values at the nearest model grid-point. Sometimes 1-SEEPS is preferred for presentational purposes as this provides a positively-oriented skill score.

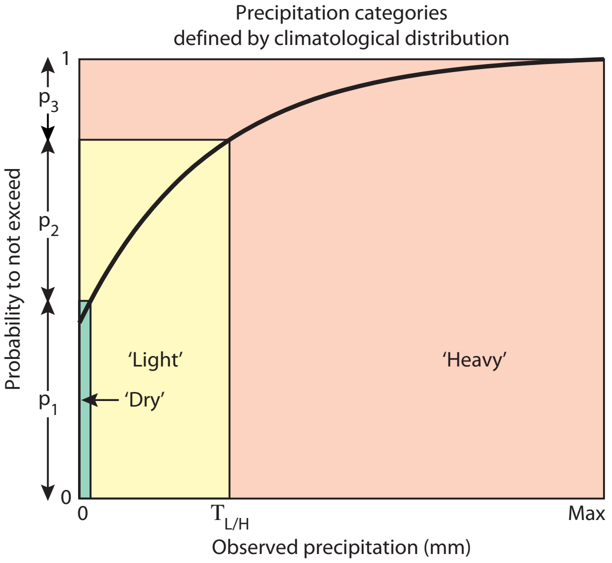
Precipitation categories
defined by climatological distribution
Probability to not exceed
1
0
p3
p2
p1
'Light'
'Heavy'
'Dry'
0
T L/H
Max
Observed precipitation (mm)

Figure 1 Schematic diagram showing how the probabilities and thresholds for the three SEEPS precipitation categories ('dry', 'light precipitation' and 'heavy precipitation') are determined from the climatological cumulative distribution (black curve).

The characteristics and benefits of SEEPS

A

Stable: SEEPS is designed to be as insensitive as possible to sampling uncertainty (for sufficiently skilful forecast systems). This allows more accurate trends to be extracted from noisy data.

Equitable Error: A perfect forecast has a SEEPS score of 0. The expected score increases linearly with the unskilled component of the forecast towards a maximum value of 1.

Probability Space: This is used to define precipitation categories; SEEPS adapts to the underlying climate to assess the pertinent aspects of local weather. It can be aggregated over heterogeneous climate regions.

Case studies

The diagnosis of short-range forecast error is particularly useful for parametrization development. Figure 2 shows how SEEPS highlights precipitation errors in a short-range forecast (the first 24 hours of the deterministic forecast initiated from 12 UTC on 22 April 2010). Although the large-scale synoptic flow was well forecast at this short-range, errors are evident in the precipitation field. For example, with the exception of a few places such as southern Sweden, most of northern Europe was dry at this time (Figure 2a) while the forecast developed up to 5 mm of precipitation within a northerly flow over Scandinavia and into Germany (Figure 2b). The forecast also developed too much precipitation within a warm front that extended from southern France to Bulgaria. Notice also that there is too much precipitation predicted along the Italian west coast associated with a second warm frontal system. Other features are well predicted such as the heavy precipitation along the Moroccan coast associated with on-shore winds.

Through use of the 30-year climatology (the climatological probability of an April day being dry is shown in Figure 2c), the precipitation fields are converted into the dry, light and heavy precipitation categories. The precipitation discrepancies highlighted above are clearly evident in the category fields (Figures 2d and 2e) and reflected in relatively large SEEPS errors (Figure 2f). Other case studies, which concentrate on medium-range forecast errors, are discussed in *Rodwell et al.* (2010).

SEEPS has been defined so that scores can be averaged over different climatic regions. To ensure that all sub-regions are correctly represented in an area-mean, the local observation density is taken into account. For example, the areas of the (small) squares in Figure 2f are proportional to the weights given to each individual score within the overall European-mean. The monitoring of area-mean scores, in order to chart progress with performance and inform development decisions, is likely to be a key use of the SEEPS score.

