Hydrol. Earth Syst. Sci. Discuss., 11, 919–944, 2014
www.hydrol-earth-syst-sci-discuss.net/11/919/2014/
doi:10.5194/hessd-11-919-2014

Hydrology and Earth System Sciences
Open Access
Discussions

Global meteorological drought – Part 2: Seasonal forecasts

E. Dutra[1], W. Pozzi[2], F. Wetterhall[1], F. Di Giuseppe[1], L. Magnusson[1], G. Naumann[3], P. Barbosa[3], J. Vogt[3], and F. Pappenberger[1,4]

[1]European Centre for Medium-Range Weather Forecasts, Reading, UK
[2]Group on Earth Observations, Geneva, Switzerland
[3]Joint Research Centre, Institute for Environment and Sustainability, Ispra, Italy
[4]College of Hydrology and Water Resources, Hohai University, Nanjing, China

Received: 16 December 2013 – Accepted: 22 December 2013 – Published: 17 January 2014

Correspondence to: E. Dutra (emanuel.dutra@ecmwf.int)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Global seasonal forecasts of meteorological drought using the standardized precipitation index (SPI) are produced using two datasets as initial conditions: the Global Precipitation Climatology Center (GPCC) and the ECMWF ERA-Interim reanalysis (ERAI); and two seasonal forecasts of precipitation: the most current ECMWF seasonal forecast system and climatologically based ensemble forecasts. The forecast skill is concentrated on verification months where precipitation deficits are likely to have higher drought impacts and grouped over different regions in the world. Verification of the forecasts as a function of lead time revealed a reduced impact on skill for: (i) long lead times using different initial conditions, and (ii) short lead times using different precipitation forecasts. The memory effect of initial conditions was found to be 1 month lead time for the SPI-3, 3 to 4 months for the SPI-6 and 5 months for the SPI-12. Results show that dynamical forecasts of precipitation provide added value, a skill similar or better than climatological forecasts. In some cases, particularly for long SPI time scales, it is very difficult to improve on the use of climatological forecasts. Our results also support recent questions whether seasonal forecasting of global drought onset was essentially a stochastic forecasting problem. Results are presented regionally and globally, and our results point to several regions in the world where drought onset forecasting is feasible and skilful.

1 Introduction

Seasonal forecasting is an essential component of an early drought forecasting system that can provide advance warning and alleviate drought impacts (Pozzi et al., 2013). The use of seasonal forecasts in such a system is mainly dependent on the actual predictability of drought conditions, that are dependent on the predictability of precipitation (Gianotti et al., 2013). Dynamical seasonal forecasting has evolved significantly in the last 20 yr, from early studies using simplified models (e.g. Cane et al., 1986)

HESSD
11, 919–944, 2014

to the modern multi-model systems (e.g. Palmer et al., 2004; Kirtman et al., 2013) which rely on coupled atmosphere-ocean models. With the increased skill of these dynamical forecasts, their use has increased, in particular in sectorial applications (e.g. Pappenberger et al., 2013),such as meteorological droughts (Yuan and Wood, 2013; Yoon et al., 2012; Mo et al., 2012; Dutra et al., 2013). Seasonal forecasting is not limited to dynamical models; several statistical techniques have been also developed (Barros and Bowden, 2008; Mishra and Desai, 2005). In this study, the European Centre for Medium-Range Weather Forecasts (ECMWF) latest dynamical seasonal forecast system is used. Different monitoring datasets are combined to the forecasted fields to generate global probabilistic meteorological drought seasonal forecasts.

Monitoring of the actual conditions is an essential part of the system, providing initial condition information (Shukla et al., 2013), and this forecasting system is initialised with the drought monitoring products which have been widely explained in the companion Part 1 paper. By extending to the global scale what was initially done in Dutra et al. (2013) in four African basins, this work tries to answer three general questions: (i) what is the importance of the monitoring in the forecast skill? (ii) what is the added value of using dynamical seasonal forecasts as compared with climatological forecasts? and (iii) what is the skill of these forecasts to predict drought onset (in aggregate, global terms)? The datasets used in this study and the skill metrics are presented in the following sections followed by the results. In the last section, the main conclusions are presented.

HESSD
11, 919–944, 2014

2 Methods

2.1 Seasonal forecasts

2.1.1 Precipitation datasets

In this study we use the ECMWF current seasonal forecasts system (System 4, hereafter S4, Molteni et al., 2011). This is a dynamical forecast system based on an atmospheric–ocean coupled model, which has been operational at ECMWF since 2011. The horizontal resolution of the atmospheric model is TL255 (about 0.7°) with 91 vertical levels in the atmosphere. S4 generates 51 ensemble members in real-time, with 30 yr (1981–2010) of back integrations (hindcasts) with 15 ensemble members with 6 months lead time. Molteni et al. (2011) provides a detailed overview of S4 performance. In addition to the dynamical seasonal forecasts, climatological forecasts (CLM) were also generated by randomly sampling past years from the observational dataset.

The reference precipitation dataset is the Global Precipitation Climatology Centre (GPCC, http://gpcc.dwd.de) full reanalysis version 6 (Schneider et al., 2011), that is available since 1901 to 2010 globally on a $1^\circ \times 1^\circ$ regular grid. In this study GPCC is used for verification. Also, the test for drought-like conditions is made by merging and blending the GPCC precipitation observations with forecast precipitation, so that GPCC also serves as an initial condition. Additionally, the ECMWF ERA-Interim reanalsysis (ERAI, Dee et al., 2011) that is available since 1979 to present expired month with the same resolution as S4 was also tested as initial conditions for the drought indicator. A detailed comparison of GPCC and ERAI for drought monitoring is presented in the companion Part 1 paper.

2.1.2 Drought indicator

As in Part 1, we selected the Standardized Precipitation Index (SPI, Mckee et al., 1993) as a meteorological drought index. The SPI is a transformation of the accumulated

HESSD
11, 919–944, 2014

precipitation in a specific time period (typically the previous 3, 6 and 12 months, denoted as SPI-3, SPI-6 and SPI-12 respectively) into a normal distribution of mean zero and standard deviation 1. The extension of the SPI from the monitoring period, i.e. past (can be also interpreted as initial conditions) to the seasonal forecasts range is performed by merging the seasonal forecasts of precipitation with the monitoring product. This study follows the same methodology that Dutra et al. (2013) applied to several basins in Africa, but in this case the SPI calculations are performed globally for each $1^{\circ} \times 1^{\circ}$ grid cell. Similar methodologies have been also used recently by Yoon et al. (2012) and Yuan and Wood (2013) using different monitoring and seasonal forecasts datasets. The SPI is a measure of incoming precipitation deficiency, and many additional factors determine the severity of drought that ensue, if any (Lloyd-Hughes, 2013).

