Hydrol. Earth Syst. Sci., 18, 2669–2678, 2014
www.hydrol-earth-syst-sci.net/18/2669/2014/
doi:10.5194/hess-18-2669-2014

Hydrology and Earth System Sciences
Open Access

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]

[1]European Centre for Medium-Range Weather Forecasts, Reading, UK
[2]Group on Earth Observations, Geneva, Switzerland
[3]European Commission, Joint Research Centre, Institute for Environment and Sustainability, Ispra, Italy

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

Received: 16 December 2013 – Published in Hydrol. Earth Syst. Sci. Discuss.: 17 January 2014
Revised: – Accepted: 11 May 2014 – Published: 24 July 2014

Abstract. Global seasonal forecasts of meteorological drought using the standardized precipitation index (SPI) are produced using two data sets as initial conditions: the Global Precipitation Climatology Centre (GPCC) and the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-Interim reanalysis (ERAI); and two seasonal forecasts of precipitation, the most recent ECMWF seasonal forecast system and climatologically based ensemble forecasts. The forecast evaluation focuses on the periods where precipitation deficits are likely to have higher drought impacts, and the results were summarized over different regions in the world. The verification of the forecasts with lead time indicated that generally for all regions the least reduction on skill was found for (i) long lead times using ERAI or GPCC for monitoring and (ii) short lead times using ECMWF or climatological seasonal forecasts. The memory effect of initial conditions was found to be 1 month of lead time for the SPI-3, 4 months for the SPI-6 and 6 (or more) months for the SPI-12. Results show that dynamical forecasts of precipitation provide added value with skills at least equal to and often above that of climatological forecasts. Furthermore, it is very difficult to improve on the use of climatological forecasts for long lead times. Our results also support recent questions of 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, which are in turn dependent on the predictability of precipitation (Gianotti et al., 2013). Dynamical seasonal forecasting has evolved significantly in the last 20 years, from early studies using simplified models (e.g. Cane et al., 1986) to 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 data sets are combined with 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 initialized with the drought monitoring products which have been widely explained in the companion Part 1 paper. By extending the global scale that was initially done by Dutra et al. (2013) in four African basins, this work tries to answer three general

questions: (i) how sensitive are drought forecasts to the monitoring data set used? (ii) What is the added value of using dynamical seasonal forecasts in comparison with climatological forecasts? (iii) What is the skill of these forecasts to predict drought onset (in aggregated, global terms)? The data sets used in this study and the skill metrics are presented in Sect. 2 followed by the results and discussion in Sect. 3 and the main conclusions in Sect. 4.

2 Methods

2.1 Seasonal forecasts

2.1.1 Precipitation data sets

In this study we use the ECMWF seasonal forecast 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 about 80 km with 91 vertical levels in the atmosphere. S4 generates 51 ensemble members in real time, with 30 years (1981–2010) of back integrations (hindcasts) with 15 ensemble members and 6 months of lead time. Molteni et al. (2011) provide 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 reference data set to match the number of ensemble members in the hindcast.

The reference precipitation data set is the Global Precipitation Climatology Centre (GPCC) full reanalysis version 6 (Schneider et al., 2011), which has been available since 1901 to 2010 globally on a $1^\circ \times 1^\circ$ regular grid. In this study GPCC is used both as a reference data set (for the forecast verification) and as a monitoring data set (providing initial conditions). Additionally, the ECMWF ERA-Interim reanalysis (ERAI, Dee et al., 2011), which has been available since 1979 up to the present with the same resolution as S4, was also tested as monitoring for the drought indicator. A detailed comparison of GPCC and ERAI for drought monitoring is presented in the companion Part 1 paper (Dutra et al., 2014).

2.1.2 Drought indicator

As in Part 1, we selected the standardized precipitation index (SPI, Mckee et al., 1993) as a meteorological drought indicator. SPI is a transformation of the accumulated precipitation amount over 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 also be interpreted as initial conditions) to the seasonal forecast range, is performed by merging the seasonal forecasts of precipitation with the monitoring product. The merging of the two products is basically a concatenation of the monitoring with the seasonal forecast of precipitation. 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 also been used recently by Yoon et al. (2012) and Yuan and Wood (2013) (denoted YW13) using different monitoring and seasonal forecast data sets. 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 forecast data sets (S4 and CLM) and two monitoring data sets (GPCC and ERAI), we generated seasonal reforecasts 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 configurations provide a 30-year hindcast period (1981–2010) with 15 ensemble members including forecasts issued every month with 6 months of lead time. The GPCC CLM and ERAI CLM configurations constitute counterparts to the ensemble streamflow prediction (ESP) method used by YW13. To investigate the role of the monitoring, we generated an extra set of reforecasts using GPCC mean climatological precipitation for monitoring and S4 forecasts (GPCC_CLM S4). This configuration will not be presented in detail in the forecast verification, but it will be compared with GPCC S4 configuration as a proxy to quantify the importance of initial conditions in the forecast skill.

