https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

Hydrology and Earth System Sciences
Discussions

Global and regional performances of SPI candidate distribution functions in observations and simulations

Patrick Pieper[1], André Düsterhus[2], and Johanna Baehr[1]

[1]Institute for Oceanography, Center for Earth System Research and Sustainability, Universität Hamburg, Hamburg, Germany
[2]ICARUS, Department of Geography, Maynooth University, Maynooth, Ireland

Correspondence: Patrick Pieper (Patrick.Pieper@uni-hamburg.de)

Abstract. The Standardized Precipitation Index (SPI) is a widely accepted drought index. Its calculation algorithm normalizes the index via a distribution function. Which distribution function to use is still disputed within literature. This study illuminates the long-standing dispute and proposes a solution which ensures the normality of the index for all common accumulation periods in observations and simulations.

We compare the normality of SPI time-series derived with the gamma, Weibull, generalized gamma, and the exponentiated Weibull distribution. Our normality comparison evaluates actual against theoretical occurrence probabilities of SPI categories, and the quality of the fit of candidate distribution functions against their complexity with Akaike's Information Criterion. SPI time-series, spanning 1983–2013, are calculated from Global Precipitation Climatology Project's monthly precipitation dataset and seasonal precipitation hindcasts from the Max Planck Institute Earth System Model. We evaluate these SPI time-series over the global land area and for each continent individually during winter and summer. While focusing on an accumulation period of 3-months, we additionally test the drawn conclusions for other common accumulation periods (1-, 6-, 9-, and 12-months).

Our results suggest to exercise caution when using the gamma distribution to calculate SPI; especially in simulations or their evaluation. Further, our analysis shows a distinctly improved normality for SPI time-series derived with the exponentiated Weibull distribution relative to other distributions. The use of the exponentiated Weibull distribution maximizes the normality of SPI time-series in observations and simulations both individual as well as concurrent. Its use further maximizes the normality of SPI time-series over each continent and for every investigated accumulation period. We, therefore, advocate to derive SPI with the exponentiated Weibull distribution, irrespective of the heritage of the precipitation data or the length of analyzed accumulation periods.

1 Introduction

Drought intensity, onset, and duration are commonly assessed with the Standardized Precipitation Index (SPI). SPI was first introduced by McKee et al. (1993) as a temporally and spatially invariant probability-based drought index. In 2011, the World Meteorological Organization (WMO) endorsed the index and recommended its use to all meteorological and hydrological services for classifying droughts (Hayes et al., 2011). Advantages of SPI are its standardization (Sienz et al., 2012), its simplicity,

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

Hydrology and Earth System Sciences
Discussions

and its variable time scale which allows its application to assess meteorological, agricultural, and hydrological drought (Lloyd-Hughes and Saunders, 2002). In contrast, the index's main disadvantage is the mean by which its standardization is realized and concerns the identification of a suitable theoretical distribution function to describe and normalize highly non-normal precipitation distributions (Lloyd-Hughes and Saunders, 2002). The choice of that suitable theoretical distribution function is a key decision in the index's algorithm (Blain et al., 2018; Stagge et al., 2015; Sienz et al., 2012). This study illuminates reasons for a missing consensus on this choice and attempts to establish such a consensus for both simulations and observations.

SPI quantifies the standardized deficit (or surplus) of precipitation over any period of interest – also called accumulation period. This is achieved by fitting a probability density function (PDF) to the frequency distribution of precipitation totals of the accumulation period – which typically spans either 1-, 3-, 6-, or 12-months. SPI is then generated by applying a Z-transformation to the probabilities and is standard normal distributed.

The choice of the PDF fitted to the frequency distribution of precipitation is essential because only a proper fit appropriately standardizes the index. While the standardization simplifies further analysis of the SPI, the missing physical understanding of the distribution of precipitation leads to a questionable basis for the fit. Therefore, the choice of the PDF is to some extent arbitrary and depicts the Achilles heel of the index.

Originally, McKee et al. (1993) proposed a simple gamma distribution – while Guttman (1999) identified the Pearson Type III distribution – to best describe observed precipitation. Both of these distributions are nowadays mostly used in SPI's calculation algorithms. As a result, many studies that use SPI directly fit the gamma (Mo and Lyon, 2015; Ma et al., 2015; Yuan and Wood, 2013; Quan et al., 2012; Yoon et al., 2012) or the Pearson type III distribution (Ribeiro and Pires, 2016) without assessing the normality of SPI's resulting distribution with goodness-of-fit tests or other statistical analyses beforehand. The selected PDF, however, is of critical importance because the choice of this PDF is the key decision involved in the calculation of SPI and indeed many authors have urged to investigate the adequacy of distribution functions for new data-sets and regions before applying them (Blain et al., 2018; Stagge et al., 2015; Touma et al., 2015; Sienz et al., 2012). Such a negligence has potentially far-reaching consequences in terms of a biased drought description (Guenang et al., 2019; Sienz et al., 2012). A biased drought description would result from an inadequacy of the fitted distribution function to describe precipitation. Such an inadequacy has been identified for the gamma (Guenang et al., 2019; Blain et al., 2018; Blain and Meschiatti, 2015; Stagge et al., 2015; Sienz et al., 2012; Touma et al., 2015; Naresh Kumar et al., 2009; Lloyd-Hughes and Saunders, 2002) as well as the Pearson type III distribution (Blain et al., 2018; Blain and Meschiatti, 2015; Stagge et al., 2015) in many parts of the world. This lead to the request for further investigations of candidate distribution functions (Blain et al., 2018; Blain and Meschiatti, 2015; Stagge et al., 2015; Touma et al., 2015; Sienz et al., 2012; Lloyd-Hughes and Saunders, 2002; Guttman, 1999).

Several studies have investigated the adequacy of PDFs fitted onto observed precipitation while focusing on different candidate distribution functions (Blain and Meschiatti, 2015), different parameter estimation methods in the fitting procedure (Blain et al., 2018), different SPI time scales (Guenang et al., 2019), general drought climatology (Lloyd-Hughes and Saunders, 2002), and even the most appropriate methodology to test different candidate distribution functions (Stagge et al., 2015). As each of these investigations analyzed different regions, different PDFs or focused on different perspectives of this highly multi-dimensional problem, they recommend different candidate PDF.

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

Nevertheless, some common conclusions can be drawn. Most investigations only analyzed 2-parameter distribution functions (Guenang et al., 2019; Blain et al., 2018; Stagge et al., 2015; Lloyd-Hughes and Saunders, 2002). Among those, they agreed depending on the accumulation period and/or the location either on the Weibull or the gamma distribution to be best suited in most cases. However, Blain and Meschiatti (2015) also investigated 3-, 4- and 5-parameter distribution functions and concluded that 3-parameter PDFs seem to be best suited to compute SPI in Pelotas, Brazil. Consequently, they advocated for a re-evaluation of the widespread use of the 2-parameter gamma distribution (see also Wu et al., 2007). Moreover, a single candidate distribution function was neither suited in each location nor for each accumulation period to properly calculate SPI time series (Guenang et al., 2019; Blain et al., 2018; Stagge et al., 2015; Lloyd-Hughes and Saunders, 2002). Further, at the accumulation period of 3-months, a critical phase transition in precipitation totals seem to manifest which complicates the overall ranking of candidate PDFs (Guenang et al., 2019; Blain et al., 2018; Stagge et al., 2015). Findings point at the Weibull distribution to be best suited for short accumulation periods (smaller than 3 months) and the gamma distribution for long accumulation periods (larger than 3 months) (Stagge et al., 2015).