Having two seasonal forecasts datasets (S4 and CLM) and two monitoring dataset (GPCC and ERAI), we generated seasonal re-forecasts of the SPI-3, 6 and 12 using four configurations:

- GPCC monitoring and S4 forecasts (GPCC S4);
- GPCC monitoring and climatological forecasts (GPCC CLM);
- ERAI monitoring and S4 forecasts (ERAI S4);
- ERAI monitoring and climatological forecast (ERAI CLM).

All four configuration provide a 30 yr hindcast period (1981–2010) with 15 ensemble members including forecasts issued every month with 6 months lead time. The GPCC CLM and ERAI CLM configurations constitute counterparts to the Ensemble Streamflow Prediction (ESP) method used by Yuan and Wood (2013). In these configurations, all the forecast skill comes from the monitoring period (or initial conditions) and they are used as reference forecasts.

HESSD
11, 919–944, 2014

2.2 Verification

2.2.1 Regions and seasons

Considering the large dataset of hindcasts (4 configurations, 3 SPI time scales, 12 initial forecast dates and 30 yr length dataset), the verification needed to be targeted to the specific drought application. Therefore, the evaluation of the forecasts is mainly focused on large regions adapted from Giorgi and Francisco (2000) (see Table 1, and Fig. S1 in the Supplement), and also used in part 1. Setting up these regions pools the grid cells together, increasing sample size, improving the quality of the verification statistics. A second point is that the seasonal forecasts relevance and skill is dependent on the different seasons for each location. Rainfall in many regions can be limited to particular seasons, so drought forecasts must be targeted to those seasons. In a global analysis, the wide variety of precipitation regimes makes it difficult to present the results synthetically for all the different initial forecasts calendar months. Since this paper is focused on drought events, the verification of the forecasts is performed for a specific calendar month where precipitation anomalies (in that month and previous months) are likely to have a higher impact. Using the mean annual cycle of GPCC precipitation in each region, we calculated the calendar month (for each region) with maximum accumulated precipitation in the previous 3 and 6 months, including the selected month (see Table 1). The calendar month with 3 months maximum accumulated precipitation is used to verify the SPI-3, while the calendar month with 6 months maximum accumulated precipitation is used to verify the SPI-6 and SPI-12. Therefore, in the results section the spatial maps displaying scores for different lead times refer to different verification calendar months. While this stratification on verification date might be more difficult to follow, it allows focusing on the season of interests and gives more emphasis on the forecast lead time.

HESSD
11, 919–944, 2014

2.2.2 Metrics

There is a large number of possible verification metrics that can be applied to probabilistic forecasts. In this paper, we focus on the root mean square (RMS) error and anomaly correlation of the ensemble mean and the relative operating characteristics (ROC) of the SPI below −0.8. The −0.8 threshold was selected as suggested by Yuan and Wood (2013) and Svoboda et al. (2002).

The RMS error of the ensemble mean for a specific region, initial forecast calendar month and lead time is calculated as:

$$\mathrm{RMS} = \frac{1}{n_\mathrm{t}} \sum_{i=1}^{n_\mathrm{t}} \left[\frac{1}{n_\mathrm{p}} \sum_{k=1}^{n_\mathrm{p}} \left(\overline{X(i,k)} - Y(i,k) \right)^2 \right]^{0.5} \quad (1)$$

where n_t is the number of years (30), n_p is the number of points in the particular regions, $Y(i,k)$ the observations for a specific year (i) and grid-point (k) and $\overline{X(i,k)}$ is the forecast ensemble mean. The RMS error confidence intervals are calculated for the temporal mean assuming a normal distribution. The time-mean of the RMS error of the ensemble mean should equal the time-mean of the ensemble spread about the ensemble-mean forecasts in a perfect forecasts (Palmer et al., 2006). The time-mean ensemble spread about the ensemble mean forecast is calculated as:

$$\mathrm{RMS\ (spread)} = \frac{1}{n_\mathrm{t}} \sum_{i=1}^{n_\mathrm{t}} \left[\frac{1}{n_\mathrm{p}} \sum_{k=1}^{n_\mathrm{p}} \left\{ \frac{1}{n_\mathrm{e}} \sum_{j=1}^{n_\mathrm{e}} \left(X(j,i,k) - \overline{X(i,k)} \right)^2 \right\} \right]^{0.5} \quad (2)$$

where n_e is the number of ensemble members (15) and $X(j,i,k)$ is the forecast ensemble member (j) in year (i) and grid-point (k). As in (1) $\overline{X(i,k)}$ represents the forecast ensemble mean of all n_e ensemble members.

HESSD
11, 919–944, 2014

For the anomaly correlation coefficient (ACC), we first calculate the grid-point Pearson correlation (r_k) in the form:

$$r_k = \frac{\sum_{i=1}^{n_t} (Y(i,k)' \overline{X(i,k)'})}{\left[\sum_{i=1}^{n_t} (Y(i,k)')^2\right]^{0.5} \left[\sum_{i=1}^{n_t} (\overline{X(i,k)'})^2\right]^{0.5}} \quad (3)$$

where $'$ denotes the temporal anomaly (after removing the temporal mean). The grid-point r_k is then averaged over the particular region with the fisher and inverse-fisher transformation:

$$\mathrm{ACC} = \tanh\left[\frac{1}{n_p}\sum_{i=1}^{n_p} \arctan h(r_k)\right]. \quad (4)$$

The confidence intervals of the anomaly correlation are calculated by a 1000 bootstrap temporal re-sampling – re-calculating Eq. (3) with random temporal sampling replacement. The ACC varies between −1 to 1 with 1 being a perfect forecast, and below 0 there is no skill to −1 were the forecasts are in anti-phase with the observations.

