

Intercomparison of global model  
precipitation forecast skill in 2010/11  
using the SEEPS score

T. Haiden, M.J. Rodwell,  
D.S. Richardson, A. Okagaki<sup>1</sup>,  
T. Robinson<sup>2</sup> and T. Hewson<sup>3</sup>

Operations Department

<sup>1</sup> Japan Meteorological Agency, Tokyo, Japan

<sup>2</sup> Canadian Meteorological Centre, Environment Canada, Dorval, Canada

<sup>3</sup> Met Office, Exeter, United Kingdom

January 2012

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



Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:

<http://www.ecmwf.int/publications/>

Contact: [library@ecmwf.int](mailto:library@ecmwf.int)

© Copyright 2012

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

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

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

## Abstract

Precipitation forecasts from five global numerical weather prediction (NWP) models are verified against raingauge observations using the new Stable Equitable Error in Probability Space (SEEPS) score. It is based on a 3×3 contingency table and measures the ability of a forecast to discriminate between ‘dry’, ‘light precipitation’, and ‘heavy precipitation’. In SEEPS, the threshold defining the boundary between the ‘light’ and ‘heavy’ categories varies systematically with precipitation climate. Results obtained for SEEPS are compared to those of more well-known scores, and are broken down with regard to individual contributions from the contingency table. It is found that differences in skill between the models are consistent for different scores, but are small compared to seasonal and geographical variations, which themselves can be largely ascribed to the varying prevalence of deep convection. Differences between the tropics and extra-tropics are quite pronounced. SEEPS scores at forecast day 1 in the tropics are similar to those at day 6 in the extra-tropics. It is found that the model ranking is robust with respect to choices in the score computation. The issue of observation representativeness is addressed using a ‘quasi-perfect model’ approach. Results suggest that just under one-half of the current forecast error at day 1 in the extra-tropics can be attributed to the fact that grid-box values are verified against point observations.

## 1 Introduction

Increases in resolution and developments in model physics have contributed to a substantial reduction of errors in global numerical weather prediction (NWP) models (Simmons and Hollingsworth 2002, Shapiro et al. 2010). Improvements are well documented for the mass and wind fields, and inter-model comparisons are performed for these fields on a regular basis (Richardson et al. 2009). Quantitative precipitation forecast (QPF) skill is less straightforward to evaluate on a global scale, due to the lack of reliable analyses and the uneven and often sparse distribution of raingauges. Model intercomparisons therefore typically focus on specific regions where high-resolution observational datasets are available (e.g., Csima and Ghelli 2008). While these studies provide valuable insight into forecast error characteristics for a given region, it is uncertain how representative the verification results, and model rankings derived from them, are for other regions. Usually weather services perform operational QPF verification in their domains of interest. Due to the large number of possibilities with regard to the choice of score, interpolation method, spatio-temporal aggregation, and observation quality control, results from different studies are rarely directly comparable.

In the mid-1990s, the World Meteorological Organization (WMO) Working Group on Numerical Experimentation (WGNE) established a common framework for QPF verification of NWP models. Some centres began verifying QPFs from several NWP models against data from their national rain gauge networks. Results of the WGNE QPF verification from 1997 through 2000 were summarized by Ebert et al. (2003). It was noted that a direct comparison of the results from different models and areas was not possible due to differences in verification methodology and rainfall observation time. To address this problem, the WMO issued recommendations for the verification and intercomparison of QPF from operational NWP models (WMO 2008).

There is considerable practical and scientific interest in more detailed information about variations in precipitation forecast skill between, for example, the tropics and extra-tropics, or between different seasons. Such comparisons are not straightforward due to geographical variations in precipitation climate. For a given fixed precipitation threshold that is used to create a contingency table, the base rate (frequency of occurrence) may differ strongly between regions. Pooling of data from a larger number of stations may render the verification statistics more robust but may also produce spurious skill if climatologically diverse regions are combined (Hamill and Juras 2006). It is important to take this into account in model intercomparison studies, as differences in overall skill between some NWP models may be rather subtle.

One way of making regions comparable is to use variable thresholds depending on precipitation climatology. Rodwell et al. (2010) followed this approach by developing the Stable Equitable Error in Probability Space (SEEPS) score which uses the categories ‘dry’, ‘light precipitation’, and ‘heavy precipitation’ based on the climatological cumulative precipitation distribution.

The objective of this study is a global model QPF skill intercomparison that covers, as far as possible, all continents, and uses a uniform data processing and verification methodology. In this way we hope to provide a more comprehensive picture of geographical and seasonal variations in precipitation forecast skill of current operational global NWP models. In order to check the robustness of the results, a re-sampling technique is applied, and sensitivity analyses are performed.

## 2 Data and methodology

### 2.1 Forecast data

Operational forecasts of 24-h precipitation totals from CMC (Canadian Meteorological Centre), JMA (Japan Meteorological Agency), NCEP (National Centers for Environmental Prediction), UKMO (United Kingdom Met Office), and ECMWF (European Centre for Medium-Range Weather Forecasts) global models (Table 1) are examined. Lead times from 24 to 144 hours (i.e. day 1 to day 6: D+1 to D+6) are analyzed. We use the convention of assigning to a precipitation value the time at the end of the interval. The D+3 forecast of 24-h precipitation, for example, corresponds to the precipitation between forecast lead times 48 and 72 h. The common verification period is 1 June 2010 – 30 April 2011 (11 months). Only 12 UTC runs are considered. Due to occasional missing forecasts from individual models in the data archive, only 319 out of the 334 days were actually used in the comparison.

The forecast datasets available in this study have grid spacings ranging from  $0.14^\circ$  to  $0.50^\circ$ . Table 1 shows the corresponding resolutions of the actual model grids. Nearest-neighbour matching between rain gauges and grid points is performed on each output grid, without additional interpolation steps. It may be argued that this reduces compatibility of the computed scores. In this study we compare the different models on their respective output grids but in addition we investigate in section 4c the sensitivity of the results to grid spacing. It turns out that the grid-spacing is not that critical to SEEPS, so the fact that some of the data was re-gridded prior to receipt (top two rows on Table 1) is probably not that critical either.

At the end of this study the authors became aware of a higher resolution version of the NCEP precipitation forecast ( $0.313^\circ$  before 28 Jul 2010,  $0.205^\circ$  afterwards) compared to the  $0.5^\circ$  data used here. Direct comparison of scores from the two datasets for a test period in August 2011 showed that the differences were small compared to the differences between the models (consistent with the sensitivity analysis in section 4c), and that the lower resolution dataset produced slightly better results. So the lower resolution dataset was used in the comparison.

*Table 1: Output grids of the models available for the intercomparison (12 UTC runs), and basic model characteristics.*

	<b>CMC</b>	<b>JMA</b>	<b>NCEP</b>	<b>UKMO</b>	<b>ECMWF</b>
<b>Output grid</b>	$0.3^\circ \times 0.3^\circ$	$0.25^\circ \times 0.25^\circ$	$0.5^\circ \times 0.5^\circ$	$0.35^\circ \times 0.23^\circ$	* $0.14^\circ \times 0.14^\circ$
<b>Model grid spacing (truncation)</b>	33 km	20 km (T959)	35 km (T382) before 28 Jul 2010, 27 km (T574) afterward	25 km	16 km (T1279)
<b>Model levels</b>	80	60	64	70	91
<b>Time step</b>	900 s	600 s	450 s	600 s	600 s
<b>Assimilation</b>	4D-Var	4D-Var	GSI 3D-Var Kleist et al. (2009)	Hybrid 4D-Var/ Ensemble	4D-Var
<b>Cloud microphysics</b>	Sundqvist et al. (1989)	Sundqvist et al. (1989)	Sundqvist et al. (1989), Zhao and Carr (1997)	Wilson et al. (2008), Abel et al. (2010)	5-species, prognostic Forbes et al. (2011)
<b>Deep convection</b>	En-/detraining plume, buoyancy sorting Kain and Fritsch (1990)	Arakawa-Schubert	Arakawa-Schubert Pan and Wu (1995)	Mass-flux Gregory and Rowntree (1990)	Tiedtke mass-flux, non-equilibrium Bechtold et al. (2008)