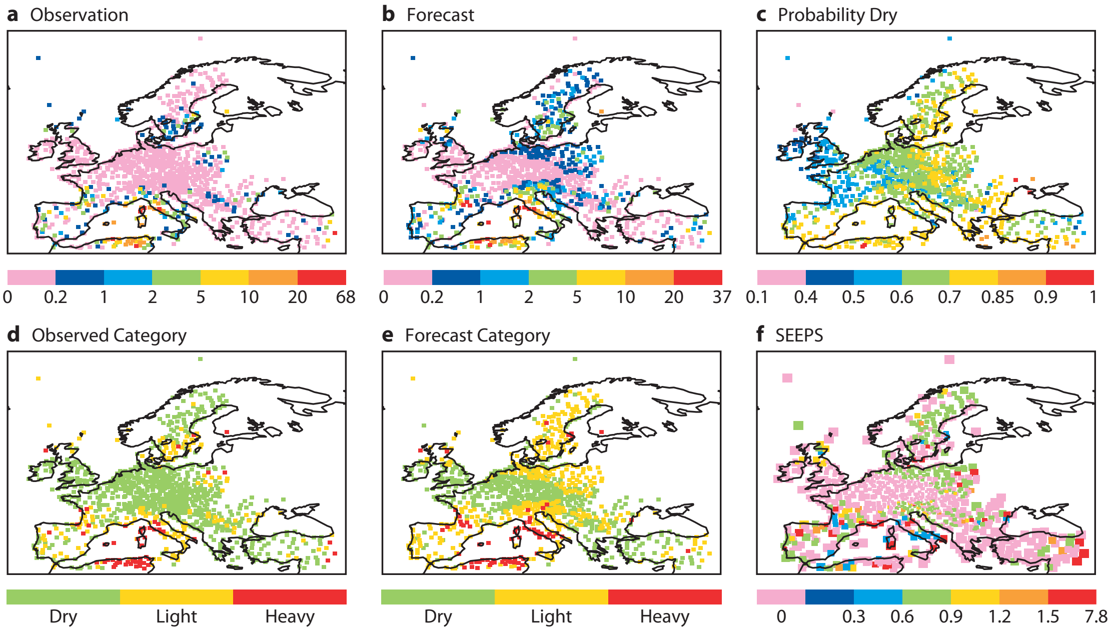

Figure 2 (a) Observed precipitation accumulated over the 24 hours to 12 UTC on 23 April 2010. (b) Forecast precipitation accumulated over lead-times 0 to 24 hours and valid for the same period as the observations. (c) Probability of a 'dry' day in April based on the 1980–2009 climatology. (d) Observed precipitation category. (e) Forecast precipitation category. (f) SEEPS. Units in (a) and (b) are mm. Squares in (f) are plotted at each observation point with areas proportional to the weight given to each station in the European area-mean score.

Score decomposition

For practical applications and further model development, it is of interest to know which kind of error ('dry' when 'light' predicted, 'light' when 'heavy' predicted etc.) contributes most to the total SEEPS. The off-diagonal panels in Figure 3 show these contributions as a function of forecast day for Europe in winter 2009/10. Large contributions are due to missed heavy events. Observed 'heavy' events which were forecast as 'light' contribute even at day 1. Observed 'heavy' events which were forecast as 'dry' contribute almost as much at long lead times, but such errors are rarer at short lead times. An error which is nearly independent of lead time is the prediction of 'light' when 'dry' was observed. The over-prediction of light precipitation is a well-known problem which can also be seen in the comparison of observed and forecast frequencies (given in the panels on the diagonal in Figure 3). Improvements in the cloud scheme aimed at alleviating this problem are currently being tested.

Score trends

Figure 4 shows the evolution of 1-SEEPS (a positively-oriented skill score) since 1993 for the extra-tropics and the tropics (the boundary defined at 30° latitude). The increase in skill has been largely the same for days 2 and 6 of the forecast, both in the extra-tropics and the tropics. It amounts to a lead-time gain of about 2 days. The difference in forecast skill between the extra-tropics and the tropics is considerable. It is equivalent to about 4 forecast days and has slightly increased over the period shown.

Since a one-year running mean filter has been applied in Figure 4, sudden improvements in skill associated with new model cycles appear as gradual ascents extending over one year, centred on the date of change. For example, the introduction of the prognostic cloud scheme in April 1995 (cycle Cy13r4) is apparent in the extra-tropics. Also major changes to the assimilation, cloud scheme and convective parametrization in January 2003 (cycle Cy25r4) are reflected in the curves of both the extra-tropics and the tropics.