There are four possible cases for successful drought detection (prediction) and unsuccessful drought prediction. In one case, called the "hit rate" a forecast of drought is made and drought is, indeed observed (the number of cases for which this holds true: case a). In a second case (case b), a drought onset is predicted, but the drought is not in fact observed. In a third case (case c), no drought is forecasted, but, in fact, a drought is actually observed. In the fourth case (case d), no drought is forecast (predicted), and none is observed. A "false alarm rate" is the number of cases for which is drought is forecast, but not observed (case b) divided by the sum of case b and the rate of no drought being forecast and no drought being observed (case d). The false alarm ratio, on the other hand, is the number of cases where a drought is predicted, but a drought

HESSD
11, 919–944, 2014

fails to appear (case b) divided by the sum of hit rate (case a) added to the false alarm rate (b). Since the number of cases of drought observed globally is relatively small (while their impacts are nonetheless severe), this means that case d, no drought and no forecast, most usually occurring weather events, is very much larger than either case a or case b (or case c, for that matter). Since the false alarm rate contains case d in the denominator, having such a large number in the denominator makes the false alarm rate very small. For this reason, false alarm ratio is used preferentially over false alarm rate. Using these definitions, the probability of detection is the hit rate divided by the sum of the hit rate (case a) added to the number of cases where no drought is predicted but one is in fact observed (case c). The probability of a false alarm, in turn, is the number of events in which a false alarm rate (a drought being forecast but not observed) (case b) occurs divided by the sum of the hit rate added to number of cases in which a drought forecast but not observed (case b).

The relative operating characteristics (ROC) measures the skill of probabilistic categorical forecasts, while the previous metrics only evaluate the ensemble mean. The ROC diagram displays the false alarm rate (F) as a function of hit rate (HR) for different thresholds (i.e. fraction of ensemble members detecting an event) identifying whether the forecast has the attribute to discriminate between an event or not. The area under the ROC curve is a summary statistics representing the skill of the forecast system. The area is standardized against the total area of the figure, such that a perfect forecast has an area of 1 and a curve lying along the diagonal (no information, HR = F) has an area of 0.5. The results presented in the paper refer to each region. This was achieved by using all the grid-points in a region when calculating the contingency tables for the F and HR estimates. The forecasts and verification were transformed into an event (or no event) based on the underlying grid-point distributions. This spatial integration has the advantage of increasing the sample size used to build the contingency table while no spatial information is retained. To estimate the uncertainty of the ROC scores and curves in the ROC diagram a 1000 bootstrap re-sampling with replacement procedure was applied. The contingency tables and the ROC scores were calculated

HESSD
11, 919–944, 2014

1000 times: in each calculation the original forecast and verification grid-point time series are randomly replaced (allowing repetition) and a new set of scores is calculated. The re-sampling was performed only on the time series, keeping all the grid-points, since the temporal sampling size (in our case 30 vales) is the largest source of uncertainty in the scores estimation. The 95 % confidence intervals are estimated from the percentiles 2.5 and 97.5 of the 1000 bootstrap values.

The skill scores measure the difference between the score of the forecast and the score of a benchmark forecast, normalized by the potential improvement and are calculated as:

$$\text{ROC skill score} = (s - s_0)/(s_1 - s_0) \quad (5)$$

where s is the actual score, s_0 is the reference (or benchmark) score and s_1 is the score of a perfect forecast. The ROC skill score, in respect to a forecast with no skill can be calculated by setting $s_0 = 0.5$ and $s_1 = 1$, or setting s_0 to the ROC score of another benchmark forecast. The skill score varies between $-\infty$ to 1 with values below 0 indicating that the forecast is worse than the reference forecast, and 1 a perfect forecast.

2.2.3 Drought onset

To permit the comparison of the ECMWF model results with the US National Multi-model Ensemble results, presented by Yuan and Wood (2013), we have used their definition of drought onset: a drought event is defined when the SPI-6 is below −0.8 for at least 3 months, and the drought onset month is the first month that the SPI-6 falls below the threshold. In the last section of the results, we present an evaluation of the drought onset forecast skill of the different configurations with a global perspective (not following the regions definitions). Some of our verification metrics also overlap with Yuan and Wood (2013).

HESSD
11, 919–944, 2014

3 Results

3.1 Regional evaluation

For each one of the regions in Table 1 a summary figure was produced displaying the evolution of the RMS, ACC and the area under the ROC curve with lead time for the specific verification date (also in Table 1) and for the SPI-3, 6 and 12 (Fig. 1 for South Africa, and Fig. S2 to S19 in the Supplement for the remaining regions). This study will not exhaustively examine forecast skill within each individual region, although results are available within the auxiliary material for scrutiny.

In all regions there is a clear difference of the RMS error of the ensemble mean at lead times 0 and 1 between the forecasts using GPCC, as opposed to using ERAI for the monitoring. ERAI has higher RMS errors. From lead time 2 onwards in the case of the SPI-3, and lead time 5 in the case of SPI-6, the forecasts using GPCC or ERAI as monitoring have the same RMS error since for these lead times only forecast precipitation is used. In East (Fig. S9, Supplement) and West East Africa (Fig. S9, Supplement) and West Africa (Fig. S20, Supplement) RMS error for ERAI merged with S4 decreases with forecast lead time, which might be contra intuitive. These results are the first indication of the importance of the monitoring quality, i.e. which precipitation dataset is chosen to merge with the forecast information, whether GPCC or ERAI, and the first indication of the importance of initial conditions on the SPI forecast skill. On the other hand, in other regions like South Africa (Fig. 1) ERAI S4 RMS errors increase with lead time, which is in line with previous findings of the quality of ERAI precipitation over South Africa when compared with East of West Africa (Dutra et al., 2013).

The comparison of the RMS error of the ensemble mean with the ensemble spread (dashed lines in Fig. 1) suggests that in general the forecasts are slightly under-dispersive. However, we do not consider the observations uncertainty (in this case the GPCC precipitation) that should be added to the ensemble spread when comparing with the RMS error of the ensemble mean. This might be also associated with the deterministic nature of the initial conditions, and the extension of the probabilistic

HESSD
11, 919–944, 2014

monitoring presented in the companion Part 1 paper could be of potential benefit to increase the spread of the forecasts. The anomaly correlation coefficient of the SPI forecasts, using GPCC or ERAI monitoring also highlight the importance of having a reliable source of precipitation for monitoring (illustrated by comparing GPCC and ERAI). The same conclusion will be shown in the analysis of the ROC scores.