\* Variable longitudinal spacing of reduced Gaussian grid

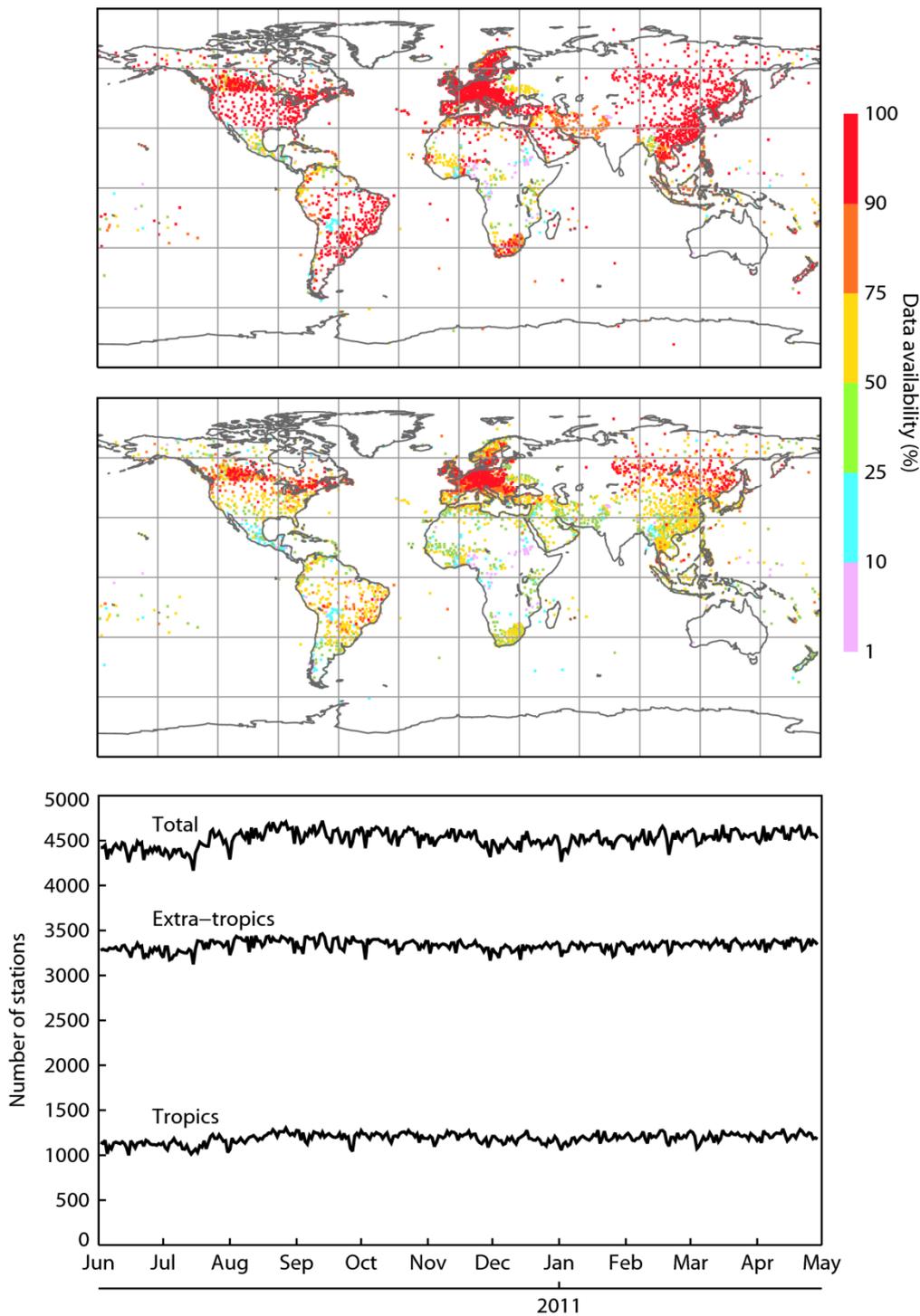


Figure 1: Geographical distribution of surface station data availability before (top) and after (center) exclusion of dry months, and time-series of the number of available stations (bottom).

## 2.2 Observations

The observational data used in this study consists of 24-h precipitation totals ending at 12 UTC, derived from 6-hourly and 12-hourly SYNOPs available from the Global Telecommunication System (GTS). Figure 1a shows the geographical distribution of available observations during the period of interest. Outside the polar regions, station density is highest in Europe and lowest in Africa. The extra-tropics, defined here as poleward of 30° latitude, are generally better represented than the tropics. Data from some regions in Eastern Europe and Asia, as well as from Australia, are missing because of non-standard reporting times. Having only 6-hourly (or longer) output available from most of the models, such observations could not be used in the comparison. The gap in central parts of Russia is due to a non-standard reporting practise which inhibited proper decoding of the GTS messages. Figure 1c shows the number of stations as a function of time within the verification period. In total, data from about 4500 stations are available of which about 1200 are from the tropics. Numbers are relatively constant over time, with typical short-term fluctuations of  $\pm 100$  stations.

## 2.3 Scores

Many different scores are in use in precipitation verification. Among the most common ones are the Equitable Threat Score (ETS), and the Peirce Skill Score [PSS, Peirce (1884)] also known as the True Skill Statistic (TSS). These scores are based on contingency tables, and the sample climatology is used as the reference forecast which appears in the denominator of the skill score. Rodwell et al. (2010) recently developed the SEEPS score so that it would have the attributes desirable for routine monitoring of weather forecast performance. It simultaneously assesses the prediction of dry weather and precipitation quantity and makes use of a global precipitation climatology.

While this verification study focuses on SEEPS, the ETS and PSS are computed as well for easier comparison with other studies. The frequency bias (FBI), although not a skill score in itself, is useful for understanding differences in behaviour of the different scores. These three dichotomous measures are defined by

$$\text{ETS} = \frac{a - a_{REF}}{a - a_{REF} + b + c} \quad \text{with} \quad a_{REF} = \frac{(a+b)(a+c)}{a+b+c+d} \quad (1)$$

$$\text{PSS} = \frac{ad - bc}{(a+c)(b+d)} \quad (2)$$

$$\text{FBI} = \frac{a+b}{a+c} \quad (3)$$

where  $a$  (hits),  $b$  (false alarms),  $c$  (misses),  $d$  (correct-nulls) are the entries in the usual  $2 \times 2$  contingency table (Wilks, 2006).

The PSS is the only 2-category score that satisfies the equitability conditions of Gandin and Murphy (1992). Gerrity (1992) demonstrated how an equitable n-category skill-score could be defined as the mean of  $(n-1)$  PSSs. The SEEPS score is based on three categories ('dry', 'light precipitation' and 'heavy precipitation') and can be constructed in a similar way, as the mean of two 2-category scores

that assess the dry/light and light/heavy thresholds, respectively. Each 2-category score is a modified PSS. Firstly, the PSS is often written as the hit rate minus the false alarm rate, but it can also be written as  $1 - (\text{miss rate} + \text{false alarm rate})$ . SEEPS is based on just the (miss rate + false alarm rate) and, as such, it is an error-score rather than a skill-score. Secondly, the miss rate and false alarm rate are calculated using the climatological base rate over the fixed 30-year period 1980-2009 derived from all SYNOP observations discussed in section 2b. The result is that SEEPS permits the construction of daily score timeseries that can be augmented as new data become available, it is asymptotically equitable (with increasing sample size), and it is asymptotically stable (with increasing forecast skill). All these attributes are important for monitoring purposes.

The ‘dry’ category (occurring with climatological probability  $p_1$ ) includes all cases with precipitation strictly less than 0.25 mm, taking into account the World Meteorological Organization (WMO) guidelines for observation reporting (Rodwell et al. 2010). To ensure that the score is applicable for different climatic regions, the threshold between the ‘light’ and ‘heavy’ categories is defined by the local climatology; the choice was made to state that ‘light’ precipitation occurs twice as often as ‘heavy’ precipitation, so that their climatological probabilities of occurrence are related by  $p_3 = p_2/2$  (see Fig. 2). Averaged over the extra-tropics, the resulting threshold between the ‘light’ and ‘heavy’ categories ( $t_{L/H}$  in Figure 2) is of the order of 5 mm in winter and 10 mm in summer. It varies strongly (typically between 2 and 20 mm) between individual locations and months.