Two additional studies analyzed the adequacy of different candidate PDFs fitted onto simulated precipitation while focusing on drought occurrence probabilities in climate projections (Touma et al., 2015; Sienz et al., 2012). Touma et al. (2015) is the only study which tested candidate PDFs globally. However, they solely provided highly aggregated results which are globally averaged for accumulation periods between 3- and 12-months and concluded that the gamma distribution is overall best suited to calculate SPI. In contrast, Sienz et al. (2012) is up to now the only study which tested candidate PDFs in simulations as well as in observations and identified notable differences in their performance in both realizations. They focused on an accumulation period of 1-month and their results also show that the Weibull distribution is well suited for SPI calculations at short accumulation periods in observations but also in simulations. Moreover, their results also hint at the phase transition mentioned above: for accumulation periods longer than 3 months their results indicate that the gamma distribution outperforms the Weibull distribution in observation as well as in simulations. More interestingly, Sienz et al. (2012) results indicate that two 3-parameter distributions (the generalized gamma and the exponentiated Weibull distribution) perform for short accumulation periods as well as the Weibull distribution and for long accumulation periods similar to the gamma distribution; in observations and simulations. Surprisingly, neither the exponentiated Weibull nor the generalized gamma distribution have been thoroughly tested since.

Testing the performance of 3-parameter distributions introduces the risk of overfitting (Stagge et al., 2015; Sienz et al., 2012) which could explain the focus on 2-parameter distributions in recent studies. As a consequence of this focus in combination with the inability of 2-parameter PDFs to perform sufficiently well in different locations and for different accumulation periods concurrently, many studies have proposed a multi-distribution approach (Guenang et al., 2019; Blain and Meschiatti, 2015; Touma et al., 2015; Sienz et al., 2012; Lloyd-Hughes and Saunders, 2002). Such an approach recommends the use of the best-suited PDF for each accumulation period and in each location. In opposition, other studies have strongly emphasized concern about this approach, because it adds complexity while reducing or even obliterating comparability across space and time (Stagge et al., 2015; Guttman, 1999). The comparability across space and time is a main advantage of SPI. Guttman (1999) even warns of using SPI widely until a single PDF is commonly accepted and established as the norm.

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

Most studies test candidate distribution functions with goodness-of-fit tests (Guenang et al., 2019; Blain et al., 2018; Blain and Meschiatti, 2015; Stagge et al., 2015; Touma et al., 2015; Lloyd-Hughes and Saunders, 2002). In this process, some studies heavily rely on the Kolmogorov-Smirnov test (Guenang et al., 2019; Touma et al., 2015). However, the Kolmogorov-Smirnov test has an unacceptably high likelihood of erroneously accepting a non-normal distribution if the parameters of the candidate PDF have been estimated from the same data on which the tested distribution bases (which is in view of scarce precipitation data availability usually always the case) (Blain et al., 2018; Blain and Meschiatti, 2015; Stagge et al., 2015). Therefore, other studies tested the goodness-of-fit either with an adaptation of the Kolmogorov-Smirnov test, the Lillieforts test (Blain et al., 2018; Blain and Meschiatti, 2015; Stagge et al., 2015; Lloyd-Hughes and Saunders, 2002), with the Anderson-Darling test (Blain et al., 2018; Stagge et al., 2015) or with the Shapiro-Wilk test (Blain et al., 2018; Blain and Meschiatti, 2015; Stagge et al., 2015). Nevertheless, the Lillieforts and Anderson-Darling tests are inferior to the Shapiro-Wilk test (Blain et al., 2018; Stagge et al., 2015) which in turn is unreliable to evaluate SPI normality (Naresh Kumar et al., 2009).

Additionally, all three of these goodness-of-fit tests are unable to produce any relative ranking of the performance of distribution functions for a specific location and accumulation period. In consequence, they are ill-suited to discriminate the best performing PDF out of a set of PDFs (Blain et al., 2018), because they are designed to deliver a binary answer. For SPI distributions, however, the question is not whether they are (or should be) normally distributed (for which goodness-of-fit tests are well suited to provide the answer). The crucial question is rather which PDF maximizes the normality of the resulting SPI distribution. As a result, those studies that rigorously analyzed candidate distribution functions or investigate an appropriate test methodology for evaluating SPI candidate PDFs advocate the use of relative assessments: mean absolute errors (Blain et al., 2018), Akaike's Information Criterion (AIC) (Stagge et al., 2015; Sienz et al., 2012), or deviations from expected SPI categories (Sienz et al., 2012). These studies also emphasize the importance of quantifying the differences between theoretical and calculated SPI values for different drought categories (Blain et al., 2018; Sienz et al., 2012). Stagge et al. (2015) who investigated appropriate methodologies to test different candidate PDFs even used AIC to discriminate the performance of different goodness-of-fit tests.

In this study, we test the adequacy of the gamma, Weibull, generalized gamma, and exponentiated Weibull distribution in SPI's calculation algorithms. The evaluation of their performance depends on the normality of the resulting SPI time-series. In this evaluation, we focus on an SPI accumulation period of 3-months (SPI_{3M}) during winter (DJF) and summer (JJA) and test the drawn conclusions for other common accumulation periods (1-, 6-, 9-, and 12-months). Our analysis conducts two different evaluations of their normality: (i) it compares actual occurrence probabilities of SPI categories (as defined by WMO's *SPI User Guide* (Svoboda et al., 2012)) against well-known theoretically expected occurrence probabilities from the standard normal distribution ($\mathcal{N}_{0,1}$), (ii) it analytically assesses with the Akaike's Information Criterion (AIC) the *optimal trade-off* between information gain against the complexity of the PDF to adhere to the risk of overfitting. During this analysis, we investigate observations and simulations, the latter are usually neglected but demand nowadays a similarly prominent role as observations because of the increasing importance of drought predictions and their evaluation. Despite this importance, the adequacy of different candidate distribution functions has to the authors' best knowledge never been tested in the output of a seasonal prediction system – although seasonal predictions constitute our most powerful tool to predict individual droughts. To

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

close that gap, this study evaluates the performance of candidate distribution functions in an output of 10 ensemble members of initialized seasonal hindcast simulations. The monthly precipitation data-set of the Global Precipitation Climatology Project (GPCP) serves as an observational product. We conduct our analysis for the period 1982 to 2013 with a global focus which also highlights regional disparities on every inhabited continent (Africa, Asia, Australia, Europe, North America, and South America).

2 Methods

2.1 Model and Data

We employ a seasonal prediction system (Baehr et al., 2015; Bunzel et al., 2018) which bases on the Max-Planck-Institute Earth System Model (MPI-ESM). MPI-ESM, also used in the Coupled Model Intercomparison Project 5 (CMIP5), consists of an atmospheric (ECHAM6) (Stevens et al., 2013), and an oceanic (MPIOM) (Jungclaus et al., 2013) component. For this study the model is initialized in May and November and runs with 10 ensemble members in the low-resolution version – MPI-ESM-LR: T63 (approx. 1.875°x1.875°) with 47 different vertical layers in the atmosphere between the surface and 0.01 hPa and 40 different vertical layers in the ocean. Except for an extension of the simulation period by 3 years (extended to cover the period 1982–2013), the investigated simulations are identical to the 10-member ensemble simulations analyzed by Bunzel et al. (2018).

