Intercomparison of Global Model Precipitation Forecast Skill in 2010/11 Using the SEEPS Score

THOMAS HAIDEN, MARK J. RODWELL, AND DAVID S. RICHARDSON
European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

AKIRA OKAGAKI
Japan Meteorological Agency, Tokyo, Japan

TOM ROBINSON
Canadian Meteorological Centre, Environment Canada, Dorval, Quebec, Canada

TIM HEWSON
Met Office, Exeter, United Kingdom

(Manuscript received 1 November 2011, in final form 19 January 2012)

ABSTRACT
Precipitation forecasts from five global numerical weather prediction (NWP) models are verified against rain gauge 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 extratropics are quite pronounced. SEEPS scores at forecast day 1 in the tropics are similar to those at day 6 in the extratropics. 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 extratropics can be attributed to the fact that gridbox 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 intermodel 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 because of the lack of reliable analyses and the uneven and often sparse distribution of rain gauges. 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. Because of the large number of possibilities with regard to the choice of score, interpolation method, spatiotemporal...
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 centers 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 because of differences in verification methodology and rainfall observation time. To address this problem, the WMO issued recommendations for the verification and intercomparison of QPFs from operational NWP models (WWRP/WGNE Joint Working Group on Verification 2008).

There is considerable practical and scientific interest in more detailed information about variations in precipitation forecast skill between, for example, the tropics and extratropics, or between different seasons. Such comparisons are not straightforward because of geographical variations in precipitation climate. For a given fixed precipitation threshold that is used to create a contingency table, the base rate (BR; 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 the precipitation forecast skill of current operational global NWP models. To check the robustness of the results, a resampling technique is applied, and sensitivity analyses are performed.

2. Data and methodology

a. Forecast data

Operational forecasts of 24-h precipitation totals from Canadian Meteorological Centre (CMC), Japan Meteorological Agency (JMA), National Centers for Environmental Prediction (NCEP), Met Office (UKMO), and European Centre for Medium-Range Weather Forecasts (ECMWF) global models (Table 1) are examined. Lead times from 24 to 144 h (i.e., from day 1 to day 6: from $D + 1$ 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 of 48 and 72 h. The common verification period is 1 June 2010–30 April 2011 (11 months). Only 1200 UTC runs are considered. Because of 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° to 0.50°. Table 1 shows the corresponding resolutions of the actual model grids. Nearest-neighbor matching between rain gauges and grid points is performed on each output grid, without additional interpolation steps. It may be argued that this reduces the 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 were regridded prior to receipt (top two rows in 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 July 2010, $0.205^\circ$ afterward) 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.

b. Observations

The observational data used in this study consist of 24-h precipitation totals ending at 1200 UTC, derived from 6-hourly and 12-hourly synoptic observations (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 extratropics, 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 nonstandard 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 nonstandard reporting practice that 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 ±100 stations.

c. 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 that 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 index (FBI), although not a skill score in itself, is useful for understanding differences in behavior of the different scores. These three dichotomous measures are defined by

\[
\text{ETS} = \frac{a - a_{\text{REF}}}{a - a_{\text{REF}} + b + c} \quad \text{with} \quad a_{\text{REF}} = \frac{(a + b)(a + c)}{a + b + c + d},
\]

(1)

\[
\text{PSS} = \frac{ad - bc}{(a + c)(b + d)}, \quad \text{and}
\]

(2)

\[
\text{FBI} = \frac{a + b}{a + c},
\]

(3)

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

The PSS is the only two-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 two-category score is a modified PSS. First, the PSS is often written as the hit rate minus the false alarm rate, but it can also be written as 1 - (miss rate + 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. Second, the miss rate and false alarm rate are calculated using the climatological base rate over the fixed 30-yr period of 1980–2009 derived from all SYNOPs discussed in section 2b. The result is that SEEPS permits the construction of daily score time series 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 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 extratropics, the resulting threshold between the light and heavy categories (\(t_{L/H}\) in Fig. 2) is on 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.