When looking at the ROC score figures, the higher the value, the greater the skill; with increasing lead time, this skill decays. There is a clear difference in the decay of the ROC scores with lead time, particularly for GPCC S4, as shown in panels g–i of Fig. 1: the decay rate is much more rapid for SPI-3 than for SPI-12. SPI-3 only contains 3 months of information, whether this is forecast precipitation or GPCC (or ERAI) "observed" precipitation. SPI-12, on the other hand, may contain many more months of monitored precipitation in the merged monitored-forecast product which is then tested against the monitored precipitation. This is intrinsic to the SPI forecasting method that uses more information from the monitoring dataset for longer SPI lead time. An additional finding is that, the ROC scores of GPCC using the S4 forecasts (GPCC S4) are higher than the same S4 forecasts used with ERAI (ERAI S4) during the first few months of lead times, after which the GPCC's higher rate decays to a rate of decay with lead time nearly identical with ERAI.

A test of the importance of the choice of monitoring dataset upon forecast skill (FS) is provided by identifying the last forecast lead time where the ROC skill score of GPCC S4 (using ERAI S4 as reference forecast) is higher than 0.05 with 95 % confidence (Fig. 2). Skill scores above 0 indicate that GPCC merged with S4 has a higher skill than ERAI S4. However, due to the sampling associated with the bootstrapping and the confidence intervals estimation, we decided to use a higher threshold of 0.05. This approach is useful for highlighting and revealing which regions the selection of ERAI for monitoring, as opposed to GPCC has a stronger detrimental effect on skill (relative to GPCC) of the seasonal forecasts. It also displays the lead time memory of the initial conditions. Referring to Fig. 2 for the SPI-3 case all regions present a lead time of 1 month, while for the SPI-12 most of the regions present a lead time of 5 months.

HESSD
11, 919–944, 2014

For the SPI-6 the memory of the initial conditions varies between 3 and 4 months, the latter mainly in the tropical regions. This SPI-6 case illustrates, once again, as shown in Part 1, that the precipitation assembled within the tropics within GPCC has higher uncertainty (and that higher disagreement is found among precipitation datasets within the tropics due to the low density of the number of gages and observations).

As opposed to testing the importance of monitored precipitation data quality on forecast skill (FS), a test of the importance of forecast information (predicted precipitation) upon forecast skill (FS) is provided by identifying the first lead time where the ROC skill score of GPCC S4 (using GPCC CLM as a reference forecast) is higher than 0.05 with 95 % confidence (Fig. 3), i.e. comparing the quality of the precipitation forecast (S4 or CLM) in the SPI forecast skill. These lead times identify the added value of using the seasonal forecasts of precipitation from S4 above the practice of simply using a climatological forecast, i.e. does the seasonal forecast add value above that of pure climatology. For the case of SPI-3 the added value of using the S4 forecast information varies between 1 to 2 months with North Eurasia regions and Australia having the lower values. For the use of SPI-12, on the other hand, the added value of using S4 can reach 5 months lead time in the Mediterranean, South Africa and Southern South America, while there is no significant improvement in Northern Europe and the Northern America – regions were the skill of the original S4 precipitation forecasts is very much reduced. For the case of SPI-3, Southern South America (SSA) and East Africa have high value (4 months), like they do for the SPI-12 case, but the values for the Mediterranean (MED), East Asia, Australia, Amazonia, and Western North America (WNA) are low (1 month). Furthermore, as Northern Europe and all three North America regions were not statistically significant for SPI12, so this is true for Central and Eastern North America for SPI-6. Even in regions where there is little more added value GPCC S4 skill is always equal or higher than climatology, i.e. GPCC CLM. In some cases, particularly for long SPI time scales (SPI-12), the proportion of monitored precipitation merged with the forecast that is tested against the monitored precipitation is very high, and the monitored precipitation is being tested against itself.

HESSD
11, 919–944, 2014

3.2 Drought onset

In order to permit the results of the ECMWF forecasting model seasonal drought forecasts to be compared to the forecast models of US National Multi-Model Ensemble drought forecast tests (Yuan and Wood, 2013) (denoted YW13), the tests made upon each one degree grid cell are combined into a global samples with global means of the probability of detection (POD), and global means of false alarm ratio (FAR) and equitable threat score (ETS) for drought onset forecasts (Table 2): the climatology case of GPCC CLM is very similar to YW13's findings, obtained using Ensemble Streamflow Prediction. This study and YW13 deployed different precipitation datasets, as well as time interval of collected hindcasts; yet, despite these differences, the results are comparable, suggesting that the results, in general, are likely robust and that the expected skill of the SPI drought onset forecasts provided only by the initial conditions, as shown in their study, as well. The climatology cases of the two studies are not only similar: the forecast of GPCC S4 matches that of some of the other models analysed within YW13's Multi-model ensemble (MME). This study also overlaps with their multi-model skill estimates (North American Multimodel Ensemble with post-processing NMME2). ERAI-based forecasts have lower skill than the GPCC CLM using the equitable threat score metric. Again, the precipitation dataset chosen, and the quality of the precipitation dataset, has a major role in the skill of SPI forecasts. As has been noted in Sect. 2.2.2, false alarm rates are very small (in the order of 10^{-2}), since the denominator term of meteorological states of no drought condition and no forecasts of drought is so high, due to the relative low incidence of drought globally.

Each ensemble member conserves the SPI characteristics, that of mean zero and standard deviation of one (which arises due to the definition of SPI), but the ensemble mean (of all the ensembles) conserves the mean of zero only, whereas, to the contrary, the standard deviation (unlike the initial calculation of SPI that makes the standard deviation of 1 within each ensemble) falls below one, contrary to the definition of SPI. The standard deviation decline below one is pronounced for long lead times as the

HESSD
11, 919–944, 2014

ensemble spread increases. Despite the change in standard deviation, the drought onset forecasts skill is based on ensemble mean, in the case of the POD, FAR, and ETS (Table 2, and YW13) statistics, with these drought onset skill metrics (POD, FAR, and ETS) depending only the SPI falling below a certain threshold. One can rescale the forecast ensemble mean to retain the unit standard deviation and arrest its decline below 1, conforming to the definition of SPI. Such ensemble mean rescaling case is presented between the brackets in Table 2. This rescaling increases the probability of drought detection (POD) (as it should), but, in exchanging for increasing the number of false alarms, the false alarm ratio, with the overall result of conferring only a slight increase of the Equitable Threat Score (ETS). To retain the SPI definition, i.e. to ensure that the criterion for drought onset condition is maintained, (alternatively stated, for skill metrics that depend on the ensemble mean and on SPI thresholds), we recommend the scaling of the ensemble mean standard deviation. This rescaling can be also interpreted as the SPI calculated directly from the ensemble mean of the precipitation forecasts. Another potential use of this rescaling is the graphical display of the ensemble mean forecasts, that was explored by Mwangi et al. (2013), and provides the users with SPI forecast maps with units/range as the SPI during the monitoring phase.