We obtain observed precipitation from the Global Precipitation Climatology Project (GPCP) which combines observations and satellite precipitation data into a monthly precipitation data-set on a 2.5°x2.5°global grid spanning 1979 to present (Adler et al., 2003). To compare these observations against our hindcasts, the precipitation output of the model is interpolated to the same grid as GPCP's precipitation data-set from which we only use the simulated time-period (1982–2013).

Depending on the accumulation period (1-, 3-, 6-, 9-, or 12-months) we calculate the frequency distribution of modeled and observed precipitation totals over 2 different seasons (August and February (1), JJA and DJF (3), MAMJJA and SONDJF (6), and so on). Because our results do not indicate major season-dependent differences in the performance of candidate PDFs for SPI_{3M}, we aggregate our results for the other accumulation periods over both seasons.

Our precipitation hindcasts are neither bias- nor drift-corrected and also not recalibrated. Such corrections usually adjust the frequency distribution of modeled precipitation in each grid-point to agree better with the observed frequency distribution. Here, we investigate the adequacy of different PDFs in describing the frequency distribution of modeled precipitation totals over each accumulation period without any correction. As a consequence, we require that SPI's calculation algorithm deals with such differing frequency distributions on its own. That requirement enables us to identify the worst possible miss-matches.

2.2 Standardized Precipitation Index

We calculate SPI (McKee et al., 1993) for our observed and modeled time-period by fitting a PDF onto sorted 3-months precipitation totals in each grid-point during both seasons of interest and for each accumulation period. Zero-precipitation

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

events are excluded from the precipitation time-series before fitting the PDF and dealt with specifically later. We optimize the parameters of our candidate PDFs in SPI's calculation algorithm with the maximum likelihood method which is also the basis for the AIC computation. Our parameter estimation method first identifies starting values for the n parameters of the candidate PDFs by roughly scanning the n-dimensional phase-space spanned by these parameters. Those starting values are then optimized (Nocedal and Wright, 1999) by three different methods: (i) a simulated annealing method (Bélisle, 1992), (ii) a limited-memory modification of the Broyden-Fletcher-Goldfarb-Shanno (also known as BFGS) quasi-Newton method (Byrd et al., 1995), and (iii) the Nelder and Mead (1965) method. After checking the convergence of the most suitable parameters of our candidate PDFs and omitting cases where convergence is not achieved, the probabilities of encountering the given precipitation totals are computed and transformed to cumulative probabilities ($G(x)$).

Since PDFs which describe the frequency distribution of precipitation totals are required to be only defined for the positive real axis, that cumulative probability ($G(x)$) is undefined for $x = 0$. Nevertheless, the time-series of precipitation totals may contain events in which zero precipitation has occurred over the entire accumulation period. Therefore the cumulative probability is adjusted:

$$H(x) = q + (1 - q)G(x) \tag{1}$$

where q is the occurrence probability of zero-precipitation events in the time-series of precipitation totals. q is estimated by the fraction of the omitted zero-precipitation events in our time-series. Next, we calculate from the new cumulative probability ($H(x)$) the likelihood of encountering each precipitation event of our time-series for every grid-point in each season of interest and each accumulation period. In the final step, analog to McKee et al. (1993), a Z-transformation of that likelihood to the standard normal (mean=0, variance=1) variable Z takes place which constitutes the time-series of SPI.

In very arid regions or those with a distinct dry season, SPI time-series are characterized by a lower bound (Pietzsch and Bissolli, 2011; Wu et al., 2007). That lower bound results from $H(x)$ dependence on q and correctly ensures that short periods without rain do not necessarily constitute a drought in these regions. Nevertheless, that lower bound also leads to non-normal distributions of SPI time-series. The shorter the accumulation period, the more likely it is for zero-precipitation events to occur – and the more likely it becomes for SPI time-series to be non-normally distributed. Stagge et al. (2015) proposed to use the *centre of mass* instead of the fraction of zero-precipitation events to estimate q. Such an adaptation leads to a lower q than the fraction-approach which distinctly increases the normality of SPI time-series and their statistical interpretability if that fraction becomes larger than approximately one third. As explained before, we want to investigate the worst possible case and, therefore, conservatively estimate q. As a consequence, SPI time-series are calculated exclusively for grid-points exhibiting zero-precipitation events in less than 34 % of the times in our time-period. This limitation restricts the SPI calculation over the Sahara for accumulation periods of 1- and 3-months, only exceptionally occurs for an accumulation period of 6-months, and does not restrict accumulation periods longer than 6-months. Current complex climate models parameterize convection and cloud micro-physics to simulate precipitation which leads to spurious precipitation amounts. Those spurious precipitation amounts prevent us from directly identifying the probability of zero-precipitation events in modeled precipitation time-series.

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Analog to Sienz et al. (2012), we prescribe a threshold of 0.035 mm month^{-1} to differentiate between months with and without precipitation in the hindcasts.

To further optimize the fit of the PDF onto modeled precipitation, all hindcast ensemble members are fitted at once – assuming that all ensemble members show in the long-term identical frequency distributions of precipitation in the same grid-point. It is, therefore, reasonable to presume that a better fit is achievable for simulated rather than for observed precipitation.

2.3 Candidate Distribution Functions

Cumulative precipitation sums are described by skewed distribution functions which are only defined for the positive real axis. We test four different distribution functions and evaluate their performance based on the normality of their resulting SPI frequency distributions. The four candidate PDFs either consist of a single shape (σ) and scale (γ) parameter or include (in the case of the two 3-parameter distributions) a second shape parameter (α). Figure 1 displays examples of those four candidate PDFs and their 95 % quantiles for 3-months precipitation totals idealized to be distributed according to the respective distribution function with $\sigma = \gamma = (\alpha) = 2$. Table 1 lists the abbreviations used for the four candidate distribution functions.

Instead of investigating the Pearson Type III distribution, which is already widely used, we analyze the simple gamma distribution. They differ by an additional location parameter which does not change the here presented results (Sienz et al., 2012). Moreover, other studies have demonstrated that the Pearson type III distribution delivers results which are virtually identical to the 2-parameter gamma distribution (Pearson's r $= 0.999$) (Giddings et al., 2005) and argued that the inclusion of a location parameter unnecessarily complicates the SPI algorithm (Stagge et al., 2015). Therefore, our 3-parameter candidate PDFs comprise a second shape parameter instead.

1. Gamma distribution

$$f(x) = \frac{1}{\sigma\Gamma(\gamma)}\left(\frac{x}{\sigma}\right)^{\gamma-1}\exp(-\frac{x}{\sigma}) \tag{2}$$

The gamma distribution (Γ being the gamma-function) is typically used for SPI calculations directly or in its location parameter extended version: the Pearson Type III distribution (Guttman, 1999). The results of the gamma distribution also serve as proxy for the performance of the Pearson Type III distribution.

2. Weibull distribution

$$f(x) = \frac{\gamma}{\sigma}\left(\frac{x}{\sigma}\right)^{\gamma-1}\exp(-\left(\frac{x}{\sigma}\right)^{\gamma}) \tag{3}$$

The Weibull distribution is usually used to characterize wind speed. Several studies identified the Weibull distribution, however, to perform well in SPI's calculation algorithm for short accumulation periods (Guenang et al., 2019; Blain et al., 2018; Stagge et al., 2015; Sienz et al., 2012).