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 emphasize 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:

\[
\mathbf{S} = \frac{1}{2} \begin{pmatrix}
0 & 1 & 4 \\
\frac{1}{p_1} & 0 & \frac{4}{1 - p_1} \\
\frac{1}{p_1} + \frac{3}{2 + p_1} & \frac{3}{2 + p_1} & 0
\end{pmatrix}.
\]

(4)
In this matrix the observed categories of 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 $1/2$ SEEPS is plotted.

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

d. Areal averaging and aggregation

SEEPS is a direct measure of error and can be averaged over a set of stations. Arithmetic averaging, however, would overemphasize 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:

$$\rho_k = \sum_l e^{-(a_{kl}/a_0)^2},$$

where the summation extends over all stations available on a given day, $a_{kl}$ is the angle subtended at the center of the earth between stations $k$ and $l$, and $a_0$ is a reference angle. The weighted area-mean SEEPS score $S$ is computed from SEEPS scores $S_k$ at individual stations by

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

A standard value of $a_0 = 0.75^\circ$ is used in this study. The sensitivity of area averages of SEEPS with respect to this parameter is analyzed in section 4b.

3. Verification results

a. 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 extratropics 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 extratropical score mainly because of the larger land area in the northern extratropics, which is represented by $\sim 3200$ stations, compared to $\sim 150$ stations in the southern extratropics. The change itself represents the fact that it is relatively difficult to predict precipitation totals in convective situations that, over most land areas, are more common in summer. The magnitude of the seasonal variation in the extratropical score is on the order of 0.2, about 4 times as large as the differences between models, which are mostly smaller than 0.05.
The highest tropical \(1 - \text{SEEPS}\) values of about 0.3 are similar to the lowest \(1 - \text{SEEPS}\) values in the extratropics.

Figure 4 shows, as a function of forecast lead time, areal averages of the \(1 - \text{SEEPS}\) score, temporally averaged over the period where data from all models are available (from 1 June 2010 to 30 April 2011). Confidence intervals for model differences were computed by resampling daily \(1 - \text{SEEPS}\) 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 detrending procedure 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, the 2.5th and 97.5th percentiles were computed, and the difference between them, averaged over all model pairs, was included in Fig. 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 detrended, intermodel \(1 - \text{SEEPS}\) differences beyond 1 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 extratropics 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 extratropical forecast on day 5. The model ranking is similar in different areas, with the ECMWF model taking the overall lead and that from UKMO generally ranking second. According to the confidence intervals derived from resampling, the ranking is rather robust globally and in the extratropics except between the CMC and JMA models, which are very close. Results for Europe show a comparatively low skill of the ECMWF model at forecast day 1. This is further investigated below (section 3c), by looking at a breakdown of SEEPS into individual contributions from the contingency table.

b. 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 over-prediction 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 \(1 - \text{SEEPS}\) curves in Fig. 4 and the ETS and PSS curves in Fig. 5 is that a difference of 0.05 roughly corresponds to one forecast day. A difference of 0.02–0.03, typically seen between models, would therefore amount to 0.5 forecast days.

c. Breakdown 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 \(3 \times 3\) contingency table. This provides some insight into the source of differences in skill between the models. Note that for the discussion of breakdown results it is more convenient to refer to the error score, SEEPS itself, rather than the skill score, \(1 - \text{SEEPS}\). Figure 6 shows the results of the breakdown for forecast day 3 for the extratropics 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 areas in Fig. 6). This is most pronounced in the tropics, and it is the main reason for the higher SEEPS values there compared to the extratropics. 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 precipitation when heavy is observed is penalized considerably more than predicting light precipitation when dry is observed. However, the latter occurs more frequently, and the overall contributions of these two errors in Fig. 6 are roughly equal. This equivalence can partly be attributed to the equitability of SEEPS.