To finalize the drought onset evaluation, the brier skill score is used, based upon the climatological frequency of drought events as reference, for the different experiments over each grid-cell of the globe (Fig. 4). The global spatial maps of the brier skill scores, for both the seasonal forecast case, GPCC S4, and the climatology case, GPCC CLM, exhibit similar spatial patterns to those observed in YW13's NMME results for the seasonal forecast POD equivalent case and the ESP climatology equivalent case. Our study and theirs with the NMME both reveal the clear benefit of a seasonal forecast over climatology, this being valid for our case of GPCC S4 when compared with GPCC CLM. The demonstration of seasonal forecast being better than climatology is found for the regions of Australia, East Africa, Northwest South America (Brazil), as well as other regions of the globe. Such a result was also found in YW13. Looking at the global brier score decomposition (Fig. S20, Supplement) shows that climatology, i.e. GPCC CLM,

HESSD
11, 919–944, 2014

has better reliability than GPCC S4 (per definition), while GPCC S4 has better resolution. The increased resolution in GPCC S4 with a small reduction of reliability (when compared with GPCC CLM) lead to better brier scores in GPCC S4. Figure 4 highlights how noisy are the individual grid cell scores globally. Assembling the grid cells into regions, on the other hand, increase the sample sizes within those regions, and permit us to investigate whether one region, as opposed to another, has consistently high skill scores (e.g. East Africa vs. West Africa).

Up until now, we have been looking at global and regional statistics of combined drought onset skill among all the hindcast samples. Here we test drought forecast skill for individual drought cases. Continuing with our example begun in companion paper Part 1, we provide SPI forecasts for the 2010/2011 drought in the Horn of Africa (Figs. S21 and S22, Supplement) and the 2012 drought in the US Great Plains (Figs. S23 and S24, Supplement). These examples also illustrate how the results of a probabilistic drought forecast would be "packaged" for skilled users (the counterpart to the probabilistic flood forecast case). The time series (Figs. S21 and S23, Supplement) show the GPCC S4 and GPCC CLM SPI forecasts issue in different initial dates and averaged over a region and overplayed with the verification. The spatial maps (Figs. S22 and S24, Supplement) comparing the actual verification SPI with four different examples of displaying a specific forecast: (i) ensemble mean, (ii) the ensemble mean rescaled, (iii) probability of the SPI > 0.8 (wet conditions); (iv) probability of the SPI < -0.8 (dry conditions).

4 Conclusions

This paper presents a general evaluation of meteorological drought seasonal forecasts using the standardized precipitation index constructed by merging different initial conditions and seasonal forecasts of precipitation. The forecast skill is targeted to verification months where precipitation deficits are likely to have higher drought impacts and grouped in 18 regions. Detailed analysis of drought forecasting skill within each

HESSD
11, 919–944, 2014

region is outside the scope of this paper, though results are made available in auxiliary material.

At the onset of this paper, three fundamental questions were posed. The first concerned the importance of the monitoring in the forecast skill.

The memory effect of initial conditions in the SPI forecasts was identified, comparing the S4 seasonal forecasts initialized with GPCC, when compared to the same S4 seasonal forecasts initialized and merged with ERA-Interim precipitation. This was found to be 1 month lead time in the case of SPI-3, 3-to-4 months for SPI-6, and 5 months for SPI-12. For earlier forecast lead times, the initial conditions of precipitation dominate the forecast skill, proving that good quality and reliable monitoring of precipitation is paramount.

The second question was centered on the added value of using ECMWF seasonal forecasts of precipitation when compared to climatological past samples of precipitation. Our results show the skill obtained in dynamical forecasts is always equal or above climatological forecasts, even for regions where the added value in terms of forecast lead time is reduced. In some cases, particularly for long SPI time scales, the improvement upon climatological forecasts is negligible. For long SPI time scales (such as SPI+12), the proportion of monitored precipitation when added to the forecast can be very high and almost the same as the monitored precipitation against which it is being tested.

Finally we questioned the skill of dynamical forecast in terms of drought onset. We have defined drought onset, based on the definition used by Yuan and Wood (2013), so that the results from the ECMWF forecasting model seasonal forecast system S4 could be compared against the drought forecasts from other forecasting ensemble models within the US National Multi-model Ensemble presented in the Yuan and Wood (2013) study. Although different datasets and different periods were used, the estimates of drought onset skill for climatological forecasts are coincident. Therefore, we suggest that they are reasonable independent from data and intrinsic to the SPI seasonal forecasting methodology. We also propose that in order to evaluate forecasts ensemble

HESSD
11, 919–944, 2014

mean in terms of SPI thresholds, the ensemble mean should be rescaled to guarantee a standard deviation of one. This is further beneficial when presenting the forecasts graphically. Yuan and Wood (2013) questioned whether seasonal forecasting of global drought onset was largely or solely a stochastic forecasting problem only. However, our results show that this is not a global result: within several regions in the world drought onset forecasting is feasible and skilful.

Supplementary material related to this article is available online at http://www.hydrol-earth-syst-sci-discuss.net/11/919/2014/hessd-11-919-2014-supplement.pdf.

Acknowledgements. This work was funded by the FP7 EU project DEWFORA (http://www.dewfora.net).

References

Barros, A. P. and Bowden, G. J.: Toward long-lead operational forecasts of drought: an experimental study in the Murray–Darling River Basin, J. Hydrol., 357, 349–367, doi:10.1016/j.jhydrol.2008.05.026, 2008.