3. Generalized gamma distribution

$$f(x) = \frac{\alpha}{\sigma\Gamma(\gamma)}\left(\frac{x}{\sigma}\right)^{\alpha\gamma-1}\exp(-\left(\frac{x}{\sigma}\right)^{\alpha}) \tag{4}$$

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

The generalized gamma distribution extends the gamma distribution by another shape-parameter (α). In the special case of $\alpha = 1$ the generalized gamma distribution becomes the gamma distribution and for the other special case of $\gamma = 1$ the generalized gamma distribution becomes the Weibull distribution. Sienz et al. (2012) identified the generalized gamma distribution as promising candidate distribution function for SPI's calculation algorithm.

4. Exponentiated Weibull distribution

$$f(x) = \frac{\alpha\gamma}{\sigma}\left(\frac{x}{\sigma}\right)^{\gamma-1}\left[1-\exp(-\left(\frac{x}{\sigma}\right)^{\gamma})\right]^{\alpha-1} \tag{5}$$

The exponentiated Weibull distribution extends the Weibull distribution by a second shape parameter (α). For $\alpha = 1$ the exponentiated Weibull distribution becomes the Weibull distribution. Sienz et al. (2012) revealed that the exponentiated Weibull distribution performs well in SPI's calculation algorithm.

2.4 Deviations from the Standard Normal Distribution

SPI time-series are supposed to be standard normally distributed ($\mu = 0$ and $\sigma = 1$). Thus, we evaluate the performance of each candidate distribution function (in describing precipitation totals) based on the normality of their resulting SPI frequency distributions. In this analysis, we calculate actual occurrence probabilities for certain ranges of events in our SPI frequency distributions and compare those actual against well-known theoretical occurrence probabilities for the same range of events. We then evaluate the performance of each candidate distribution function and their resulting SPI time-series based on the magnitude of deviations from the standard normal distribution ($\mathcal{N}_{0,1}$). These deviations are henceforth referred to as deviations from $\mathcal{N}_{0,1}$.

According to WMO's *SPI User Guide* (Svoboda et al., 2012) (see Table 2), SPI distinguishes between seven different SPI categories. These seven different categories with their pre-defined SPI intervals serve as analyzed ranges of possible events in our analysis. It is noteworthy here, that these seven SPI categories differ in their occurrence probabilities. The occurrence of normal conditions (N0) is more than twice as likely than all other six conditions put together. Therefore, any strict normality analysis of SPI time-series would weight each classes' identified deviation from $\mathcal{N}_{0,1}$ with the occurrence probability of the respective class. However, when analyzing droughts with SPI, one is usually interested in extreme precipitation events. Thus, it seems less important for the center of SPI's distribution to be normally distributed. Instead, it is intuitively particularly important for the tails (especially the left-hand tail) of the distribution to adhere to the normal-distribution. The better the tails of our candidate PDF's SPI distributions agree with $\mathcal{N}_{0,1}$, the better is our candidate PDF's theoretical description of extreme precipitation events. For this reason, we treat all seven SPI categories equally, irrespective of their theoretical occurrence probability.

The 3-parameter candidate distribution functions contain the 2-parameter candidate distribution functions for special cases. Given those special cases, the 3-parameter candidate distribution functions will in theory never be inferior to the 2-parameter candidate distribution functions they contain when analyzing deviations from $\mathcal{N}_{0,1}$ – assuming a sufficient quantity of input data which would lead to a sufficient quality of our fit. Thus, the question is rather whether deviations from $\mathcal{N}_{0,1}$ reduce enough to justify the 3-parameter candidate distribution functions' requirement of an additional parameter. An additional parameter

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

which needs to be fitted increases the risk of overfitting (Stagge et al., 2015; Sienz et al., 2012). On the one hand, the final decision on this trade-off might be subjective and influenced by computational resources available or by the length of the time-series which is to be analyzed because fitting more parameters requires more information. Moreover, it might well be wiser to employ scarce computational resources in optimizing the fit rather than increasing the complexity of the PDF. On the other hand, assuming computational resources and data availability to be of minor concern, there exists an analytical way to tackle this trade-off: Akaike's Information Criterion (Akaike, 1974).

2.5 Akaike's Information Criterion

Our aim is twofold. First, we want to maximize the normality of our SPI time-series by choosing an appropriate distribution function. Second, we simultaneously aspire to minimize the parameter-count of the distribution function to avoid unnecessary complexity which decreases the risk of overfitting. The objective is to identify the necessary (minimal) complexity of the PDF which prevents the PDF from being too simple and lose explanatory power. Or in other words: we are interested in the so-called *optimal trade-off* between bias (model too simple) and variance (model too complex). Akaike's information criterion (AIC) performs this trade-off analytically (Akaike, 1974). AIC estimates the value of information gain (acquiring an improved fit) and penalizes complexity (the parameter count) directly by estimating the Kullback-Leibler information (Kullback and Leibler, 1951):

$$AIC = -2\ln\mathcal{L}(\hat{\theta}|y) + 2k \tag{6}$$

$\mathcal{L}(\hat{\theta}|y)$ describes the likelihood of specific model-parameters ($\hat{\theta}$) with given data from which these parameters were estimated (y). k describes the degrees of freedom of the candidate PDF (the parameter-count which equates dependent on the candidate PDF either to 2 or 3). Analogue to Burnham and Anderson (2002), we modified the last term from $2k$ to $2k+(2k(k+1))/(n-k-1)$ in order to improve the AIC calculation for small sample sizes ($n/k < 40$), whereas in our case n corresponds to the sample size of the examined period (31 for observations and 310 for simulations). The modified version approaches the standard version for large n.

In our case, AIC's first term evaluates the performance of candidate PDFs in describing the given frequency distributions of precipitation totals. The second term penalizes candidate PDFs based on their parameter-count. The best performing distribution function attains a minimum AIC value (AIC_{min}) because the first term is negative and the second one is positive.

Further, the absolute AIC value is often of little information – especially in contrast to relative differences between AIC values derived from different distribution functions (henceforth we index different distribution functions with an i and name the corresponding AIC values AIC_i accordingly). These relative differences inform us about superiority in the optimal trade-off between bias and variance. Thus, we use AIC differences (AIC-D) in our further assessment:

$$AIC\text{-}D_i = AIC_i - AIC_{min} \tag{7}$$

For our analysis, AIC-D are well suited to compare and rank different candidate PDFs based on their trade-off between bias and variance. The best performing distribution function is characterized by a minimum AIC value (AIC_{min}) which translates

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

to an AIC-D value of 0. It seems noteworthy here that any evaluation of (or even any discrimination between) candidate distribution functions which exhibit a sufficiently small AIC-D is unfeasible as a consequence of our rather small sample size (particularly in observations, but also in simulations). AIC-D values below two should be in general interpreted as an indicator of substantial confidence in the performance of the model (here, the PDF). In contrast, AIC-D values between four and seven indicate considerable less confidence and values beyond ten essentially none (Burnham and Anderson, 2002).

2.6 Aggregation of Results over Domains

For each candidate distribution function, accumulation period, domain, and during both seasons, we compute deviations from $\mathcal{N}_{0,1}$ separately for observations and simulations as schematically depicted on the left-hand side in Fig. 2. First, we count the events of each SPI category in every land grid-point globally. For each category, we then sum the category counts over all grid-points which belong to the domain of interest. Next, we calculate actual occurrence probabilities through dividing that sum by the sum over the counts of all seven SPI categories (per grid-point there are 31 total events in observations and 310 in simulations). In a final step, we compute the difference to theoretical occurrence probabilities of $\mathcal{N}_{0,1}$ (provided in Table 2) for each SPI category and normalize that difference – expressing the deviation from $\mathcal{N}_{0,1}$ as percent of the theoretically expected occurrence probability.