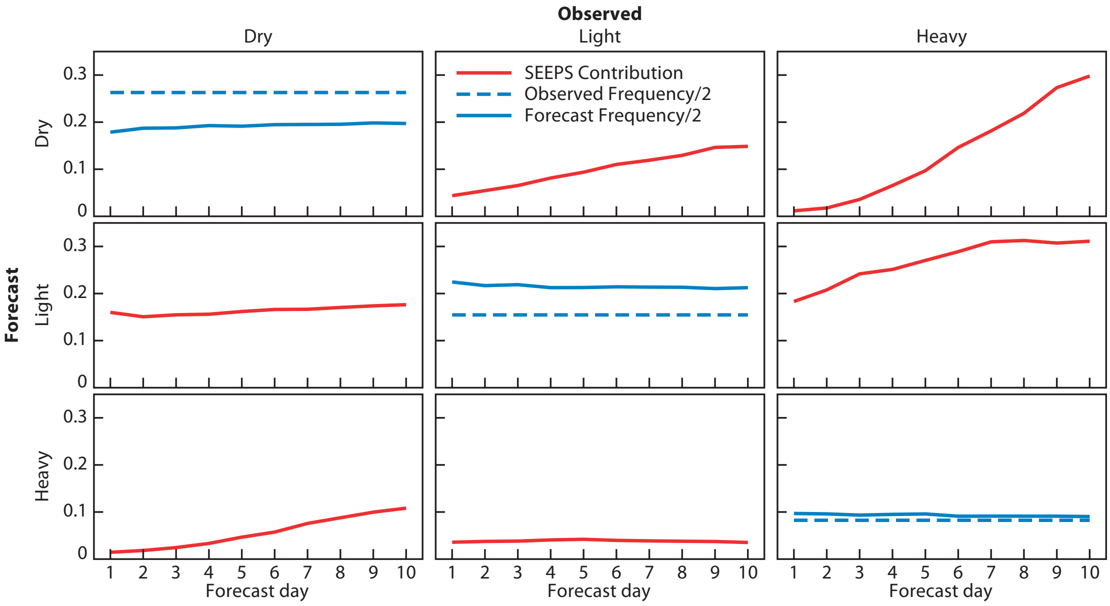

Figure 3 Off-diagonal panels show the contributions to SEEPS from each kind of forecast error as a function of forecast day. Panels on the diagonal show observed and forecast frequency of events. Results are for Europe during the period 1 December 2009 to 28 February 2010 (12 UTC forecasts).

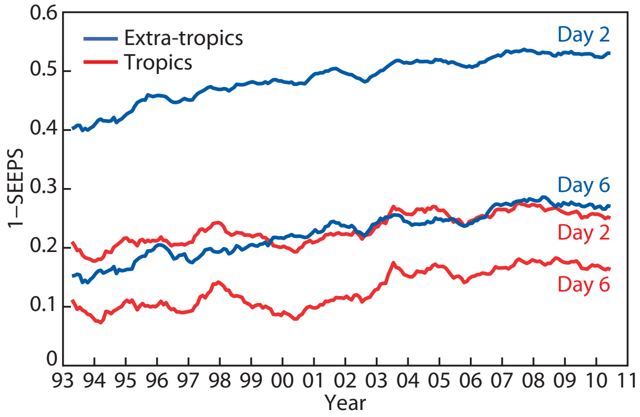

Figure 4 Long-term evolution of 1-SEEPS for the ECMWF model for forecast days 2 and 6 in the extra-tropics and the tropics with a one-year running-mean filter applied.

Model inter-comparison

Model inter-comparisons provide important information for both users and developers, and are part of the operational verification at ECMWF. Since March 2010 comparisons have been made between the skill of precipitation forecasts from the global models of the Japan Meteorological Agency (JMA), National Centers for Environmental Prediction (NCEP), UK Met Office and ECMWF. Verification against observations offers a large number of possibilities with regard to the choice of score, interpolation method, spatio-temporal aggregation, verification period, verification domain, and observation quality control. As a consequence, results from different studies are rarely directly comparable (*Ebert et al.*, 2003). Here we use the same methodology with regard to data preprocessing, interpolation, and score computation for all available models, ensuring maximum compatibility of results.

Figure 5 shows a time-series of 1-SEEPS of the four models (NCEP data is available from June 2010 only) for forecast day 4 for the extra-tropics. Day-to-day variations are smoothed by the weekly averaging but strong variations are present also on the weekly to seasonal timescales and shared by all the models. The reduction of skill during the northern hemisphere convective season (May to August) is noticeable in the global score because there are many fewer SYNOP stations in the southern hemisphere (the weighting methodology does not completely compensate for this lack of observations). Skill differences between models are comparable in size to the weekly and monthly variations. The ECMWF model shows a robust and statistically significant lead.

Analysis of results for individual continents and for other lead times confirms the general ranking seen in Figure 5, although the differences are not always as large. In the shortest range (forecast days 1 and 2), the UK Met Office and ECMWF models exhibit very similar SEEPS values.