2.2 Verification

2.2.1 Regions and seasons

Considering the large size of the hindcast data sets (4 configurations, 3 SPI timescales, 12 initial forecast dates and 30 years), the verification was 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. Setting up these regions pools the grid cells together, increases sample size and improves the robustness of the verification statistics. A second point is that the seasonal forecast 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 forecast calendar months.

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 (inclusive) is presented and was calculated from the mean annual cycles of GPCC.

Name	Acronym	Max 3 months	Max 6 months
Australia	AUS	March	April
Amazon Basin	AMZ	March	May
Southern South America	SSA	August	October
Central America	CAM	September	October
Western North America	WNA	January	March
Central North America	CNA	July	September
Eastern North America	ENA	August	October
Mediterranean Basin	MED	January	March
Northern Europe	NEU	September	November
Western Africa	WAF	September	October
East Africa	EAF	May	August
Southern Africa	SAF	February	April
Southeast Asia	SEA	December	December
East Asia	EAS	August	September
South Asia	SAS	August	October
Central Asia	CAS	April	May
Tibet	TIB	August	September
North Asia	NAS	August	October

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 the maximum 3-month accumulated precipitation was used to verify SPI-3, while the calendar month with the maximum 6-month accumulated precipitation was used to verify SPI-6 and SPI-12. Consequently, the spatial maps of scores for different lead times refer to different verification calendar months. While this stratification on verification date is somewhat arbitrary, it allows focusing on the season of interest and gives more emphasis on the forecast lead time.

2.2.2 Metrics

The root mean square (rms) error of the ensemble mean for a specific region, initial forecast calendar month and lead time is calculated as

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

where n_t is the number of years (30), n_p the number of points in the particular regions, $Y(i,k)$ the observations for a specific year (i) and grid point (k), and $\overline{X(i,k)}$ 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 forecast (Palmer et al., 2006). The time-mean ensemble spread about the ensemble mean forecast is calculated as

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

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 Eq. (1) $\overline{X(i,k)}$ represents the forecast ensemble mean of all n_e ensemble members.

The anomaly correlation coefficient (ACC) of the ensemble mean is calculated as in Molteni et al. (2011). First the grid-point Pearson correlation (r_k) is calculated in the following form:

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

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:

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

The confidence interval of the anomaly correlation was calculated by a 1000-bootstrap temporal resampling and recalculating Eqs. (3) and (4) with random temporal sampling replacement. The ACC varies between -1 and 1 with 1 being a perfect forecast, and below 0 there is no skill to -1 where the forecasts are in antiphase with the observations.

The relative operating characteristic (ROC) measures the skill of probabilistic categorical forecasts, while the previous two metrics only evaluate the ensemble mean. The ROC diagram displays the false alarm rate (F) as a function of hit rate (HR) for different fractions of ensemble members detecting an event. 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 F and HR. The forecasts and verification were transformed into an event (or no event) by determining

if SPI is below −0.8 as suggested by YW13 and Svoboda et al. (2002). The 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 resampling with replacement procedure was applied. The contingency tables and the ROC scores were calculated 1000 times: in each calculation the original forecast and verification grid-point time series were randomly replaced (allowing repetition), and a new set of scores was calculated. The resampling was performed only on the time series, keeping all the grid points, since the temporal sampling size (in our case 30 values) is the largest source of uncertainty in the score 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 calculated as

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

where s is the ROC score of the forecast, s_0 the ROC score of a benchmark forecast and s_1 the ROC score of a perfect forecast. The ROC skill score, with 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 $-\infty$ and 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 compare the ECMWF model results with the US National Multi-Model Ensemble results, presented by YW13, 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 YW13.

3 Results

3.1 Regional evaluation

For each 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 Figs. S2–S19 in the Supplement for the remaining regions). This study will not exhaustively examine forecast skill within each individual region, although results are available in the Supplement for scrutiny.