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
FIG. 4. Values of $1 - \text{SEEPS}$ for the CMC, JMA, NCEP, UKMO, and ECMWF models in different regions, averaged over the period 1 Jun 2010–30 Apr 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).
produces too many light cases, mostly at the expense of the number of dry cases. This is most evident at the subtropical station of Naha Airport. At the moist, midlatitude station of Connaught Airport, the problem does not occur in the winter period. Overall, the results suggest that it is mostly the convective precipitation that produces the overprediction of the light category in the forecast.
The too frequent prediction of light precipitation, as illustrated by the purple and orange areas in Fig. 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 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).

Also contributing to the higher SEEPS values in the tropics are events predicted as heavy when no precipitation was observed (dark blue areas in Fig. 6). Interestingly, the opposite [i.e., observed heavy events that were completely missed (red)] does not contribute more in the tropics than in the extratropics. One type of error contribution that is consistently smaller in the tropics is light observed precipitation when none was forecast (yellow).

Looking at differences between the models in the extratropics, it is evident that the higher total SEEPS of NCEP compared to ECMWF is mainly due to the larger number of underforecast events. The UKMO model 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 nonzero precipitation forecasts when dry was observed. The overforecasting of light precipitation in the UKMO model has been noted...
by Hamill (2012) in his examination of multimodel probabilistic forecasts using the Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) data, and is also an issue known to operational forecasters at UKMO, which they regularly correct for. The CMC and JMA models have a higher SEEPS value than ECMWF mainly because of a higher number of two-category errors (dry observed versus heavy forecast and vice versa). In the tropics, the relative differences in the total SEEPS values between models 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 overforecasting 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 precipitation 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.

a. 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 underforecasting 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 (2012) in the verification of probabilistic quantitative precipitation forecasts. To assess the effect on scores and model rankings, the calculation of SEEPS is repeated for when 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 extratropics, and from about 600 to 650 in the tropics. For a dry probability of 0.95 the numbers increase to 2380 in the extratropics 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 extratropics. The ranking of the models remains more or less unchanged. In the extratropics, 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 (2012), 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 (2012) do not have this difference in weighting, and the UKMO model is more heavily penalized for the over-forecasting of precipitation in dry regions.

It is interesting that the skill in the tropics increases when more stations from dry climates are included. An analysis of individual SEEPS contributions reveals that this is because the increased number of complete misses (dry forecast when heavy precipitation was observed) is more than compensated for by a reduced number of light forecasts when dry was observed. In the extratropics, 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.

b. Station density weighting

The weighting that 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° (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° and 3.0°, the weight of Europe in the global average score decreases from 19% to 13% and 10%, respectively. 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%, respectively, 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 (~12%) from \( \alpha_0 = 0.75° \) to \( \alpha_0 = 3.0° \) (Fig. 7b). Scores for the extratropics and tropics decrease by about 0.02–0.03. The global score decreases more strongly than the extratropical 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 in different models is nearly parallel.

c. 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. Upscaling 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°), 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

\[
P_{ij} = \frac{1}{(m_2 - m_1 + 1)(n_2 - n_1 + 1)} \sum_{m=m_1}^{m_2} \sum_{n=n_1}^{n_2} P_{mn},
\]

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 box centered 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 Fig. 8. The skill of the ECMWF model is slightly increased by the averaging, whereas for the JMA and UKMO models there is little change, and 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 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 Fig. 6) reveals that the averaging has a similar effect on all five forecasts, and that it primarily affects the dry versus “nondry” distinction. As expected from a smoother field, we find a decrease in the number

---

**Fig. 8.** For the global domain, 1 – SEEPS 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° 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° grid (red).
of cases where the models erroneously predict dry, but at the same time there is an increased number of observed dry cases that the models fail to predict as such. The change of SEEPS seen in Fig. 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.

d. Observation representativeness

Even for a perfect model and perfectly accurate observations, verification against surface stations would show a nonzero error because of 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 gridscale averages of high-density observations. Verification of these forecasts showed the maximum point forecast skill that 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 \( \approx 0.2 \) (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 gridbox mean. However, they can be considered “quasi perfect” in the sense that they produce the smallest possible mean absolute error between gridbox and rain gauge values for the given observations, grid spacing, and spatial matching method (nearest grid point). Hence we prefer to use the term quasi perfect for this method.