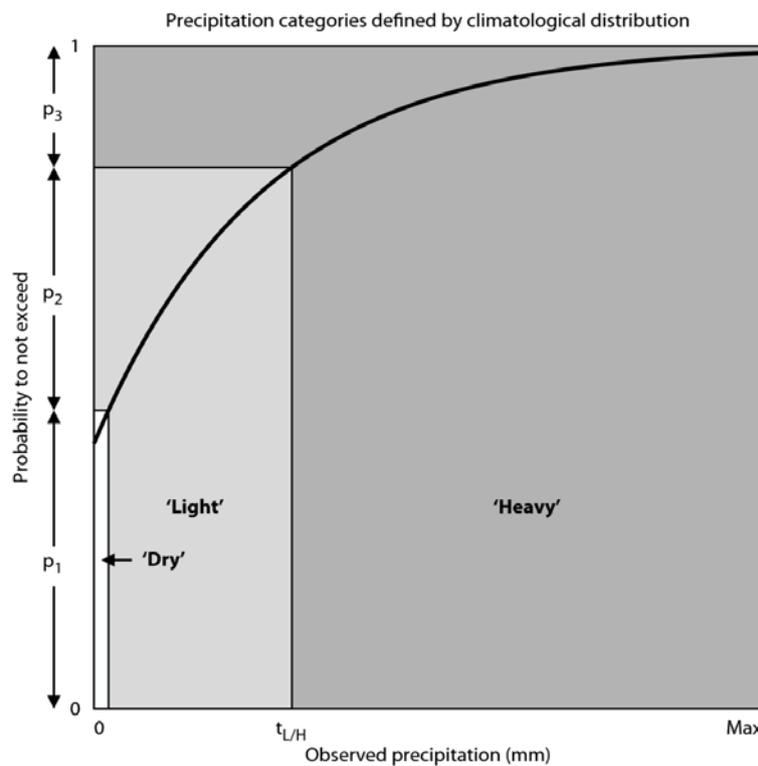


Figure 2: Schematic diagram showing a typical cumulative distribution for 24-hour precipitation at a given location. There is a probability  $p_1$  of it being ‘dry’. The ‘wet’ days are divided in the ratio 2:1 with a fraction  $p_2$  of all days being said to have ‘light’ precipitation and a fraction  $p_3$  ( $=p_2/2$ ) of all days said to have ‘heavy’ precipitation. Hence the threshold (in mm precipitation) between the ‘light’ and ‘heavy’ categories depends on the local climatology.

Following Gandin and Murphy (1992), SEEPS is written as the scalar product of the  $3 \times 3$  contingency table and a scoring matrix. Rodwell et al. (2010) wrote the scoring matrix in terms of  $p_1$  and  $p_3$  to emphasise its symmetry with respect to dry and heavy categories. Here, the relationships  $p_3 = p_2/2$  and  $p_1 + p_2 + p_3 = 1$  are used to write the scoring matrix in terms of  $p_1$  alone

$$\mathbf{S} = \frac{1}{2} \begin{Bmatrix} 0 & \frac{1}{1-p_1} & \frac{4}{1-p_1} \\ \frac{1}{p_1} & 0 & \frac{3}{1-p_1} \\ \frac{1}{p_1} + \frac{3}{2+p_1} & \frac{3}{2+p_1} & 0 \end{Bmatrix}. \quad (4)$$

In this matrix the observed categories ‘dry’, ‘light’, and ‘heavy’ run from left to right, and the corresponding forecast categories from the top down. Hence, a forecast for light precipitation when it is observed to be dry incurs a penalty of  $1/(2p_1)$ . SEEPS is a negatively oriented score with values between 0 and a maximum expected value of 1 (over a sufficiently long period of time) for an unskilled forecast. If used as a skill score the value of  $1 - \text{SEEPS}$  is plotted.

Months for which there is a very low climatological frequency of non-dry days ( $p_1$  close to 1, leading to large weights in the upper-right half of the scoring matrix) are excluded from the verification in order to reduce sensitivity to sampling uncertainty. This reduces the effective observation availability especially in the tropics and sub-tropics, and also in arctic and sub-arctic areas (Fig. 1b). The sensitivity of the results to the threshold defining a ‘climatologically very dry month’ is studied in section 4a.

## 2.4 Areal averaging and aggregation

SEEPS is a direct measure of error and can be averaged over a set of stations. Arithmetic averaging, however, would over-emphasize regions with high station density. To alleviate this effect, a weighting factor has been introduced by Rodwell et al. (2010). The station density in the vicinity of the  $k$ -th station,  $\rho_k$ , is calculated using a Gaussian kernel

$$\rho_k = \sum_l e^{-(\alpha_{kl}/\alpha_0)^2}, \quad (5)$$

where the summation extends over all stations available on a given day,  $\alpha_{kl}$  is the angle subtended at the centre of the Earth between stations  $k$  and  $l$ , and  $\alpha_0$  is a reference angle. The weighted area-mean SEEPS score  $S$  is computed from SEEPS scores  $S_k$  at individual stations by

$$S = \frac{\sum_k \frac{S_k}{\rho_k}}{\sum_k \frac{1}{\rho_k}}. \quad (6)$$

A standard value of  $\alpha_0=0.75^\circ$  is used in this study. The sensitivity of area-averages of SEEPS with respect to this parameter is analysed in Section 4b.

### 3 Verification results

#### 3.1 SEEPS results

Figure 3 illustrates the temporal variation of precipitation forecast skill at D+3 as measured by 1-SEEPS (higher values indicating more skill) in the extra-tropics and the tropics. Note that variations are considerably smoothed by the monthly averaging. A strong increase in skill at the transition from the northern hemisphere summer to winter seasons is noticeable in the extra-tropical score mainly because of the larger land area in the northern extra-tropics, which is represented by ~3200 stations, compared to ~150 stations in the southern extra-tropics. The change itself represents the fact that it is relatively difficult to predict precipitation totals in convective situations which, over most land areas, are more common in summer. The magnitude of the seasonal variation in the extra-tropical score is on the order of 0.2, about four times as large as the differences between models, which are mostly smaller than 0.05. The highest tropical 1-SEEPS values of about 0.3 are similar to the lowest 1-SEEPS values in the extra-tropics.

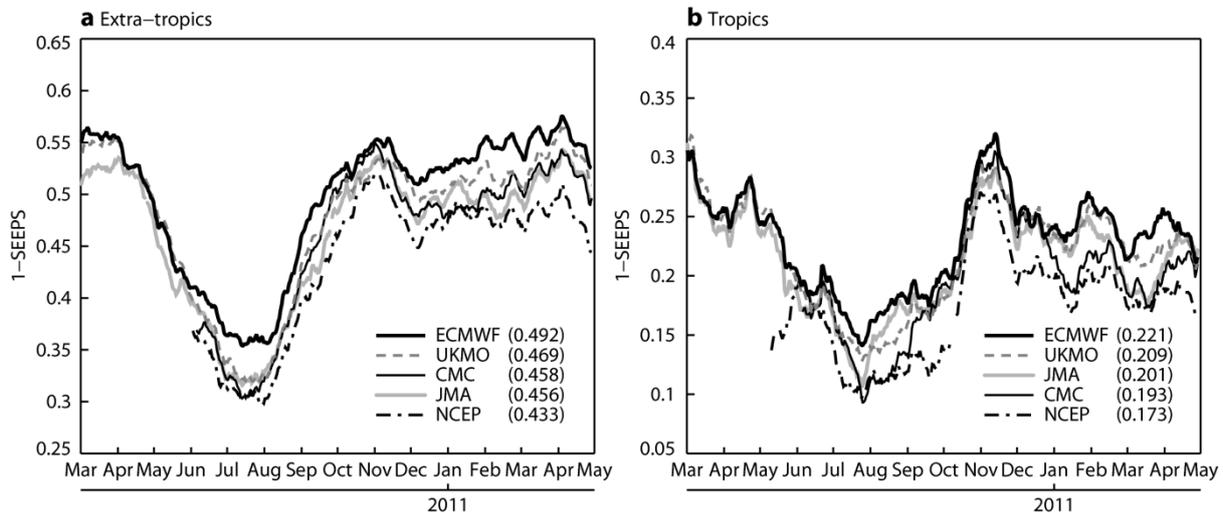


Figure 3: Time evolution of the running monthly mean of the 1-SEEPS score in the extra-tropics (a) and the tropics (b) on forecast day 3 for the CMC, JMA, NCEP, UKMO, and ECMWF models. CMC and NCEP data was available only from 1 June 2010 onwards. Numbers in parentheses give mean values over the period 1 Jun 2010 – 30 Apr 2011.