In all regions there is a clear difference of the rms error of the ensemble mean for lead times 0 and 1 months between the forecasts using GPCC, in comparison with using ERAI for the monitoring, for example; ERAI has higher rms errors. From lead time 2 (5) months and onwards SPI-3 (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. The forecasts using ERAI for monitoring are penalized when compared with the forecasts using GPCC for monitoring, since GPCC is used as a reference data set (for the forecast verification). These results do not consider the uncertainties in GPCC that are discussed in more detail in the companion Part 1 paper, in particular the changes in the number of rain gauges used in the data set. In East Africa (Fig. S7 in the Supplement) and West Africa (Fig. S18 in the Supplement) the rms error for ERAI merged with S4 decreases with forecast lead time, which might be counter-intuitive, and is associated with the problems of ERAI's inter-annual precipitation over those regions (Dutra et al., 2013). These results are the first indication of the importance of the monitoring quality (i.e. whether GPCC or ERAI was merged with the forecast information) and subsequently 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. This is in line with previous findings of the quality of ERAI precipitation over South Africa when compared with East or West Africa (Dutra et al., 2013).

In general the forecasts are slightly under-dispersive, which can be seen from the dashed lines in Fig. 1. However, we do not consider the observation uncertainty (in this case the GPCC precipitation), which 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 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 highlights 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.

There is a clear difference in the decay of the ROC scores with lead time, particularly for GPCC S4, as shown in Fig. 1g–i: 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 data set for longer SPI lead times. Additionally, the ROC scores of GPCC using the S4 forecasts (GPCC S4)