Cane, M. A., Zebiak, S. E., and Dolan, S. C.: Experimental forecasts of El Nino, Nature, 321, 827–832, 1986.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. Roy. Meteorol. Soc., 137, 553–597, doi:10.1002/qj.828, 2011.

Dutra, E., Di Giuseppe, F., Wetterhall, F., and Pappenberger, F.: Seasonal forecasts of droughts in African basins using the Standardized Precipitation Index, Hydrol. Earth Syst. Sci., 17, 2359–2373, doi:10.5194/hess-17-2359-2013, 2013.

HESSD
11, 919–944, 2014

Gianotti, D., Anderson, B. T., and Salvucci, G. D.: What do rain gauges tell us about the limits of precipitation predictability?, J. Climate, 26, 5682–5688, doi:10.1175/jcli-d-12-00718.1, 2013.
Giorgi, F. and Francisco, R.: Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM, Clim. Dynam., 16, 169–182, doi:10.1007/pl00013733, 2000.
Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M. P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippett, M. K., Barnston, A. G., Li, S., Rosati, A., Schubert, S. D., Rienecker, M., Suarez, M., Li, Z. E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W. J., Denis, B., and Wood, E. F.: The North American Multi-Model Ensemble (NMME): Phase-1 Seasonal to Interannual Prediction, Phase-2 Toward Developing Intra-Seasonal Prediction, B. Am. Meteorol. Soc., doi:10.1175/bams-d-12-00050.1, in press, 2013.
Lloyd-Hughes, B.: The impracticality of a universal drought definition, Theor. Appl. Climatol., doi:10.1007/s00704-013-1025-7, in press, 2013.
Mckee, T. B., Doesken, N. J., and Kleist, J.: The relationship of drought frequency and duration to time scales, Eight Conference on Applied Climatology, Anahaim, California, 179–184, 1993.
Mishra, A. K. and Desai, V.: Drought forecasting using stochastic models, Stoch. Env. Res. Risk A., 19, 326–339, doi:10.1007/s00477-005-0238-4, 2005.
Mo, K. C., Shukla, S., Lettenmaier, D. P., and Chen, L.-C.: Do Climate Forecast System (CFSv2) forecasts improve seasonal soil moisture prediction?, Geophys. Res. Lett., 39, L23703, doi:10.1029/2012gl053598, 2012.
Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnunson, L., Mogensen, K., Palmer, T., and Vitart, F.: The new ECMWF seasonal forecast system (System 4), ECMWF Tech. Memo. 656, ECMWF, Reading, UK, 49 pp., 2011.
Mwangi, E., Wetterhall, F., Dutra, E., Di Giuseppe, F., and Pappenberger, F.: Forecasting droughts in East Africa, Hydrol. Earth Syst. Sci. Discuss., 10, 10209–10230, doi:10.5194/hessd-10-10209-2013, 2013.
Palmer, T. N., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., and Smith, L.: Ensemble prediction: a pedagogical prespective, ECMWF Newslett., 106, 10–17, 2006.

HESSD
11, 919–944, 2014

Palmer, T. N., Doblas-Reyes, F. J., Hagedorn, R., Alessandri, A., Gualdi, S., Andersen, U., Feddersen, H., Cantelaube, P., Terres, J. M., Davey, M., Graham, R., Délécluse, P., Lazar, A., Déqué, M., Guérémy, J. F., Díez, E., Orfila, B., Hoshen, M., Morse, A. P., Keenlyside, N., Latif, M., Maisonnave, E., Rogel, P., Marletto, V., and Thomson, M. C.: Development of a european multimodel ensemble system for seasonal-to-interannual prediction (demeter), B. Am. Meteorol. Soc., 85, 853–872, doi:10.1175/bams-85-6-853, 2004.

Pappenberger, F., Wetterhall, F., Dutra, E., Di Giuseppe, F., Bogner, K., Alfieri, L., and Cloke, H. L.: Seamless forecasting of extreme events on a global scale, in: Climate and Land Surface Changes in Hydrology, edited by: Boegh, E., Blyth, E., Hannah, D. M., Hisdal, H., Kunstmann, H., Su, B., and Yilmaz, K. K., IAHS Publication, Gothenburg, Sweden, 3–10, 2013.

Pozzi, W., Sheffield, J., Stefanski, R., Cripe, D., Pulwarty, R., Vogt, J. V., Heim, R. R., Brewer, M. J., Svoboda, M., Westerhoff, R., van Dijk, A. I. J. M., Lloyd-Hughes, B., Pappenberger, F., Werner, M., Dutra, E., Wetterhall, F., Wagner, W., Schubert, S., Mo, K., Nicholson, M., Bettio, L., Nunez, L., van Beek, R., Bierkens, M., de Goncalves, L. G. G., de Mattos, J. G. Z., and Lawford, R.: Toward global drought early warning capability: expanding international cooperation for the development of a framework for monitoring and forecasting, B. Am. Meteorol. Soc., 94, 776–785, doi:10.1175/bams-d-11-00176.1, 2013.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., and Ziese, M.: GPCC Full Data Reanalysis Version 6.0 at 1.0°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data, [Data set], DWD, Germany, doi:10.5676/DWD_GPCC/FD_M_V6_100, 2011.

Shukla, S., Sheffield, J., Wood, E. F., and Lettenmaier, D. P.: On the sources of global land surface hydrologic predictability, Hydrol. Earth Syst. Sci., 17, 2781–2796, doi:10.5194/hess-17-2781-2013, 2013.

Svoboda, M., LeComte, D., Hayes, M., Heim, R., Gleason, K., Angel, J., Rippey, B., Tinker, R., Palecki, M., Stooksbury, D., Miskus, D., and Stephens, S.: The Drought Monitor, B. Am. Meteorol. Soc., 83, 1181–1190, 2002.

Yoon, J.-H., Mo, K., and Wood, E. F.: Dynamic-model-based seasonal prediction of meteorological drought over the contiguous United States, J. Hydrometeorol., 13, 463–482, doi:10.1175/jhm-d-11-038.1, 2012.

Yuan, X., and Wood, E. F.: Multimodel seasonal forecasting of global drought onset, Geophys. Res. Lett., 40, 4900–4905, doi:10.1002/grl.50949, 2013.