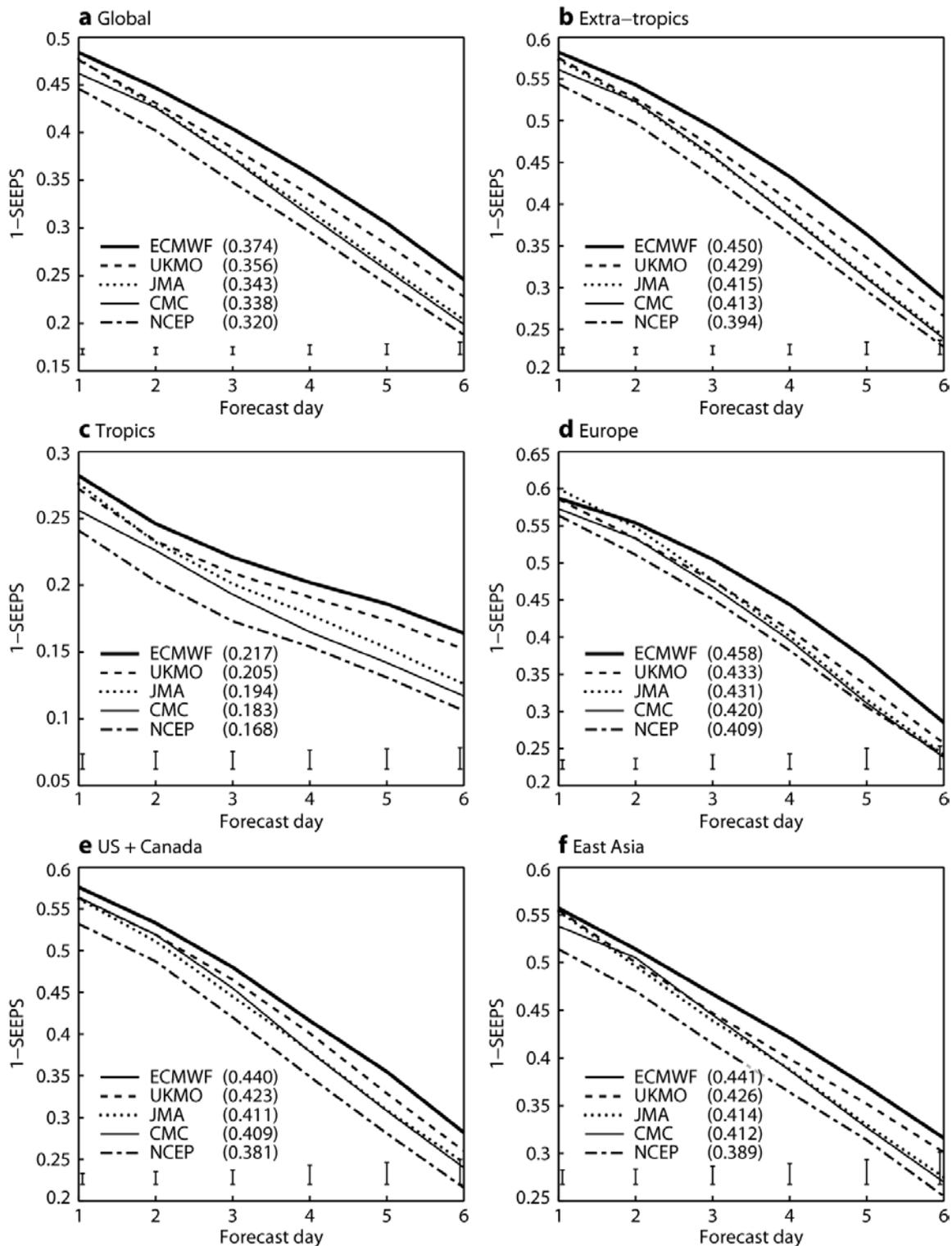


Figure 4: Values of 1-SEEPS for the CMC, JMA, NCEP, UKMO, and ECMWF models in different regions, averaged over the period 1 June 2010 – 30 April 2011, as a function of lead time. Mean values are given in parentheses. Error bars show width of 95% confidence intervals for model differences, derived from resampling of daily scores (see text for details).

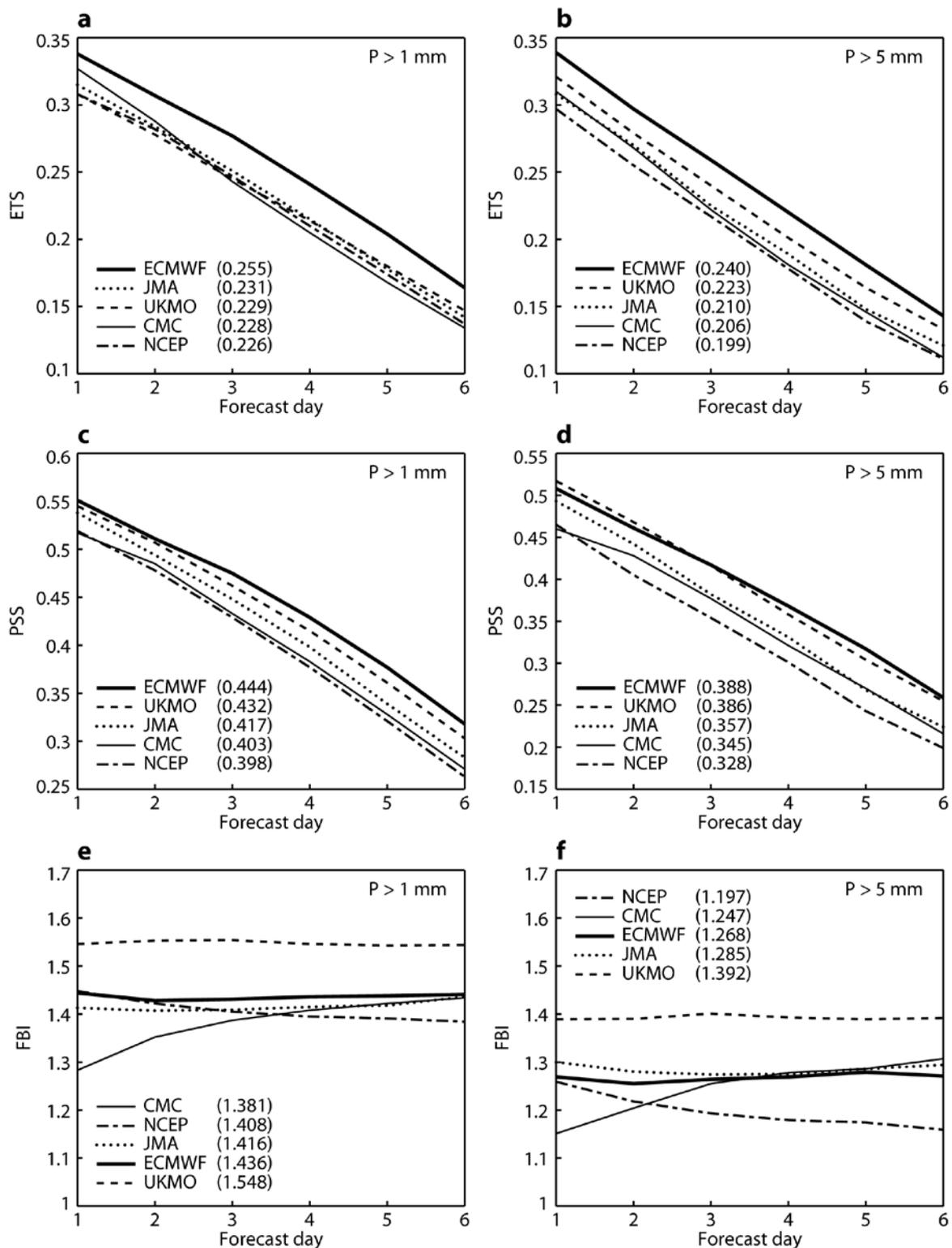


Figure 5: Global ETS (a), (b), PSS (c), (d), and FBI values (e), (f) for the CMC, JMA, NCEP, UKMO, and ECMWF models, averaged over the period 1 June 2010 – 30 April 2011, as a function of lead time. Thresholds are 1 mm (left panels) and 5 mm (right panels). Mean values are given in parentheses.