Again for each candidate distribution function, accumulation period, domain, and both seasons, we aggregate AIC-Ds over several grid-points into a single graph separately for observations and simulations as depicted on the right-hand side of the flow chart in Fig. 2. For each domain, we compute the fraction of total grid-points of that domain for which each candidate PDF displays an AIC-D value equal to or below a specific AIC-D$_{max}$ value. That calculation is iteratively repeated for infinitesimally increasing AIC-D$_{max}$ values. In this representation, the probabilities of all PDFs at the specific AIC-D$_{max}$ value of 0 sum up to 100 % because only one candidate PDF can perform best in each grid-point. Thus, we arrive at a summarized AIC-D presentation in which those candidate distribution functions which approach 100 % the fastest (preferably before the specific AIC-D$_{max}$ value of 4; ideally even before the AIC-D$_{max}$ value of 2) are better suited than the others.

2.7 Regions

We investigate the normality of SPI time-series derived from each candidate PDF first for the entire global land area and analyze subsequently region-specific disparities. For this analysis we focus on the land area over six regions scattered over all six inhabited continents: Africa (0°–30°S; 10°E–40°E), Asia (63°N–31°N; 86°E–141°E), Australia (16°S–38°S; 111°E–153°E), Europe (72°N–36°N; 10°W–50°E), North America (50°N–30°N; 130°W–70°W), and South America (10°N–30°S; 80°W–35°E) (Fig. 3).

Examining frequency distributions of precipitation totals over smaller domains than the entire globe reduces the risk of encountering opposite deviations from $\mathcal{N}_{0,1}$ for the same category which then balance each other in different grid-points. This statement is based on either one of the following two assumptions. First, the sum over less grid-points is less likely to produce deviations which balance each other. Second, the frequency distribution of precipitation totals is likely to be more uniform for grid-points that belong to the same region (and therefore exhibit similar climatic conditions) than when they are scattered

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

around the entire globe. One could continue along this line of reasoning because the smaller the area of the analyzed regions, the more impactful are both of these assumptions. However, comparing actual against theoretically expected occurrence probabilities with a scarce database (31 events in observations) will inevitably produce deviations. In observations, we would expect in each grid-point that 0.7 extremely wet/dry and 1.4 severely wet/dry events occur during 31 years. Thus, deviations in different grid-points need to balance each other to some extent, to statistically evaluate and properly compare candidate PDFs. The crucial performance requirement demands that they balance each other also when averaged over sufficiently small domains with similar climatic conditions.

For a first overview, it is beneficial to cluster as many similar results as possible together to minimize the level of complexity of the regional dimension. The choice of sufficiently large/small domains is still rather subjective. Which size of regions is most appropriate? This subjective nature becomes apparent in studies which identify differing borders for regions which are supposed to exhibit rather uniform climatic conditions (Giorgi and Francisco, 2000; Field et al., 2012). Instead of using *Giorgi-Regions* (Giorgi and Francisco, 2000) or *SREX-Regions* (Field et al., 2012), we opt here for a broader and more continental picture.

3 Results

3.1 SPI Accumulation Period of 3-Month

3.1.1 Global

In agreement with prior studies (Blain et al., 2018; Lloyd-Hughes and Saunders, 2002; McKee et al., 1993), the 2-parameter gamma distribution (GD2) describes on global average the observed frequency distribution of SPI_{3M} rather well during the boreal winter (DJF) and summer (JJA) (Fig. 4, (a)). Contrary to Sienz et al. (2012), who investigated SPI_{1M} time-series, the 2-parameter Weibull distribution (WD2) delivers a poor frequency distribution of SPI_{3M} during both seasons (Fig. 4, (b)). Further, GD2 leads to a better agreement between the frequency distribution of SPI_{3M} time-series and $\mathcal{N}_{0,1}$ than any of the here investigated 3-parameter PDFs over both seasons of interest. Still, GD2 represents the especially important left-hand tail of SPI_{3M} time-series' frequency distribution (D3) in JJA relatively poor. Here, the investigated 3-parameter distributions, GGD3 and the exponentiated Weibull distribution (EWD3), perform better (Fig. 4, (c) and (d)). Despite these minor differences, and in agreement with Sienz et al. (2012), GGD3 and EWD3 perform overall similar to GD2 (compare Fig. 4, (a) against (c) and (d)).

In theory, since the 3-parameter generalized gamma distribution (GGD3) encompasses GD2 as a special case, GGD3 should not be inferior to GD2. In reality, however, the applied optimization methods appear to be too coarse for GGD3 to lead to an identical or better optimum than the one identified for GD2 with the given length of the time-series. When optimizing 3 parameters it is more likely to miss a specific constellation which would further optimize the fit; especially when limited computational resources impede the identification of the actual optimal fitting parameters. Additionally, a limited database (our database spans 31 years) obscures the frequency distribution of precipitation totals which poses another obstacle to the

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

fitting methods. This results in missed optimizations opportunities which impact GGD3 stronger than GD2 because of GGD3's complexity. As a result, the weighted sum over the absolute values of deviations from $\mathcal{N}_{0,1}$ along all SPI categories weighted by their theoretical occurrence probability (see Table 2) is lowest for GD2 in both analyzed seasons (see legend in Fig. 4, (a)–(d)).

In agreement with Sienz et al. (2012), who identified notable differences in the performance of candidate PDFs between observations and simulations, this general ranking changes when we consider modeled instead of observed SPI_{3M} time-series (Fig. 4, (e)–(h)). While GD2, GGD3, and EWD3 perform similar in their representation of the observed frequency distribution of SPI_{3M} time-series (Fig. 4 (a), (c), and (d)), a noticeable difference emerges in simulations (Fig. 4 (e), (g), and (h)). GD2's performance distinctly deteriorates in simulations (Fig. 4, (e)) relative to observations. In contrast, both 3-parameter candidate PDFs excel in describing the frequency distribution of 3-months precipitation totals in both seasons (Fig. 4, (g) and (h)). Any distinction between both 3-parameter candidate distribution functions is still difficult (Fig. 4, (g) and (h)), Given the absolute deviations of GD2, one might most likely dismiss the need for any adjustment in SPI_{3M}'s calculation algorithm as of yet. However, since Fig. 4 shows the sum of deviations from $\mathcal{N}_{0,1}$ over all land grid-points of the entire globe, distribution functions might be oppositely wrong for the same SPI category in different grid-points resulting in deviations which balance each other across different grid-points.

In simulations, the fit onto 3-months precipitation totals is performed on all ten ensemble members at once. This leads to unequal databases (i.e. lengths of time-series) between observations and simulations. These unequal databases obscure any direct comparison between observed and modeled SPI_{3M} deviations. Therefore, deviations from $\mathcal{N}_{0,1}$ derived by different PDFs were compared separately for observations and simulations up to now. Such separate comparisons base on equally long time-series. Yet, deviations reduce non-identically along our four candidate distribution functions as a result of 10-folding the database of their fit. These irregular reductions provide us with the opportunity to analytically compare by how much deviations decrease for the same PDF as a result of 10-folding their database. The magnitude of this reduction should be notable for candidate distribution functions which are adequately suited to describe modeled 3-months precipitation totals – assuming an imperfect fit for the 31 events spanning our observational time-series.