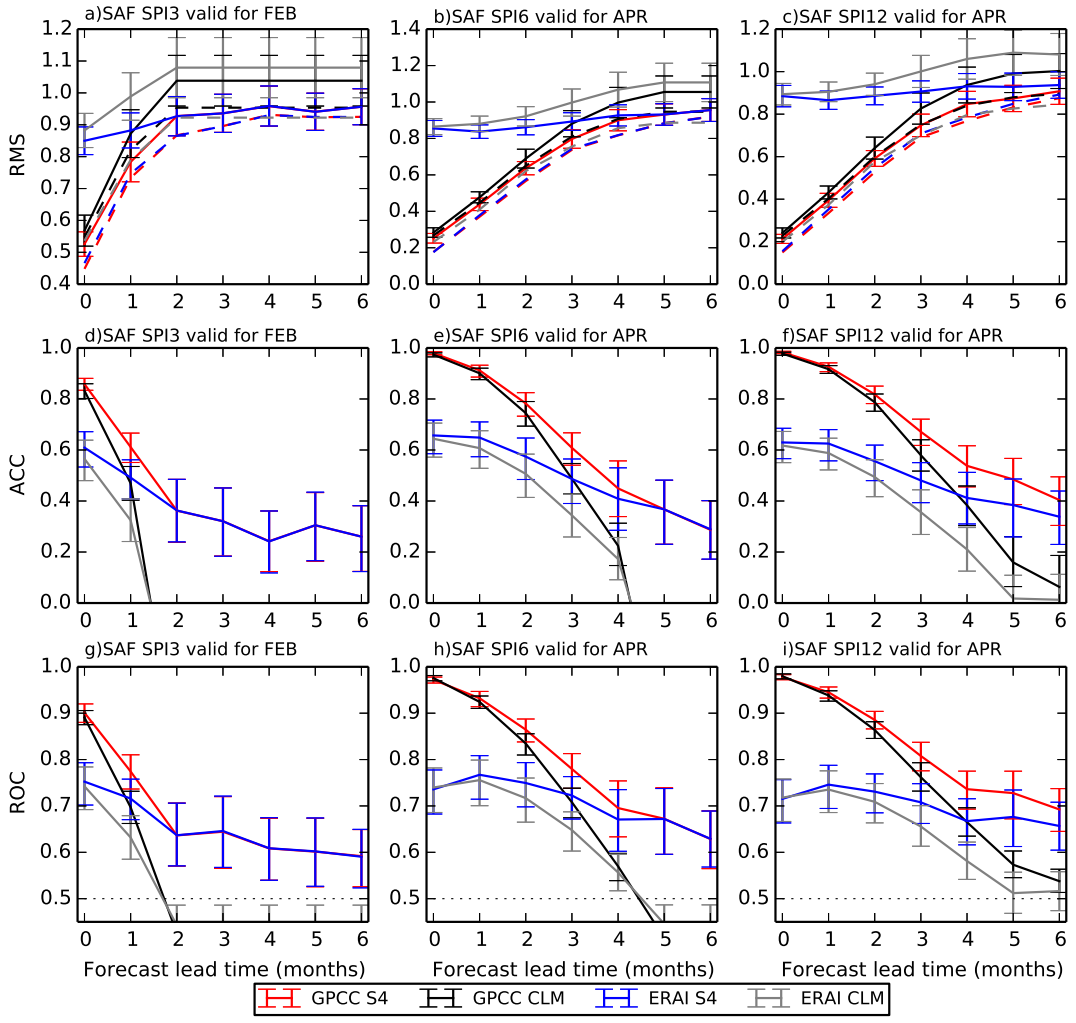

Figure 1. Seasonal forecast evaluation resume for the South Africa region (SAF) for the SPI-3 **(a, d, g)**, SPI-6 **(b, e, h)**, and SPI-12 **(c, f, i)**. For each SPI timescale the evaluation consist of three panels displaying a specific score as a function of lead time (horizontal axis) for a specific 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 coefficient; **(g–i)** area under the ROC curve for SPI forecasts below −0.8. The error bars in all panels denote the 95 % confidence intervals computed from 1000-sample bootstrapping with resampling.

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 to ERAI. This is again due to the use of GPCC as a reference data set that penalized the scores of the forecasts using ERAI for monitoring.

A test of the importance of the monitoring data set upon forecast skill is provided by identifying the last forecast lead time where the ROC skill score of GPCC S4 (using ERAI S4 as a 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 interval estimation, a higher threshold of 0.05 was selected. This approach is useful for highlighting and revealing those regions where the selection of ERAI for monitoring has a stronger detrimental effect on skill (relative to GPCC) of the seasonal forecasts. To quantify the lead time memory of the initial conditions, GPCC S4 was compared with GPCC_CLM S4 (Fig. 3), and it was 1 and 6 or more months for SPI-3 and SPI-12 respectively. For the SPI-6 the memory of the initial conditions varied between 3 and 4 months. The main difference of ranking GPCC S4 with ERAI S4 or GPCC_CLM S4 is for the SPI-6 within the tropics. This shows that a higher disagreement is found among precipitation data sets within the tropics due to the low density of the number of observations.

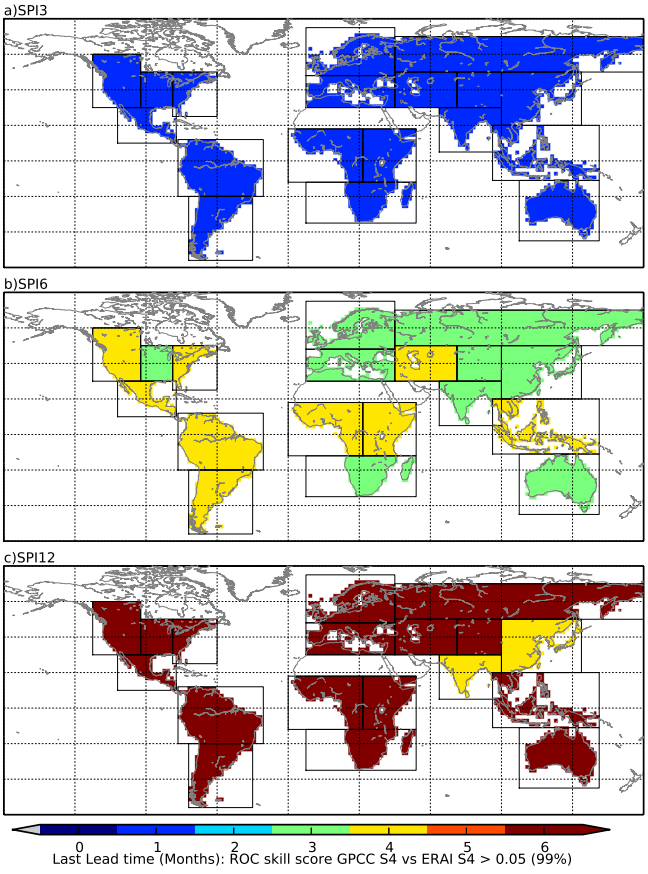

Figure 2. Last forecast lead time (months) where the ROC skill score of GPCC S4 (using ERAI S4 as reference forecasts) is higher than 0.05 with 95 % confidence and the ROC of GPCC S4 is higher than 0 with 95 % confidence. Seasonal forecasts of the **(a)** SPI-3, **(b)** SPI-6, and **(c)** SPI-12. The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