Figure 4 shows, as a function of forecast lead time, areal averages of the 1-SEEPS score, temporally averaged over the period where data from all models are available (1 June 2010 to 30 April 2011). Confidence intervals for model differences were computed by resampling daily 1-SEEPS differences between models, creating a large number ( $M=1000$ ) of synthetic time series for each model pair and region. The method, also known as bootstrapping, does not require assumptions about the shape of the error distribution (Wilks 1997). Because of the seasonal variation in the data (not just in the original score but also in the model differences), a de-trending using a 2-month running mean was applied prior to the resampling. The variance around this mean showed little variation on the seasonal time-scale. From the resulting distribution of the  $M$  period-means, 2.5% and 97.5% percentiles were computed, and the difference between them, averaged over all model pairs, was included in Figure 4. If two curves are separated by more than half the length of the confidence interval, they are significantly different at the 5% significance level (Lanzante 2005). The block-length used in the resampling was 10 days, but the results were insensitive to the exact choice of this value, as there was virtually no serial correlation in the de-trended, inter-model 1-SEEPS differences beyond one day (Hamill 1999).

The decrease of skill with forecast lead time is approximately linear within the range of lead times considered here. There is a very large difference in skill between the tropics and extra-tropics due to the generally much lower predictability of convective precipitation. The skill on forecast day 1 in the tropics is roughly equal to the skill of the extra-tropical forecast on day 5. The model ranking is similar in different areas, with the ECMWF taking the overall lead, and the UKMO generally ranking second. According to the confidence intervals derived from resampling, the ranking is rather robust globally, and in the extra-tropics, except between the CMC and JMA models which are very close. Results for Europe show comparatively low skill of the ECMWF model at forecast day 1. This is further investigated below (section c), by looking at a break-down of SEEPS into individual contributions from the contingency table.

### 3.2 Comparison with other scores

To compare the model ranking by the SEEPS score to the rankings deduced from more commonly used scores, the ETS and PSS have been computed as well. Figure 5 shows results for thresholds of 1 and 5 mm for the global domain. The overall model ranking is the same as in SEEPS. Interestingly, the PSS places the UKMO and ECMWF models much closer together than the ETS. This may partly be because the ETS penalizes overprediction more strongly than the PSS (Brill 2009).

The decrease with lead time of SEEPS skill is very similar to the decrease with lead time of ETS and PSS for a threshold of 1 mm. For example from forecast day 1 to 5 the skill given by all three scores decreases by about 50%. For a threshold of 5 mm the relative decrease of ETS and PSS is slightly larger (50-60%).

The relevance to forecast users of the differences in skill seen between models will depend on their individual cost-loss relationships and may not necessarily be measured by the scores used here. What we can infer from the slopes of the 1-SEEPS curves in Figure 4 and the ETS and PSS curves in Figure 5 is that a difference of 0.05 roughly corresponds to 1 forecast day. A difference of 0.02-0.03, typically seen between models, would therefore amount to 0.5 forecast days.

### 3.3 Break-down of the SEEPS score

Because of its linearity, the SEEPS score can be broken down into individual contributions from the six off-diagonal elements of the 3×3 contingency table. This provides some insight into the sources of differences in skill between the models. Note that for the discussion of break-down results it is more convenient to refer to the error score SEEPS itself, rather than the skill score 1-SEEPS. Figure 6 shows the results of the break-down for forecast day 3 for the extra-tropics and tropics. All models get their largest SEEPS error contribution from predicting the ‘light’ category, when either ‘dry’ or ‘heavy’ was observed (blue and orange in Figure 6). This is most pronounced in the tropics, and it is the main reason for the higher SEEPS values there compared to the extra-tropics. Note that the SEEPS scoring matrix (4) is quite asymmetric along the middle row. For stations with moderate-to-dry climatologies ( $p_1 > 0.5$ ), predicting ‘light’ when ‘heavy’ is observed, is penalized considerably more than predicting ‘light’ when ‘dry’ is observed. However, the latter occurs more frequently, and the overall contributions of these two errors in Figure 6 are roughly equal. This equivalence can partly be attributed to the equitability of SEEPS.

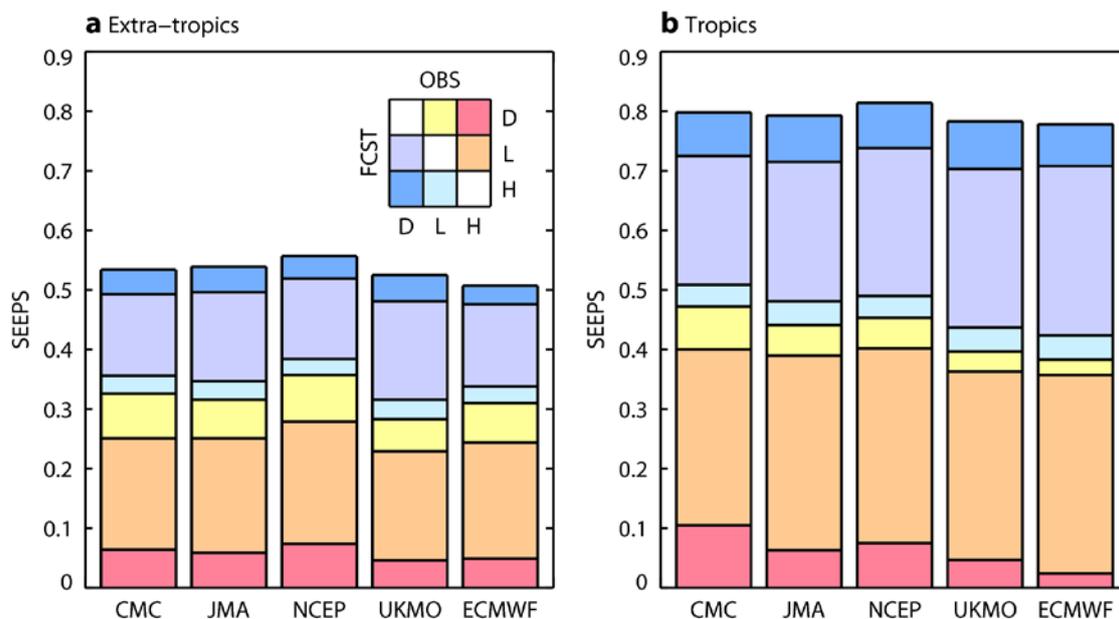


Figure 6: Break-down of SEEPS for forecast day 3 into individual contributions from the contingency table, for the extra-tropics (a), and the tropics (b). The insert shows the correspondence between color and position in the contingency table (D=dry, L=light, H=heavy). Warm colors indicate insufficient precipitation in the forecast, cold colors indicate too much precipitation.

Table 2 shows for three stations the fraction of days in the forecast, observations, and climatology that fall into each of the three SEEPS categories. The stations were selected to cover a wide range of ‘dry’ climatological probabilities  $p_1$ , from about 0.3 at the moist station (Connaught Airport) to 0.8-0.9 at the dry station (Albuquerque International Airport), and include both continental and maritime locations. In general, the forecast produces too many ‘light’ cases, mostly at the expense of the

number of ‘dry’ cases. This is most evident at the subtropical station Naha Airport. At the moist, mid-latitude station Connaught Airport the problem does not occur in the winter period. Overall, the results suggest that it is mostly the convective precipitation which produces the overprediction of the ‘light’ category in the forecast.

The too frequent prediction of light precipitation, as illustrated by the blue and orange areas in Figure 6, is a well-known issue in NWP. The feature-oriented verification study by Wittmann et al. (2010) suggests that the spatial structure of precipitation forecasts improves with increasing model resolution for grid spacings from 25 km down to 2.5 km, and with a change from parameterized to explicit deep convection. No clear signal of reduced overprediction of light precipitation with higher resolution is seen here, where all models employ deep convection parameterizations, and resolutions range from 16 to 35 km (Table 1).