For the 2-parameter PDFs, the weighted deviations from $\mathcal{N}_{0,1}$ either stay constant (for GD2 in DJF) or increase in simulations relative to observations (shown in the legend of Fig. 4, compare the left against the right column). GD2's weighted deviations increase by more than 120 % in JJA, while WD2's increase by more than 25 % in JJA and 80 % in DJF. The most plausible explanation for these weighted deviations to increase when 10-folding the database are different frequency distributions between observed and modeled 3-months precipitation totals. The 2-parameter PDFs are better suited to describe observed than modeled 3-months precipitation totals. In contrast, the 3-parameter candidate distribution functions benefit strongly from the artificial increase of our time-series. Their weighted deviations from $\mathcal{N}_{0,1}$ are substantially larger in observations than in simulations. GGD3's (EWD3's) are larger by 210 % (500 %) and 58 % (200 %) during DJF and JJA, respectively. These findings strongly hint at the presence of different frequency distributions between observed and modeled 3-months precipitation totals. Both 2-parameter candidate PDFs seem inadequately suited for describing modeled 3-months precipitation totals. In contrast,

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

the 3-parameter candidate distribution functions perform distinctly better in describing modeled 3-months precipitation totals than the 2-parameter candidate PDFs in both of our investigated seasons.

In this section, we have analyzed global deviations from $\mathcal{N}_{0,1}$ thus far and identified:

- GD2, GGD3, and EWD3 describe similarly well the overall frequency distribution of observed 3-months precipitation totals.
- WD2 performs overall poorly and is in every regard inferior to any other candidate distribution function.
- GGD3 and EWD3 describe the frequency distribution of modeled 3-months precipitation totals distinctly better than any 2-parameter candidate distribution.
- GD2 still describes the frequency distribution of modeled 3-months precipitation totals sufficiently well on global average.
- Both 2-parameter candidate distribution functions are unable to benefit from the increased length of the database in simulations relative to observations, while both 3-parameter PDFs strongly benefit from that increase.

As mentioned before, investigating deviations from $\mathcal{N}_{0,1}$ over the entire globe contains the risk of encountering deviations which balance each other in different grid-points. On the one hand, we can reduce that risk by analyzing these deviations only over specific regions, which is done later. On the other hand, we eliminate this risk next by examining AIC-D frequencies: aggregating AIC-D values over the entire globe evaluates the performance of PDFs in each grid-point and normalizes these evaluations by (rather than adding them over) the total number of grid-points of the entire globe.

In general, each of the candidate distribution functions perform similarly well in winter and summer in their depiction of the frequency distribution of observed 3-months precipitation totals (compare Fig. 5, (a) against (b)). In agreement with our previous results and prior studies (Blain et al., 2018; Lloyd-Hughes and Saunders, 2002; McKee et al., 1993), GD2 is in most grid-points of the global land area best suited to describe observed 3-months precipitation totals in DJF and JJA (Fig. 5, (a) and (b)). GD2 displays AIC-D values of less than 2 in approximately 84.5 % of the global land area in DJF and 83.5 % in JJA. That should be interpreted as substantial confidence in GD2's performance in these grid-points. However, beyond an AIC-D_{max} value of 2, EWD3 and GGD3 approach 100 % coverage considerably faster than GD2. The 3-parameter candidate distribution functions compensate rather quickly for their increased penalty imposed by AIC through a distinctly better performance in virtually every global land grid-point. GGD3 and EWD3 both show in more grid-points than GD2 an AIC-D_{max} value of approximately 2.5 (exactly 2.47 for EWD3 in both seasons and 2.51 (2.58) for GGD3 in DJF (JJA)) (see intersect between the yellowish and the bluish as well as the yellowish and black lines in Fig. 5, (a) and (b)). Further, once they compensate their penalty, they quickly approach 100 % coverage for the entire globe. For EWD3 more than 98 % of the land area is characterized in both seasons by an AIC-D_{max} value of less than 3 (98 % coverage is exactly fulfilled for an AIC-D_{max} value of 2.65 (2.95) in DJF (JJA)).

Contrarily, both 2-parameter candidate distribution functions display considerably less confidence in their description of observed 3-months precipitation totals in more than 10 % of the global land grid-points (apparent by AIC-D values beyond 4 in

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

these grid-points). In consequence, they need considerably longer to reach 98 % coverages – even allowing AIC-D$_{max}$ values as high as 6 does not lead to 98 % coverage for neither one of our 2-parameter candidate PDFs in any season (98 % coverage is for GD2 (WD2) exactly fulfilled for an AIC-D$_{max}$ value of 6.39 (6.46) in JJA and 6.68 (6.66) in DJF). As a reminder: AIC-D values between 4 and 7 indicate already considerably less confidence in the distribution function's performance. As a side note, EWD3 performs better than GGD3 but only by a few grid-points increased coverage for each AIC-D$_{max}$ value. Each candidate distribution function exhibits only in a minor fraction of grid-points essentially no confidence (AIC-D values of 10 and beyond) in their description of observed 3-months precipitation totals. GD2 (WD2) fails in its description in 0,41 % (0,49 %) and 0.59 % (0.26 %) of grid-points of the global land area in DJF and JJA, respectively. GGD2 only fails in 0.08 % (0.26 %) of grid-points of the global land area in DJF (JJA), while EWD3 does not fail in a single grid-point during both investigated seasons.

The confidence in GD2 drastically diminishes further when we analyze the performance of the four candidate PDFs in describing the frequency distribution of modeled 3-month precipitation totals. EWD3 is superior to any other distribution function in JJA and DJF for each AIC-D$_{max}$ value beyond 1.52 in DJF and 0.73 in JJA (see intersect between yellowish and blueish lines in Fig. 5, (c) and (d)). Assuming those AIC-D$_{max}$ values to be sufficiently small (AIC-D values of less than 2 are practically indistinguishable from each other in their performance), EWD3 performs best among all candidate PDFs in general. We interpret EWD3's description of the frequency distribution of modeled 3-months precipitation totals with substantial confidence in approximately 84.8 % of the global land area in DJF and 86.4 % in JJA. For AIC-D$_{max}$ values beyond 2, EWD3 again quickly approaches 100 % coverage in both seasons. Our results are again rather stable for all investigated distribution functions between summer and winter (compare Fig. 5, (c) against (d)). All distribution functions display in both seasons the same distinct ranking of their performance for AIC-D$_{max}$ values of 2 and beyond. EWD3 outperforms GGD3 which is better than GD2, while WD2 performs especially poor. In winter GGD3 performs better than GD2 for AIC-D$_{max}$ values beyond 1.99 (See intersect between blueish and black lines in Fig. 5, (c)). Here, both distributions functions performance should be interpreted with substantial confidence in almost 70 % (exactly 68.45 % for GD2 and 69.04 % for GGD3) of the global land area. However, for an AIC-D$_{max}$ value of just 2.1, GGD3 already out-performs GD2 in 7.92 % (11.75 %) of the global land area during winter (summer).