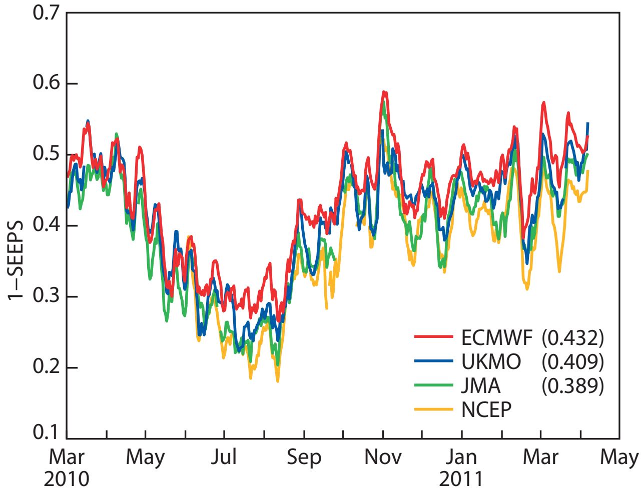

Figure 5 Precipitation forecast model inter-comparison for the extra-tropics for day 4 using 1-SEEPS. The verification period is 1 March 2010 to 5 April 2011 (12 UTC forecasts), with NCEP data available from 1 June 2010 onwards. Shown are running weekly averages of 1-SEEPS for the global models of ECMWF, UK Met Office (UKMO), Japan Meteorological Agency (JMA) and National Centers for Environmental Prediction (NCEP). Numbers in parentheses are period averages.

doi:10.21957/x49qpn1o

Evaluation of parallel suites

Before each change to the forecasting system, the proposed new model cycle is tested in parallel with the operational system. Cy36r4 (which became operational in November 2010) involved several changes that could have directly affected precipitation forecasts. It included a change to a five species prognostic microphysics scheme, with cloud rainwater content and cloud ice water content as new model variables. There was also a retuning and simplification of convective entrainment/detrainment and a land/sea dependent threshold for precipitation formation. Cy36r4 was tested over the period 1 July 2010 to 8 November 2010 in parallel with the operational cycle at the time (Cy36r2). Figure 6 shows the positive impact on 1-SEEPS scores. The most pronounced and highly statistically significant increase in skill was found for the extra-tropics at short lead times. In the tropics the improvement was seen to persist to longer lead times, but not to reach the same level of statistical significance.

Spread–error relationship

The SEEPS score has also been tested with regard to its usefulness in the analysis of the spread–error relationship in the EPS. The approximate equivalence of long-term mean spread and error is usually established by tuning the specification of uncertainties in the initial conditions and sub-grid scale processes with regard to 500 hPa geopotential height and 850 hPa temperature. Consequently, it is of some interest to complement this by looking at the spread–error relationship for surface fields such as precipitation. SEEPS may be useful for this purpose because of the way it handles the difficult distribution of precipitation and its normalizing characteristics with regard to climatology; also, importantly, SEEPS places emphasis on the dry/wet boundary. Ensemble error is calculated here as the mean of the SEEPS of each ensemble member against the observations. Ensemble spread is calculated as the mean of the SEEPS of each ensemble member against each other ensemble member.

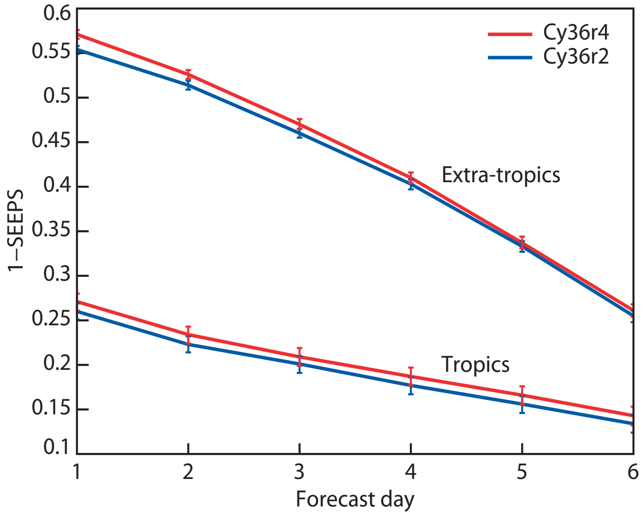

Figure 6 1-SEEPS scores for Cy36r4 (red) and Cy36r2 (blue) for the extra-tropics and the tropics as a function of lead time, averaged over the period 1 July to 9 November 2010 (12 UTC forecasts). Error bars show 95% confidence intervals calculated by re-sampling.