*Table 2: Fraction of days falling into individual SEEPS categories at three airport stations with strongly differing climatological probabilities of ‘dry’ days (Connaught, Ireland; Albuquerque, USA; Naha, Japan) during a summer and a winter season.*

		JJA 2010			DJF 2010/11		
		Dry	Light	Heavy	Dry	Light	Heavy
<b>Connaught Airport (moist) 54°N, 9°W</b>	Forecast	0.16	0.61	0.23	0.43	0.41	0.16
	Observed	0.32	0.47	0.21	0.42	0.47	0.11
	Climate	0.33	0.45	0.22	0.28	0.48	0.24
<b>Albuquerque International Airport (dry) 35°N, 107°W</b>	Forecast	0.70	0.28	0.02	0.86	0.13	0.01
	Observed	0.86	0.04	0.10	0.94	0.05	0.01
	Climate	0.80	0.13	0.07	0.88	0.08	0.04
<b>Naha Airport (moderate) 26°N, 128°E</b>	Forecast	0.06	0.74	0.20	0.22	0.72	0.06
	Observed	0.41	0.36	0.23	0.58	0.30	0.12
	Climate	0.62	0.25	0.13	0.64	0.24	0.12

Also contributing to the higher SEEPS values in the tropics are events predicted as ‘heavy’ when no precipitation was observed (purple). Interestingly the opposite, that is, observed ‘heavy’ events which were completely missed (red), do not contribute more in the tropics than in the extra-tropics. One type of error contribution which is consistently smaller in the tropics is light observed precipitation when none was forecast.

Looking at differences between the models in the extra-tropics, it is evident that the higher total SEEPS of NCEP compared to ECMWF is mainly due to the larger number of underforecast events. UKMO has the smallest underforecasting contribution of all models (consistent with it having the largest FBIs of all models (Fig. 5)), but also the largest number of non-zero precipitation forecasts when ‘dry’ was observed. The over-forecasting of light precipitation in the UKMO model has been noted by Hamill (2011) in his examination of multi-model probabilistic forecasts using Thorpex Interactive Grand Global Ensemble (TIGGE) data, and is also an issue known to operational forecasters at the UKMO, which they regularly correct for. The CMC and JMA models have a higher SEEPS than ECMWF mainly because of a higher number of 2-category errors (‘dry’ observed vs

‘heavy’ forecast and vice versa). In the tropics the relative differences between models of the total SEEPS values are generally smaller, but individual contributions differ considerably. The CMC and ECMWF models, for example, arrive at a similar total skill with very different characteristics regarding under- and over-forecasting of events.

The lower skill of the ECMWF model compared to the other models in Europe at forecast day 1 is mainly due to a larger number of ‘light’ forecasts when ‘heavy’ was observed (not shown). The JMA model ranks first in this case primarily because it has a smaller number of ‘light’ forecasts than the other models in cases when no precipitation was observed.

## 4 Sensitivity analysis

In this section we assess the robustness of the results with respect to choices made in the computation of the SEEPS score; specifically the exclusion of arid climates, the areal density weighting, and the direct verification on the different grids of the models. The issue of observation representativeness is also addressed, using a perfect model approach.

### 4.1 Exclusion of arid climates

In arid climates, where the probability of the ‘dry’ category approaches 1, the error matrix becomes very asymmetric, strongly penalizing the under-forecasting of precipitation. The value of 0.85 as an upper limit to the ‘dry’ probability was suggested by Rodwell et al. (2010) to reduce uncertainty in area-mean scores. However, it also leads to the exclusion of stations particularly in areas where station density is already low. Similar difficulties were noted by Hamill (2011) in the verification of probabilistic quantitative precipitation forecasts. In order to assess the effect on scores and model rankings, the calculation of SEEPS is repeated for limiting dry probabilities of 0.90 and 0.95. For a dry probability of 0.90 the number of available stations, averaged over the verification period, increases from about 2100 to 2240 in the extra-tropics, and from about 600 to 650 in the tropics. For a dry probability of 0.95 the numbers increase to 2380 in the extra-tropics and 720 in the tropics, respectively. The latter is a considerable relative increase of 20%.

Figure 7a shows that changing the value of the probability limit does change the scores by about 10% in the tropics, and by 2-3% in the extra-tropics. The ranking of the models remains more or less unchanged. In the extra-tropics the UKMO score appears to benefit slightly from including a larger number of ‘dry’ stations, unlike the other models. This result differs from the conclusion of Hamill (2011), mainly because in a climatologically dry region SEEPS penalizes an individual dry forecast when wet was observed more strongly than an individual wet forecast when dry was observed (Rodwell et al. 2010). The probabilistic norms used by Hamill (2011) do not have this difference in weighting, and the UKMO model is more heavily penalized for the overforecasting of precipitation in dry regions.

It is interesting that the skill in the tropics increases when more stations from dry climates are included. Analysis of individual SEEPS contributions reveals that this is because the increased number of complete misses (‘dry’ forecast when ‘heavy’ was observed) is more than compensated for by a

reduced number of ‘light’ forecasts when ‘dry’ was observed. In the extra-tropics, meanwhile, the former of the two effects dominates (except in the UKMO model), leading to a slight worsening of the SEEPS score when more arid stations are included.

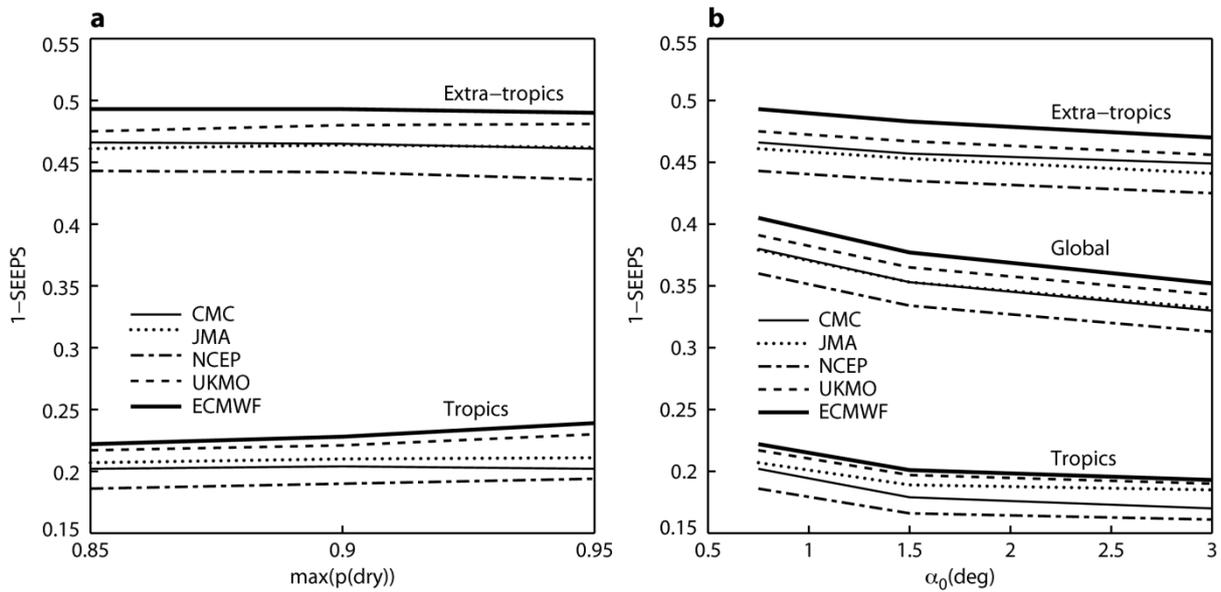


Figure 7: Dependence of 1-SEEPS for forecast day 3 (a) on the upper limit of the probability of ‘dry’ days used in the exclusion of arid climates, and (b) on the length scale used in the density weighting.

## 4.2 Station density weighting

The weighting which is used in the areal averaging to partially compensate for inhomogeneous station density, has one adjustable parameter  $\alpha_0$ . It is a reference angle with a standard value of  $0.75^\circ$  (83 km). An increase of this parameter increases the weight of data-sparse regions in global or continental-scale averages. When  $\alpha_0$  is doubled and quadrupled from its reference value to  $1.5^\circ$  and  $3.0^\circ$ , the weight of Europe in the global average score decreases from 19% to 13%, and 10%. The last value is close to the actual land surface fraction taken up by Europe which is about 8% (not counting Antarctica). The weight of Africa increases from 5% to 6% and 9%, which is still less than half of its actual land surface fraction (20%).