While EWD3 does not display a deteriorating performance in simulations in more than 1 % of grid-points, there is season-dependent considerably less confidence in GD2's performance in about one-third to one-fourth of the global land grid-points (apparent by AIC-D values beyond 4 in these grid-points). Most telling might be the fraction of grid-points in which the candidate PDFs display AIC-D values of 10 and beyond and thus show no confidence in their depiction of 3-months precipitation totals. GD2 and WD2 fail in their description during DJF (JJA) in 9.87 % (14.95 %) and 57.84 % (56.57 %) of the global land area, respectively. While GGD3 still fails in 3.61 % (4.23 %) of grid-points, EWD3 only fails in 0.59 % (0.71 %) during DJF (JJA). Ergo, EWD3 reduces the count of grid-points in which it's description of modeled 3-months precipitation totals is without any skill by over one magnitude (by a factor of roughly 20) relative to GD2.

Table 3 summarizes our findings from the investigation of AIC-D values over the entire global land area. While not even a single candidate PDF performs ideally with substantial confidence around the globe (AIC-D ≤ 2 in 95 or more % of land

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

grid-points) in either realization, EWD3 performs well with substantial confidence around the globe (AIC-D ≤ 4 in 95 or more % of land grid-points) in both realizations. The other analyzed candidate PDFs perform substantially worse than EWD3 in simulations and slightly worse in observations. It seems worth elaborating on the combination between EWD3's increased penalty relative to our 2-parameter candidate PDFs and the fact that EWD3 does not perform ideally with substantial confidence around the globe. On the one side, EWD3's increased complexity justifies the increased penalty when evaluating whether that increased complexity is necessary. However, the results justify the necessity for this increased complexity. The risk of underfitting by using 2-parameter PDFs is higher than the risk of overfitting by using 3-parameter PDFs. In particular, when we demand that a single candidate PDF should be suited for observations and simulations concurrently, 2-parameter candidate PDFs seem ill-posed for the task at hand. On the other side, once the need for 3-parameter candidate PDFs is established, their remaining competition against 2-parameter PDFs biases the analysis; especially for the ideal AIC-D category. EWD3's increased penalty relative to 2-parameter candidate PDFs depends on the sample size and amounts to 2.46 in observations and 2.04 in simulations. This penalty is also approximately the AIC-D$_{max}$ value where EWD3 reaches a coverage close to 100 % (Fig. 5 (a)–(d)). Indeed, if EWD3 solely competes with GGD3, EWD3 performs ideally (AIC-D ≤ 2) over both seasons in observations (simulations) in 99 % (100 %) of the global land grid-points (not shown). Thus, EWD3 already performs at least on par with the best-performing candidate PDF in both realizations at virtually every grid-point.

These characteristics stay valid in all investigated regions except Australia. Here, GD2 performs better than any other analyzed PDF during DJF in observations. In contrast during JJA-observations, GD2 performs worse than any other investigated candidate PDFs (even WD2). Additionally, WD2 and the other candidate PDFs also out-perform GD2 during DJF in simulations. Since these are the only minor regional particularities evident in regional AIC-D frequencies, we will during the regional focus in the remaining analysis of SPI$_{3M}$ solely display, explain, and concentrate on deviations from $\mathcal{N}_{0,1}$.

3.1.2 Regional Deviations from $\mathcal{N}_{0,1}$

We investigated thus far deviations from $\mathcal{N}_{0,1}$ for the entire global land area. That analysis might be blurred by deviations which balance each other over totally different regions with unrelated climatic characteristics. Thus, we will reduce the area analyzed in this subsection and perform a further aggregated investigation for each continental region individually. That further aggregation of results dismisses the dimension of different SPI categories because their analysis revealed a rather uniform relation over each region: extreme SPI categories show the largest deviations, while normal conditions exhibit the smallest. As a consequence, we display from now on only unweighted sums over the absolute values of these deviations from all SPI categories. To provide a more intuitive number for these unweighted sums, we normalize them by our SPI category count (7). Consequently, our analysis will investigate mean deviations per SPI category, henceforth.

In observations (Fig 6. (a) and (b)), WD2 performs in all analyzed regions again worst of all candidate PDFs in describing a proper frequency distribution of SPI$_{3M}$ during both investigated seasons. Over all analyzed regions and seasons, EWD3 displays the smallest deviations from $\mathcal{N}_{0,1}$, while GD2 and GGD3 perform only slightly worse. Some minor region-dependent differences emerge. E.g. in Africa, a distinct ranking of the performance of all four candidate distribution functions emerges

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

during JJA – EWD3 outperforms GGD3 which performs better than GD2. Aside, all candidate PDFs perform almost identical in their attempt to describe observed precipitation over Australia during DJF.

In simulations (Fig 6. (c) and (d)), the ranking of the performance of different PDFs becomes more distinct than it is in observations during both analyzed seasons and investigated domains, except Australia. This compared to observations easier distinction over almost every region of the globe results from increased mean deviations for GD2, while they stay comparable low for GGD3 and EWD3, relative to the global analysis. As showed before, 2-parameter PDFs are inadequately suited to properly describe modeled precipitation totals. In consequence, during both seasons, GGD3 and EWD3 perform in each region exceptionally well, while GD2 performs overall average at best, whereas WD2 performs still poor in general. The performances of GD2 and WD2 are only in Africa during DJF equally poor which impedes any clear ranking. Similar difficult is any distinction of their performance in North America during JJA as a consequence of one of WD2's best performances (as also identified by Sienz et al. (2012) for SPI_{1M}). Furthermore poses Australia an exception to the identified ranking pattern of candidate PDFs for simulations. During the austral summer (DJF), WD2 distinctly outperforms GD2 which exhibits the largest mean deviations. Interestingly, analog to the performance of candidate PDFs over Australia in observations during DJF, we identify over Australia also in simulations a season when the performance of all four candidate distribution functions is rather similar. However, this occurs in simulations during JJA.

These insights about the candidate PDFs performance in observations and simulations are even more obvious at first glance when displayed in an image plot (Fig. 7 (a) and (b)). The poor performance of WD2 in observations and simulations is obvious over all domains and in both investigated seasons. Also, the exception to this pattern for Australia during the austral summer (Fig. 7 (a)) in simulations is distinctly visible. Evident are further the overall similar performances of GD2, GGD3 and EWD3 in observations over all domains and both analyzed seasons. Further, the general improved performance of 3-parameter candidate distribution functions (GGD3 and EWD3) relative to 2-parameter candidate PDFs in simulations is distinctly palpable. Aside, even the better performance of EWD3 relative to GGD3 in Africa generally or in observations over Europe is easily discernible.

The regional analysis confirms the overall insights from the global analysis in observations for each region. In simulations, the regional analysis additionally corroborates the finding of the AIC-D analysis that our 3-parameter candidate distribution functions perform in simulations noticeably better than our 2-parameter PDFs. The corroboration of this finding substantiates support for the 3-parameter candidate PDFs.

3.1.3 Improvement relative to a multi-PDF Approach and a Baseline

In the following, we investigate deviations from $\mathcal{N}_{0,1}$ for a multi-PDF SPI calculation algorithm which uses in each grid-point that distribution function which yields for this respective grid-point the minimum AIC value (whose AIC-D value equates to 0). An analog SPI calculation algorithm has been repeatedly proposed in literature (Guenang et al., 2019; Blain and Meschiatti, 2015; Touma et al., 2015; Sienz et al., 2012; Lloyd-Hughes and Saunders, 2002). We analyze the impact of such an SPI calculation algorithm and compare those results against a baseline comparison and against the most suitable calculation algorithm identified in this study which uses EWD3 as PDF. We label the results obtained from the multi-distribution function calculation algorithm AIC_{min}-analysis. As a baseline comparison, we choose the calculation algorithm and optimization method of

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

the frequently used R-package from Beguería and Vicente-Serrano (2017) and refer to these results as baseline. To maximize comparability of SPI time-series calculated with our baseline, we employ the simple 2-parameter gamma distribution as a calculation algorithm and estimate the parameters of the PDF again with the *maximum-likelihood method*. It seems noteworthy that our parameter estimation method takes about 60 times longer to find optimal parameters of GD2 than the baseline. The comparison between the performance of our baseline against GD2's performance (see Fig. 7 (a) and (b)) thus also serves as an indicator for the impact of very similar parameter estimation methods which only differ by their optimization procedure.