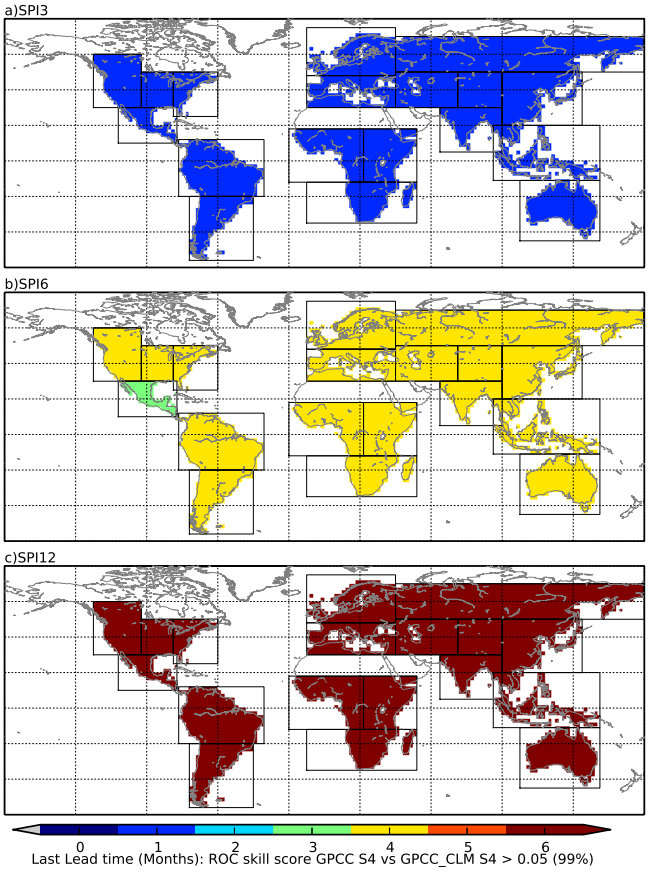

Figure 3. As Fig. 2 but using GPCC_CLM S4 as reference forecasts.

As opposed to testing the importance of monitored precipitation data quality on forecast skill, a test of the importance of forecast information (predicted precipitation) upon forecast skill 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. 4), 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. That is, the seasonal forecast adds 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 where northern Eurasia regions and Australia have the lower values. For SPI-12, the added value of using S4 can reach 5 months of lead time as in the Mediterranean, South Africa and southern South America, while there is significant improvement in northern Europe and the North America – regions where 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 values (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, northern Europe and all three North America regions were not statistically significant for SPI12. Even in regions where there is little more added value to the reduction of lead time, GPCC S4 skill is always equal to or higher than climatology. In some cases, particularly for long SPI timescales (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, so that this is the same as the climatology.

3.2 Drought onset

In order to compare our SPI seasonal forecasts with the forecast models using the US National Multi-Model Ensemble drought forecast in YW13, the tests made upon each 1 degree grid cell are combined into global samples with global means of the probability of detection (POD), and global means of