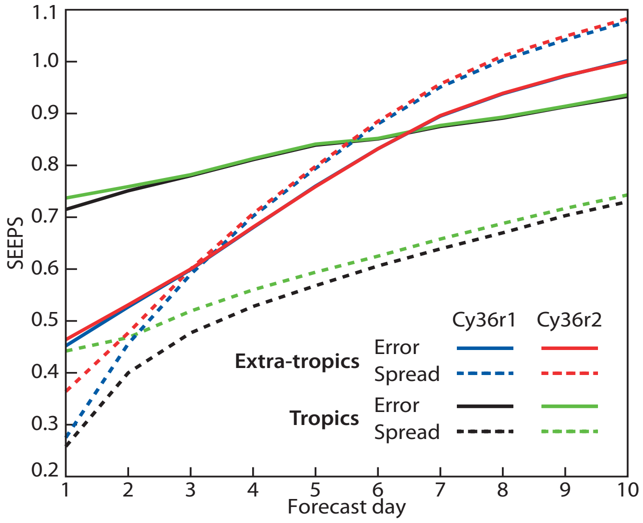

Figure 7 SEEPS error and spread of EPS precipitation forecasts from 00 UTC runs for a period of 56 days in the first half of 2010 for the extra-tropics and tropics. The operational suite at the time was Cy36r1 and the e-suite containing the EDA was Cy36r2.

doi:10.21957/x49qpn1o

Figure 7 shows the SEEPS spread–error relationship for Cy36r1 and Cy36r2. The difference between the two cycles is that Cy36r2 uses the Ensemble of Data Assimilations (EDA) as well as singular vectors to create the initial perturbations for the EPS. It became operational in June 2010. In the extra-tropics, there is reasonable correspondence between spread and error at Cy36r1 (blue lines). Interestingly the apparent under-dispersion at short lead times and over-dispersion at longer lead times is not seen in the upper-air fields. Further work is required to understand if SEEPS is indicating a true mismatch in spread and error. The EDA improves the spread-error relationship in the extra-tropics mostly on forecast day 1 (red lines). In the tropics the correspondence between spread and error at Cy36r1 is poorer (black lines). Although the increase of spread with lead time parallels that of the error, it does so at too low a level. This under-dispersion is also seen in the upper-air fields. The EDA (green lines) again helps to improve the spread at short ranges.

Future developments

To improve the coverage and robustness of global precipitation verification, it should be attempted to close remaining gaps in the areal distribution of precipitation observations obtained from the GTS. As model output frequency increases (currently 3-hourly for the ECMWF model), and with algorithm developments, it will be possible to verify against observations at times other than 0 and 12 UTC (such as from Finland, India, and Australia, for example).

The impact of observation uncertainty and representativeness on scores was quantified for 24-hour accumulations based on rain gauge data in Rodwell et al. (2010), but there are plans to extend this analysis. For example, high-resolution precipitation analyses combining rain gauge and radar data (*Haiden et al.*, 2011) will be used to better assess sub-grid scale variability and shorter accumulation periods. The hope being that the diurnal cycle can be partially resolved, and the spread–error relationship better assessed.

The SEEPS categories can also be used within a proper score (such as the Ranked Probability Score) for the probabilistic verification of the EPS. The combined approach provides a natural and 'seamless' way of applying the attributes of equitability and propriety to the entire Integrated Forecasting System. It also permits the assessment of the dry/wet boundary within the probabilistic system, and thus complements the frequently used Continuous Ranked Probability Score. Additional tests, sensitivity studies and theoretical work will be carried out to assess the utility of this approach.

Further reading

Ebert, E.E., U. Damrath, W. Wergen & **M.E. Baldwin,** 2003: The WGNE assessment of short-term quantitative precipitation forecasts. *Bull. Am. Meteorol. Soc.*, **84**, 481–492.

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

Gneiting, T. & **A.E. Raftery,** 2007: Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.*, **102**, 359–378.

Haiden, T., A. Kann, C. Wittmann, G. Pistotnik & **C. Gruber,** 2011: The Integrated Nowcasting through Comprehensive Analysis (INCA) system and its validation over the eastern alpine region. *Wea. Forecasting*, **26**, 16–183.

Richardson, D.S., J. Bidlot, L. Ferranti, A. Ghelli, T. Hewson, M. Janousek, F. Prates & **F. Vitart,** 2010: Verification statistics and evaluations of ECMWF forecasts in 2009–2010. *ECMWF Tech. Memo. No. 635*, ECMWF, Reading, UK.

Rodwell, M.J., 2006: Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurrence better. *ECMWF Newsletter* **No.106**, 17–23.

Rodwell, M.J., D.S. Richardson, T.D. Hewson & **T. Haiden,** 2010: A new equitable score suitable for verifying precipitation in numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **136**, 1344–1363.

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

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

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

doi:10.21957/x49qpn1o