Global 1-SEEPS scores for forecast day 3 drop by about 0.05 ( $\approx 12\%$ ) from  $\alpha_0 = 0.75^\circ$  to  $\alpha_0 = 3.0^\circ$  (Fig. 7b). Scores for the extra-tropics and tropics decrease by about 0.02-0.03. The global score decreases more strongly than the extra-tropical and tropical ones because a higher value of  $\alpha_0$  increases the weight of the data-sparse, lower-skill, tropical areas in the global score. The model ranking is not affected, as the drop of 1-SEEPS values of different models is nearly parallel.

### 4.3 Grid spacing

For assessing the sensitivity of the results to grid spacing we aggregate the forecasts to the same reference grid to see how this affects the SEEPS score. Up-scaling model results to a coarser grid is not attempting to run the model on a coarser grid, but it is done to give some indication of the sensitivity of the verification results to grid spacing. Since the NCEP output grid is the one with the largest spacing ( $0.5^\circ$ ) it is used as the common reference. We aggregate the CMC, JMA, UKMO, and ECMWF precipitation forecasts to the NCEP grid using arithmetic averaging. The precipitation value  $P_{ij}$  at point  $(i,j)$  on the coarse grid is calculated from

$$P_{ij} = \frac{1}{(m_2 - m_1 + 1)(n_2 - n_1 + 1)} \sum_{m=m_1}^{m_2} \sum_{n=n_1}^{n_2} p_{mn} \quad (7)$$

where  $p_{mn}$  is the precipitation at point  $(m,n)$  on the fine grid, and the summation extends over all points within a coarse-grid gridbox centred around point  $(i,j)$ . Arithmetic averaging has been chosen as the aggregation method because it conserves the total amount of precipitation. For the NCEP forecast no aggregation is performed since (7) in this case simply reproduces the original data.

The effect of the interpolation is summarized in Figure 8. The skill of the ECMWF model is slightly increased by the averaging, whereas for the JMA and UKMO models there is little change, for CMC it is slightly negative. Since the ECMWF original grid is the one with the highest resolution, this could suggest that the double penalty effect [getting bad scores at two times (locations) when the forecast is

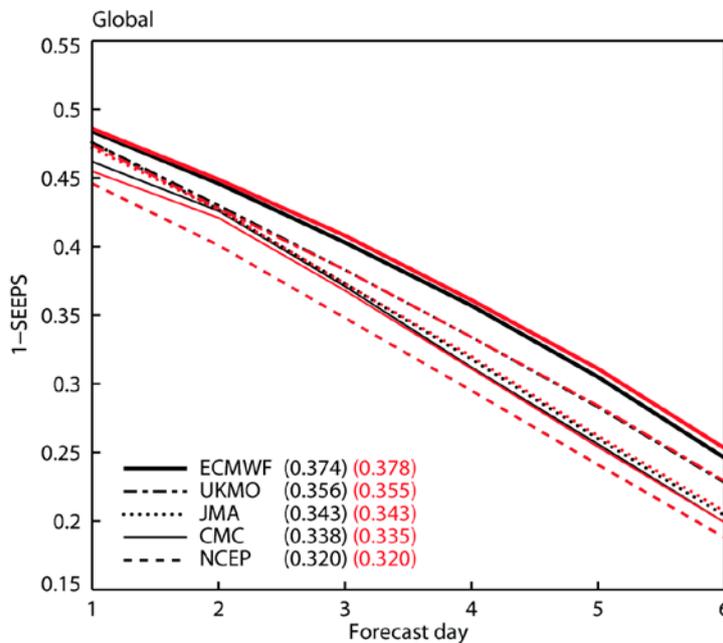


Figure 8: 1-SEEPS of the global domain as a function of lead time for verification on the original output grids (see Table 1) (black lines), and after arithmetic averaging to the  $0.5^\circ$  NCEP grid (red lines). Numbers in parentheses give values of 1-SEEPS averaged over all lead times on each model's original output grid (black), and on the  $0.5^\circ$  grid (red).

phase-shifted in time (space)] starts to be relevant in global models as well. So far it has mainly been an issue in precipitation forecasts of limited area models. Analysis of the individual contributions to SEEPS (as in Figure 6) reveals that the averaging has a similar effect on all five forecasts, and that it primarily affects the ‘dry’ vs ‘non-dry’ distinction. As expected from a smoother field, we find a decrease in the number of cases where the models erroneously predict ‘dry’, but at the same time there is an increased number of observed ‘dry’ cases which the models fail to predict as such. The change of SEEPS seen in Figure 8 is the net result of these larger opposing changes, so its sign is sensitive to the characteristics of the precipitation forecasts, particularly the frequency bias. The ECMWF model benefits most from the averaging because the number of missed precipitation cases decreases more strongly than in the other models.

#### 4.4 Observation representativeness

Even for a perfect model and perfectly accurate observations, verification against surface stations would show a non-zero error due to the fact that a quantity at grid-scale is compared to point observations of that quantity, representing a different, and varying, spatial scale. Göber et al. (2008) investigated the problem by constructing ‘perfect’ forecasts in the form of grid-scale averages of high-density observations. Verification of these forecasts showed the maximum point forecast skill which can be reached by a model at a given resolution. The value of this attainable score will vary with the quantity considered and its heterogeneity, so it is different in different geographic areas. For Europe, Rodwell et al. (2010) used the same approach to estimate the minimum error for a ‘perfect’ forecast as a function of grid resolution. For a grid spacing of 25 km this value was found to be  $\sim 0.2$  (i.e.  $1 - \text{SEEPS} = 0.8$ ). Here, a similar approach is taken, and applied to the larger range of scores considered.

Gridded fields created by averaging over a limited number of rain gauges per grid box can never be perfect, since they are only approximations to the true grid-box mean. However, they can be considered ‘quasi-perfect’ in the sense that they produce the smallest possible mean absolute error between grid box and rain gauge values for the given observations, grid spacing, and spatial matching method (nearest gridpoint). Hence we prefer to use the term ‘quasi-perfect’ for this method.

Quasi-perfect forecasts were provided by the gridded 24-hour precipitation analyses derived from high-resolution raingauge data available for the European area (Ghelli and Lalaurette 2000). The gridding method is grid-box averaging. The grid spacing is that of the ECMWF T799 (25 km) reduced Gaussian grid (N400), roughly corresponding to the average of grid spacings used by the different models (Table 1). Results presented below are based on data from one year (2007).

Table 3 shows scores as a function of  $n_{\min}$ , the minimum number of observations required within a grid-box in order for a perfect forecast value to be calculated. As  $n_{\min}$  increases, the model becomes more ‘perfect’ but the corresponding decrease in  $N$  (the number of stations in the domain at which a quasi-perfect forecast is available) leads to increased uncertainty in the mean score. Table 3 demonstrates that the scores are not very sensitive to the precise value of  $n_{\min}$ . In agreement with Rodwell et al. (2010),  $1 - \text{SEEPS}$  values of around 0.8 are obtained for the quasi-perfect forecast. Hence, observation representativeness accounts for a little less than half of the difference between current extra-tropical  $1 - \text{SEEPS}$  on forecast day 1 (0.5-0.6; see Fig. 3 and Fig. 4) and the theoretical maximum value of 1. The other half of the skill deficit can be ascribed to actual forecast error and

shows the potential for model improvement at the current resolution. Classical dichotomous verification measures and scores (for a threshold of 1 mm/day) have also been computed. According to these results a frequency bias of about 5% is produced by the grid-point vs. raingauge comparison. This is small compared to the frequency bias seen in real forecasts (Fig. 5) and leaves plenty of room for further reductions in FBI, even at the current model resolution.