HESSD
11, 919–944, 2014

Table 1. List of regions used in this study. Adapted from Giorgi and Francisco (2000) (Fig. S1 in the Supplement and also Part 1). For each region, the calendar month with maximum accumulated precipitation in the previous 3 and 6 months (inclusie) is presented and was calculated from the mean annual cycles of GPCC.

Name	Acronym	Max 3 month	Max 6 month
Australia	AUS	Mar	Apr
Amazon Basin	AMZ	Mar	May
Southern South America	SSA	Aug	Oct
Central America	CAM	Sep	Oct
Western North America	WNA	Jan	Mar
Central North America	CNA	Jul	Sep
Eastern North America	ENA	Aug	Oct
Mediterranean Basin	MED	Jan	Mar
Northern Europe	NEU	Sep	Nov
Western Africa	WAF	Sep	Oct
East Africa	EAF	May	Aug
Southern Africa	SAF	Feb	Apr
Southeast Asia	SEA	Dec	Dec
East Asia	EAS	Aug	Sep
South Asia	SAS	Aug	Oct
Central Asia	CAS	Apr	May
Tibet	TIB	Aug	Sep
North Asia	NAS	Aug	Oct

HESSD
11, 919–944, 2014

Table 2. Global mean values of probability of detection (POD), false alarm ratio (FAR) and equitable threat score (ETS) for drought onset forecasts. The scores between brakets were calculated after scalling the ensemble mean (see Sect. 3.1 for details). The 95 % confidence intervals, estimated from 1000 samples bootrapping with replacement, returned similar values for all scores and models of approximatly ±0.01. The ESP and NMME2 models scores are included in this table for comparison purposes only and were retrived from Yuan and Wood (2013, see Table 1).

Model	POD	FAR	ETS
GPCC CLM	0.17 (0.27)	0.40 (0.57)	0.15 (0.21)
GPCC S4	0.30 (0.42)	0.47 (0.57)	0.25 (0.29)
ERAI CLM	0.14 (0.25)	0.85 (0.87)	0.09 (0.12)
ERAI S4	0.22 (0.31)	0.82 (0.84)	0.13 (0.14)
ESP	0.16	0.36	0.14
NMME2	0.32	0.42	0.24

HESSD
11, 919–944, 2014

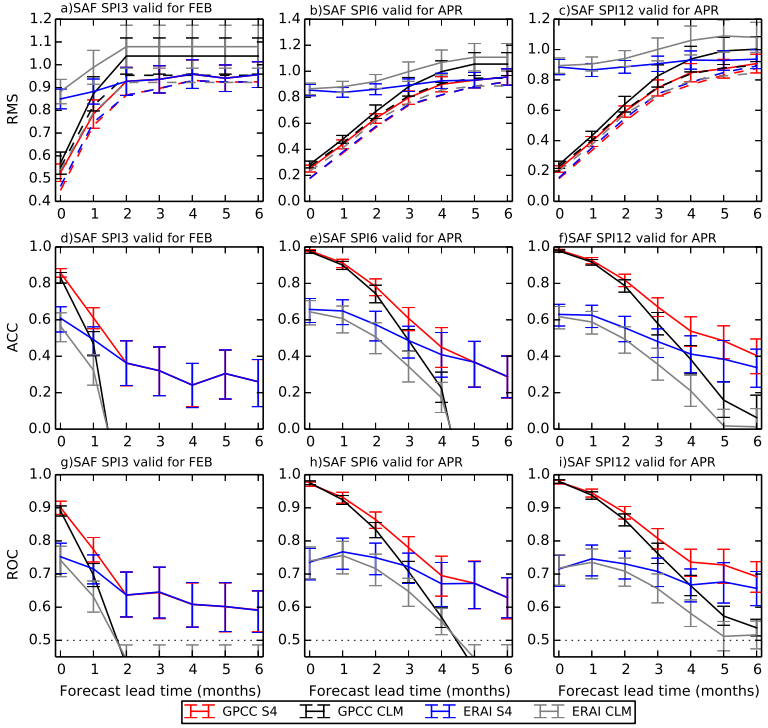

Fig. 1. Seasonal forecasts evaluation summary for the South Africa region (SAF) for the SPI-3 **(a, d, g)**, SPI-6 **(b, e, h)** and SPI-12 **(c, f, i)**. For each SPI time-scale the evaluation consist of threee panels displaying a specific score as a function of lead time (horizontal axis) for a specific verification date (in the title) for the GPCC S4 forecasts (red), GPCC CLM (black), ERAI S4 (blue) and ERAI CLM (grey). **(a)**–**(c)** RMS error of the ensemble mean and ensemble spread about the ensemble-mean in dashed; **(d)**–**(f)** Anomaly correlation coeficient; **(g)**–**(i)** area under the ROC curve for SPI forecasts below −0.8. The error bar in all panels denote the 95 % confidence intervals computed from 1000 samples bootstraping with re-sampling.

HESSD
11, 919–944, 2014

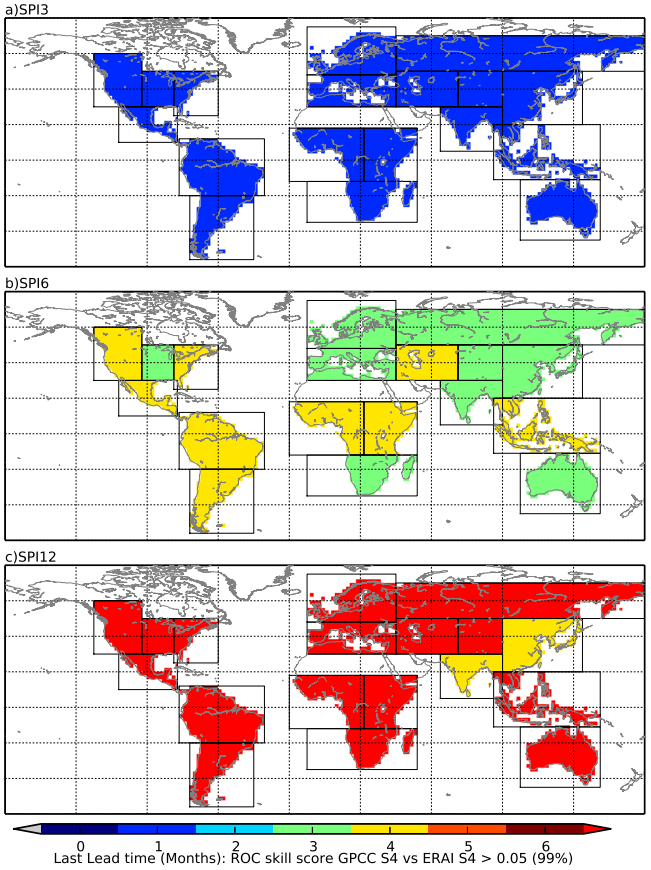

Fig. 2. Last forecast lead time (months) where the ROC skill score of GPCC S4 (using ERAI S4 as reference forecasts) is higher than 0.05 with 95 % confidence and the ROC of GPCC S4 is higher than 0 with 95 % confidence. Seasonal forecasts of the **(a)** SPI-3, **(b)** SPI-6 and **(c)** SPI-12. The forecast are verified in each region for the calendar month presented in Table 1.