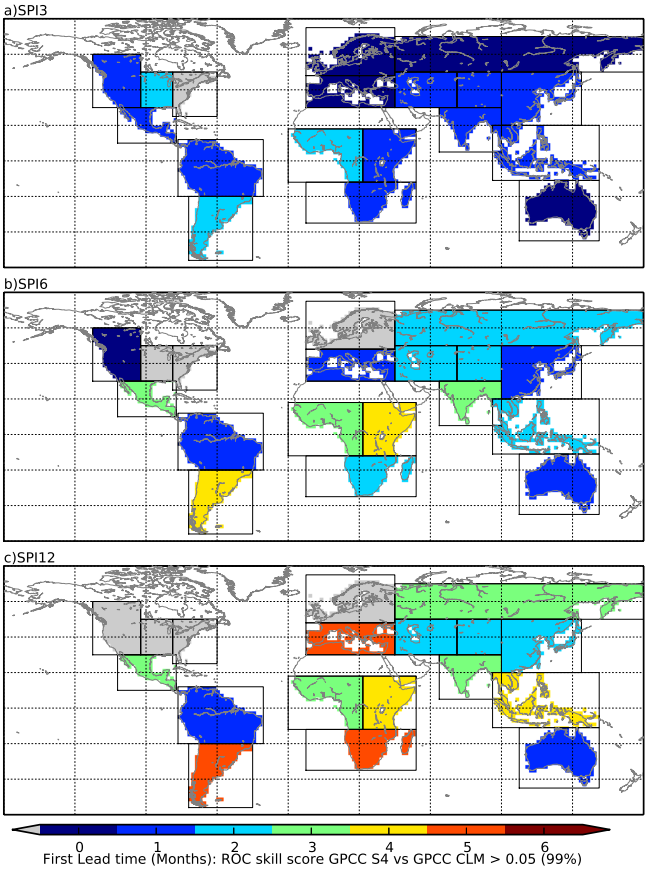

Figure 4. First forecast lead time (months) where the ROC skill score of GPCC S4 (using GPCC CLM as reference forecasts) is higher than 0.05 with 95 % confidence and the ROC of GPCC S4 is higher than 0 with 95 % confidence. Seasonal forecasts of the **(a)** SPI-3, **(b)** SPI-6, and **(c)** SPI-12. The forecasts are verified in each region for the calendar month presented in Table 1.

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 ESP. This study and YW13 deployed different precipitation data sets, as well as time interval of collected hindcasts. 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 data set chosen, and the quality of the precipitation data set, has a major role in the skill of SPI forecasts.

Each ensemble member conserves the SPI characteristics, that of mean zero and standard deviation of one (which arises due to the definition of SPI as a standardized variable). The ensemble mean (of all the ensemble members) conserves the

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 brackets were calculated after scaling the ensemble mean. The 95 % confidence intervals, estimated from 1000-sample bootstrapping with replacement, returned similar values for all scores and models of approximately ±0.01. The ESP and NMME2 model scores are included in this table for comparison purposes only and were retrieved 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

mean of zero, whereas the standard deviation falls below one due to the aggregation procedure. The standard deviation decline below one is pronounced for long lead times as the ensemble spread increases. Despite the change in standard deviation, the drought onset forecast 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 an ensemble mean rescaling case is presented between the brackets in Table 2. This rescaling increases the probability of drought detection (as it should) but, in exchange 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, which was explored by Mwangi et al. (2014) 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 (Wilks, 2006) was used, based upon the climatological frequency of drought events as a reference, for the different experiments over each grid cell of the globe (Fig. 5). 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

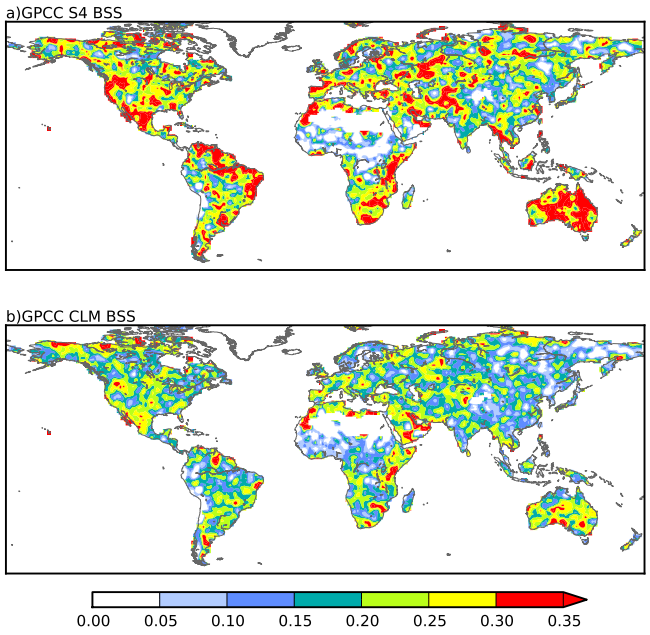

Figure 5. Brier skill scores for the drought onset forecasts of **(a)** GPCC S4 and **(b)** GPCC CLM. The reference forecast for the skill score was the climatological frequency of the verification data set. The original maps at $1^\circ \times 1^\circ$ were smoothed with a 3×3 window.