Quasi-perfect forecasts were provided by the gridded 24-h precipitation analyses derived from high-resolution rain gauge data available for the European area (Ghelli and Lalaurette 2000). The gridding method is gridbox averaging. The grid spacing is that of the ECMWF T799 (25 km) reduced Gaussian grid (N400), roughly corresponding to the average of the 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 \( n_{\text{min}} \), the minimum number of observations required within a grid box for a perfect forecast value to be calculated. As \( n_{\text{min}} \) 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 \( n_{\text{min}} \). 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 extratropical \( 1 - \text{SEEPS} \) values on forecast day 1 (0.5–0.6; see Figs. 3, 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\(^{-1}\)) have also been computed. According to these results a frequency bias of about 5% is produced by the gridpoint versus rain gauge 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.

5. Conclusions

The aim of this study has been to provide a direct comparison of the precipitation forecast skills of five operational global NWP models in 2010/11. Use was made of the new three-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 extratropics, differences in the 24-h precipitation forecast skills of several global models are on the order of one forecast day. This applies to the new SEEPS score, as well as to ETS and PSS.

<table>
<thead>
<tr>
<th>( n_{\text{min}} )</th>
<th>( n )</th>
<th>( N )</th>
<th>( 1 - \text{SEEPS} )</th>
<th>( \text{BR} )</th>
<th>( \text{HR} )</th>
<th>( \text{FBI} )</th>
<th>( \text{ETS} )</th>
<th>( \text{PSS} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.82</td>
<td>863</td>
<td>0.826</td>
<td>0.328</td>
<td>0.929</td>
<td>1.040</td>
<td>0.763</td>
<td>0.875</td>
</tr>
<tr>
<td>2</td>
<td>4.63</td>
<td>670</td>
<td>0.811</td>
<td>0.332</td>
<td>0.926</td>
<td>1.052</td>
<td>0.742</td>
<td>0.863</td>
</tr>
<tr>
<td>4</td>
<td>6.23</td>
<td>390</td>
<td>0.800</td>
<td>0.331</td>
<td>0.925</td>
<td>1.061</td>
<td>0.731</td>
<td>0.858</td>
</tr>
<tr>
<td>6</td>
<td>7.90</td>
<td>198</td>
<td>0.796</td>
<td>0.333</td>
<td>0.922</td>
<td>1.059</td>
<td>0.725</td>
<td>0.853</td>
</tr>
<tr>
<td>8</td>
<td>10.22</td>
<td>79</td>
<td>0.781</td>
<td>0.344</td>
<td>0.910</td>
<td>1.049</td>
<td>0.705</td>
<td>0.837</td>
</tr>
</tbody>
</table>
• In all models, differences between the tropics and extratropics are large. SEEPS scores at \(D + 1\) in the tropics are comparable to those at \(D + 6\) in the extratropics.

• There is a marked decrease of extratropical 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 precipitation” category, when either “dry” or “heavy precipitation” was observed, highlighting the two distinct problems of (i) overprediction of drizzle and (ii) difficulty in predicting heavy precipitation.

• Choices of parameter values in the verification (like the threshold for the 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 extratropics can be attributed to the fact that gridbox 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 resampling, and by performing sensitivity studies, we have attempted to increase confidence in the validity of the results.

Acknowledgments. We 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


Ghelli, A., and F. Lalalurette, 2000: Verifying precipitation forecasts using upscaled observations. ECMWF Newsletter, No. 87, ECMWF, Reading, United Kingdom, 9–17.


Sundqvist, H., E. Berge, and J. E. Kristjansson, 1989: Condensation and cloud parameterization studies with a mesoscale


WWRP/WGNE Joint Working Group on Verification, 2008: Recommendations for the verification and intercomparison of QPFs and PQPFs from operational NWP models. WMO Tech Doc. 1485, WWRP 2009-1, revision 2, 37 pp.