HESSD
11, 919–944, 2014

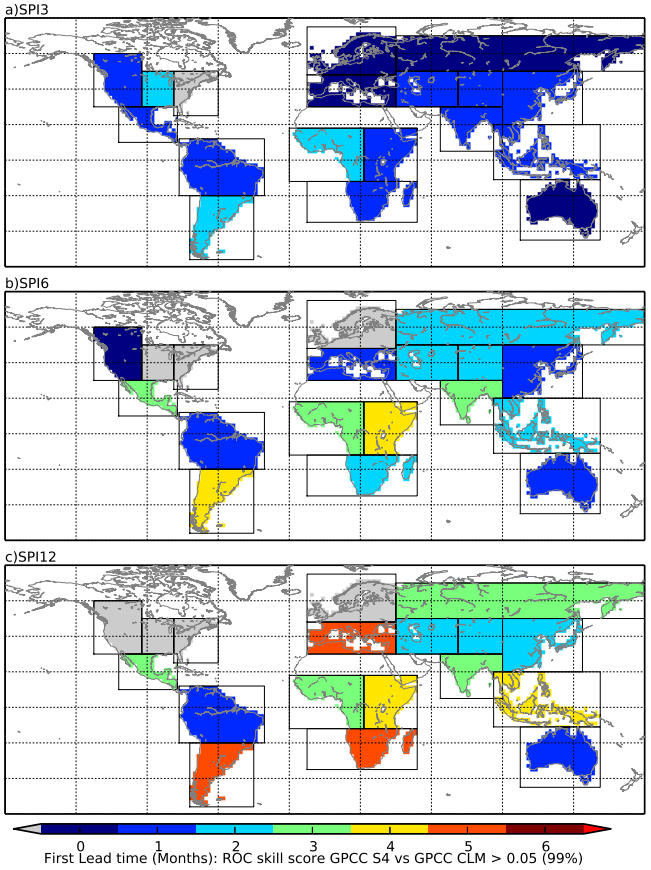

Fig. 3. First forecast lead time (months) where the ROC skill score of GPCC S4 (using GPCC CLM as reference forecasts) is higher than 0.05 with 95 % confidence and the ROC of GPCC S4 is higher than 0 with 95 % confidence. Seasonal forecasts of the **(a)** SPI-3, **(b)** SPI-6 and **(c)** SPI-12. The forecast are verified in each region for the calendar month presented in Table 1.

HESSD
11, 919–944, 2014

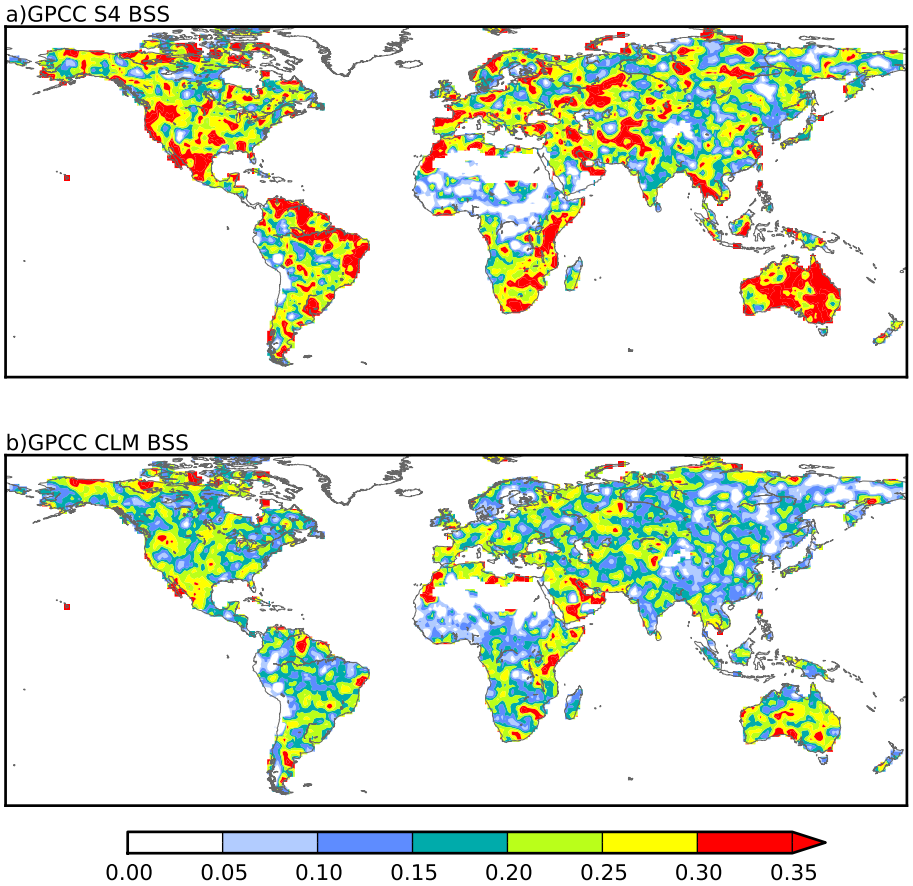

Fig. 4. Brier skill score for the drough onset forecasts of **(a)** GPCC S4 and **(b)** GPCC CLM. The reference forecasts for the skill score was the climatological frequency of the verification dataset. The original maps at 1° × 1° were smoothed with a 3 × 3 window.

HESSD
11, 919–944, 2014