POD equivalent case and the ESP climatology equivalent case. Our results support the clear benefit of a seasonal forecast over climatology, this being valid for our case of GPCC S4 when compared with GPCC CLM. Seasonal forecasts were better than climatology in Australia, East Africa, northwest South America (Brazil), as well as other regions of the globe, which also corroborates the results in YW13. Looking at the global Brier score decomposition (Fig. S20 in the Supplement) shows that climatology (i.e. GPCC CLM) 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) leads to better Brier scores in GPCC S4. Figure 5 highlights how noisy the individual grid cell scores are globally. Assembling the grid cells into regions, on the other hand, increases the sample sizes within those regions and permits 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. Continuing with the examples in the companion paper Part 1, the system is tested by producing SPI forecasts for the 2010/11 drought in the Horn of Africa (Figs. S21 and S22 in the Supplement) and the 2012 drought in the US Great Plains (Figs. S23 and S24 in the 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 in the Supplement) show the GPCC S4 and GPCC CLM SPI forecast issue on different initial dates and averaged over a region and overlaid with the verification. The spatial maps (Fig. S22 and S24 in the Supplement) compare the actual verification SPI with four different examples of displaying a specific forecast: (i) ensemble mean, (ii) the ensemble mean rescaled (see previous paragraph), (iii) probability of the SPI > 0.8 (wet conditions), and (iv) probability of the SPI < -0.8 (dry conditions).

4 Conclusions and outlook

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 skill of the forecasts is targeted to verification months where precipitation deficits are likely to have higher drought impacts, as well as 18 regions. Detailed analysis of drought forecasting skill within each region is outside the scope of this paper, but all the results are made available in the Supplement. In the course of the study, several comparisons were made between forecast skill and drought frequency on a global scale, but none returned informative results. Further investigations could be performed by following a similar approach to YW13 by conditioning the analysis on El Niño/La Niña events and restricting the comparison to particular regions.

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

The memory effect of initial conditions in the SPI forecasts has been identified, comparing the S4 seasonal forecasts initialized with GPCC to the same S4 seasonal forecasts initialized and merged with GPCC climatological precipitation. This was found to be 1 month of lead time in the case of SPI-3, 3–4 months for the SPI-6 case, and 6 (or more) months for the SPI-12 case. For earlier forecast lead times, the initial conditions of precipitation dominate the forecast skill, proving that good quality and reliable monitoring of precipitation is of paramount importance

The second question was the added value of using ECMWF seasonal forecasts of precipitation when in comparison with climatological sampling. Even in regions where the added value in terms of forecast lead time is reduced, our results show that the skill of dynamical forecasts is always equal to or above to climatological forecasts. In some cases, particularly for long SPI timescales, it is very difficult to improve the climatological forecasts. For long SPI timescales (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, so, in the limit, it is like testing the monitored precipitation against itself.

Finally we posed the following question: what is the skill of dynamical forecast in terms of drought onset? The definition of drought onset followed that of YW13 in order to be able to compare our results against the drought forecasts from other forecasting ensemble models within the US National Multi-Model Ensemble in the YW13 study. Although different data sets and periods were used, the estimates of drought onset skill for climatological forecasts are similar, and therefore we suggest they are reasonably independent of data and intrinsic to the SPI seasonal forecasting methodology. We recommend that when evaluating only the forecast's ensemble 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. YW13 raised the question as to whether seasonal forecasting of global drought onset was largely or solely a stochastic forecasting problem. Our results are coherent with their findings, but our regional analysis highlights that within several regions in the world drought onset forecasting is feasible and skilful.

The Supplement related to this article is available online at doi:10.5194/hess-18-2669-2014-supplement.

Acknowledgements. This work was funded by the European Commission Seventh Framework Programme (EU FP7) in the framework of the Improved Drought Early Warning and Forecasting to Strengthen Preparedness and Adaptation to Droughts in Africa (DEWFORA) project under grant agreement 265454 (http://www.dewfora.net).

Edited by: M. Werner

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. Meteor. 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.

Dutra, E., Wetterhall, F., Di Giuseppe, F., Naumann, G., Barbosa, P., Vogt, J., Pozzi, W., and Pappenberger, F.: Global meteorological drought – Part 1: Probabilistic monitoring, Hydrol. Earth Syst. Sci., 18, 2657–2667, doi:10.5194/hess-18-2657-2014, 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., 18, 611-620, doi:10.5194/hess-18-611-2014, 2014.

Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., and Smith, L.: Ensemble prediction: A pedagogical prespective, ECMWF Newsletter, 106, 10–17, 2006.

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], 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.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences., International Geophysics Series, Academic Press, 648 pp., 2006.

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.