*Table 3: SEEPS and other scores (threshold 1 mm day<sup>-1</sup>) for a quasi-perfect forecast on a grid with 25 km grid spacing in Europe for the year 2007.*

$n_{\min}$  = Minimum number of observations in a grid-box required for creating a quasi-perfect forecast  
 $n$  = Average number of observations in a grid-box from which quasi-perfect forecast is created  
 $N$  = Number of stations in the whole domain for which a quasi-perfect forecast is available  
 BR = Base rate (climatological frequency of occurrence)  
 HR = Hit rate (probability of detection)  
 FBI = Frequency bias index  
 ETS = Equitable threat score  
 PSS = Peirce skill score

$n_{\min}$	$n$	$N$	1-SEEPS	BR	HR	FBI	ETS	PSS
1	3.82	863	0.826	0.328	0.929	1.040	0.763	0.875
2	4.63	670	0.811	0.332	0.926	1.052	0.742	0.863
4	6.23	390	0.800	0.331	0.925	1.061	0.731	0.858
6	7.90	198	0.796	0.333	0.922	1.059	0.725	0.853
8	10.22	79	0.781	0.344	0.910	1.049	0.705	0.837

## 5 Conclusions

The aim of this study has been to provide a direct comparison of the precipitation forecast skill of five operational global NWP models in 2010/11. Use was made of the new 3-category SEEPS score, which measures skill in probability space, thus allowing the aggregation of climatologically diverse regions. The more widely used dichotomous scores ETS and PSS have been calculated as well. The main findings can be summarized as follows:

- For lead times between D+1 and D+6 in the extra-tropics, differences in 24-h precipitation forecast skill of several global models are on the order of 1 forecast day. This applies to the new SEEPS score, as well as to ETS and PSS.
- In all models, differences between the tropics and extra-tropics are large. SEEPS scores at D+1 in the tropics are comparable to those at D+6 in the extra-tropics.

- There is a marked decrease of extra-tropical scores during the northern hemisphere summer such that, for example, winter skill at D+6 is similar to summer skill at D+3.
- All models get their largest SEEPS error contribution from predicting the ‘light’ category, when either ‘dry’ or ‘heavy’ was observed, highlighting the two distinct problems of (a) over-prediction of drizzle and (b) the difficulty to predict heavy precipitation.
- Choices of parameter values in the verification (like the threshold for exclusion of dry climates) affect the absolute values of SEEPS but not the overall model ranking.
- Just under one-half of the current forecast error at day 1 in the extra-tropics can be attributed to the fact that grid-box values are verified against point observations.
- Upscaling of the forecasts to a common grid, using averaging, has little effect on the results and does not change the overall model ranking.

The main limitation of global precipitation verification against surface stations is their uneven geographical distribution. Some of the gaps could be filled by more strongly standardized reporting procedures. On a more fundamental level, land areas represent only about 30% of the Earth’s surface. This limits the validity of the verification in terms of monitoring the geophysical accuracy of a forecast system. However, the results are clearly highly relevant for the human population. By providing estimates of uncertainty based on re-sampling, and by performing sensitivity studies, we have attempted to increase confidence in the validity of the results.

## Acknowledgements

We would like to thank Tom Hamill (NOAA) and an anonymous reviewer for their insightful and constructive comments which helped to improve the manuscript. The help of Anabel Bowen (ECMWF) in improving the graphics is appreciated.

## References

- Abel, S. J., D. N. Walters, and G. Allen, 2010: Evaluation of stratocumulus cloud prediction in the Met Office forecast model during VOCALS-Rex. *Atmos. Chem. Phys.*, **10**, 10541–10559.
- Bechtold, P., M. Köhler, T. Jung, F. Doblas-Reyes, M. Leutbecher, M. J. Rodwell, F. Vitart, and G. Balsamo, 2008: Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. *Q. J. R. Meteorol. Soc.*, **134**, 1337-1351.
- Belair, S., M. Roch, A.-M. Leduc, P. A. Vaillancourt, S. Laroche, and J. Mailhot, 2009: Medium-range quantitative precipitation forecasts from Canada’s new 33-km deterministic global operational system. *Wea. Forecasting*, **24**, 690-708.

- Brill, K. F., 2009: A general analytic method for assessing sensitivity to bias of performance measures for dichotomous forecasts. *Wea. Forecasting*, **24**, 307-318.
- Cherubini, T., A. Ghelli, and F. Lalaurette, 2002: Verification of precipitation forecasts over the Alpine region using a high-density observing network. *Wea. Forecasting*, **17**, 238-249.
- Csima, G., and A. Ghelli, 2008: On the use of the intensity-scale verification technique to assess operational precipitation forecasts. *Meteorol. Appl.*, **15**, 145-154.
- Ebert, E. E., U. Damrath, W. Wergen, and M. E. Baldwin, 2003: The WGNE assessment of short-term quantitative precipitation forecasts. *Bull. Amer. Meteor. Soc.*, **84**, 481-492.
- Forbes, R. M., A. M. Tompkins, and A. Untch, 2011: A new prognostic bulk microphysics scheme for the IFS. *ECMWF Tech. Memo.*, 649, 28 pp.
- Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361-370.
- Gerrity, J. P., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.*, **120**, 2709-2712.
- Ghelli, A., and F. Lalaurette, 2000: Verifying precipitation forecasts using upscaled observations. *ECMWF Newsletter 87*, 9-17.
- Göber, M., E. Zsoster, and D. S. Richardson, 2008: Could a perfect model ever satisfy a naïve forecaster? On grid box mean versus point verification. *Meteorol. Appl.*, **15**, 359-365.
- Gregory, S. D. and P. R. Rowntree, 1990: A massflux convection scheme with representation of cloud ensemble characteristics and stability dependent closure. *Mon. Wea. Rev.*, **118**, 1483-1506.
- Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.
- Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Q. J. R. Meteorol. Soc.*, **132**, 2905-2923.
- Hamill, T. M., 2011: Verification of TIGGE multi-model and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the conterminous US. *Mon. Wea. Rev.*, **139** (accepted pending minor revisions).
- Kain, J. S., and J. M. Fritsch, 1990: A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.*, **47**, 2784-2802.
- Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, R. M. Errico, and R. Yang, 2009: Improving incremental balance in the GSI 3DVAR analysis system. *Mon. Wea. Rev.*, **137**, 1046-1060.

- Lanzante, J. R., 2005: A cautionary note on the use of error bars. *J. Climate*, **18**, 3699-3703.
- Pan, H.-L. and W.-S. Wu, 1995: Implementing a mass flux convective parameterization package for the NMC Medium-Range Forecast Model. NMC Office Note 409, 40 pp. [Available from NOAA/NWS/NCEP, Environmental Modeling Center, WWB, Room 207, Washington, DC 20233.]
- Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453-454.
- Richardson, D. S., J. Bidlot, L. Ferranti, A. Ghelli, C. Gibert, T. Hewson, M. Janousek, F. Prates, and F. Vitart, 2009: Verification statistics and evaluations of ECMWF forecasts in 2008-2009. *ECMWF Technical Memorandum*, No. 606, ECMWF, Reading, United Kingdom, 45pp.
- Rodwell, M. J., D. S. Richardson, T. D. Hewson, and T. Haiden, 2010: A new equitable score suitable for verifying precipitation in numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **136**, 1344-1363.
- Shapiro, M., and co-authors, 2010: An Earth-system prediction initiative for the twenty-first century. *Bull. Amer. Meteor. Soc.*, **91**, 1377-1388.
- Simmons, A. J., and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **128**, 647-677.
- Sundqvist, H., E. Berge, and J. E. Kristjansson, 1989: Condensation and cloud parameterization studies with a mesoscale numerical weather prediction model. *Mon. Wea. Rev.*, **117**, 1641-1657.
- Wilks, D. S., 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate*, **10**, 65-82.
- Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 59, Academic Press, 629 pp.
- Wilson, D. R. and S. P. Ballard, 1999: A microphysically based precipitation scheme for the UK Meteorological Office Unified Model. *Q. J. R. Meteorol. Soc.*, **125**, 1607-1636.
- Wittmann, C., T. Haiden, and A. Kann, 2010: Evaluating multi-scale precipitation forecasts using high-resolution analysis. *Adv. Sci. Res.*, **4**, 89-98.
- WMO, 2008: Recommendations for the verification and intercomparison of QPFs and PQPFs from operational NWP models. WWRP 2009-1, Revision 2, WMO TD No. 1485, 37pp.
- Zhao, Q., and F. H. Carr, 1997: A prognostic cloud scheme for operational NWP models. *Mon. Wea. Rev.*, **125**, 1931-1953.