The AIC_{min}-analysis performs generally almost identical to EWD3 over each domain and in both realizations (observations and simulations). Further, deviations are not necessarily minimal when computing SPI with the AIC_{min}-analysis (Fig. 8, (a) and (b)). This results from the dependence of AIC's punishment on the parameter count of the distribution function. It is simply not sufficient for EWD3 to perform best by a small margin in order to yield a lower AIC value than GD2/WD2. EWD3 needs to perform sufficiently better to over-compensate its by AIC imposed punishment. Or in other words, EWD3 is expected to perform distinctly better than GD2/WD2 because of its increased complexity. As a consequence, EWD3 is only selected by AIC as best performing distribution function if it fulfills that expectation.

In contrast to previous results (Stagge et al., 2015), which showed no seasonal differences in the performance of candidate PDFs, our baseline performs overall better in JJA than in DJF (compare in Fig. 8, (a) against (b)). Relative to our findings in the previous subsection (Fig 7.), our baseline performs similar to GD2 in JJA but worse than WD2 in DJF (compare Fig. 7 against Fig. 8,). This reveals a substantial impact of the optimization method, at least for DJF-precipitation totals. Further, our baseline performs especially poor in describing the frequency distribution of SPI_{3M} in simulations during the austral summer. It is important to note that our baseline over-estimates modeled extreme droughts during DJF over Australia by more than 240 % (not shown). That is by a huge margin the largest deviation we encountered during our analysis and highly undesirable when analyzing droughts. Contrary to Blain et al. (2018), who investigated the influence of different parameter estimation methods on SPI's normality and identified only barely visible effects, the massive difference between our baseline and GD2 in DJF is severely concerning; especially given that the here used parameter estimation methods are almost identical and only differ by their optimization procedure. Since GD2 and our baseline both use the maximum likelihood method to estimate the PDF's parameters, main differences do not only emerge when using different estimation methods but rather manifest already in the applied procedure by which these methods are optimized.

Unsurprisingly the same deficit as identified before for both 2-parameter candidate PDFs also emerges in our baseline's performance: the by each classes' likelihood of occurrence weighted sum over the absolute values of deviations from $\mathcal{N}_{0,1}$ increases as a result of 10-folding our database (not shown). Although our baseline already performs especially poor when analyzing weighted deviations during DJF in observations, it performs even worse in simulations; although the performance deteriorates only marginally. Such an increase of weighted deviations is a strong indicator of our baseline's inability to sufficiently describe the frequency distribution of modeled SPI_{3M}. In our baseline, these weighted deviations increase globally by 2 % in DJF and 40 % in JJA (as a reminder: the weighted deviations stay constant for GD2 in DJF and increase by more than 120 % in JJA). In contrast, these weighted deviations decrease for the AIC_{min}-analysis by 70 % in DJF and by 60 % in JJA around the entire globe (not shown).

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

Moreover, identifying the maximum deviation from $\mathcal{N}_{0,1}$ for 196 different analyses which range along each SPI category (7), region (7), both seasons (2), as well as differentiating between observation and simulation (2) (not shown), our baseline performs worst in 79 out of those 196 analyses, while WD2 performs worst in 103 of these analyses. It is noteworthy that out of those 79 analyses in which our baseline performs worst, 63 analyses occur during DJF. As a side note, GD2 performs with our optimization overall worst six times, while GGD3 and EWD3 each perform worst four times.

3.2 Other SPI Accumulation Periods

A similar pattern as identified for SPI_{3M} also emerges in the evaluation of AIC-D-based performances of our candidate PDFs for accumulation periods of 1-, 6-, 9-, and 12-months (Table 4). No candidate PDF performs ideally (AIC-D values below 2) with substantial confidence around the globe. The reasons for this shortcoming are distribution-dependent. GD2 performs too poor in too many grid-points (e.g. apparent by too low percentages for covering AIC-D values even below 4) and EWD3 excels only for AIC-D values beyond 2 because it first needs to over-compensate its AIC-imposed complexity-penalty (as explained before). Equally apparent is the striking inability of the 2-parameter candidate PDFs to adequately perform in simulations for all analyzed accumulation periods which we have also seen for SPI_{3M} before.

In agreement with prior studies (Stagge et al., 2015; Sienz et al., 2012), we also identify the apparent phase transition between short (less than 3-months) and long (more than 3-months) accumulation periods for the 2-parameter candidate PDFs. While WD2 performs well for short accumulation periods (only in observations though), GD2 performs better than WD2 for longer accumulation periods. Nevertheless, the results for the 3-parameter candidate PDFs do not display such a phase transition.

Most interesting, EWD3 performs well almost everywhere around the entire globe for each accumulation period and in both realizations. EWD3 shows the highest percentages of all candidate PDFs for each analysis (each row of Table 4) beyond AIC-D values of 2; except for an accumulation period of 12-months in simulations. While there is not even a single candidate PDF that seems sufficiently well suited for an accumulation period of 12-months in simulations, GD2 and EWD3 both perform equally adequate; despite EWD3's higher AIC-penalty compared to GD2. If EWD3 only competes against GGD3, EWD3 performs ideal (AIC-D ≤ 2) in 88 % and shows no skill (AIC-D > 10) in less than 5 % of the global land grid-points. Moreover, EWD3 performs best in 32 out of all 40 analyses (all rows of Table 3 and Table 4), and in 30 of those 32 analyses, we consider EWD3's performance to display at least average confidence (indicated by a yellow or green background color in the table). In contrast, GD2 (WD2) only performs 2 (1) times best while also performing with at least average confidence and GGD2 never performs best.

4 Discussion

Previous studies have emphasized the importance of using a single PDF to calculate SPI for each accumulation period and location (Stagge et al., 2015; Guttman, 1999) to ensure comparability across space and time which is one of the index's main advantages (Lloyd-Hughes and Saunders, 2002). However, any 2-parameter distribution function seems in observations already ill-suited to deliver adequately normally distributed SPI time-series for both short (less than 3-months) and long (more

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

than 3-months) accumulation periods (Stagge et al., 2015; Sienz et al., 2012). Introducing simulations as another level of complexity exacerbates the problem additionally. Yet, the importance of accepting and solving this problem becomes increasingly pressing as a result of a growing interest in dynamical drought predictions and their evaluation against observations. To properly evaluate drought predictability of precipitation hindcasts against observations, the distribution function used in SPI's calculation algorithm needs to capture sufficiently well both frequency distributions mutually: those of observed and modeled precipitation totals. In this study, we show that the 3-parameter exponentiated Weibull distribution (EWD3) is very promising in solving this problem virtually everywhere on the entire globe in both realizations (observations and simulations) for all common accumulation periods (1-, 3-, 6-, 9-, and 12-months).

Other studies have pessimistically dismissed the possibility of such a solution to this problem and proposed instead a multi-PDF approach (Guenang et al., 2019; Blain and Meschiatti, 2015; Touma et al., 2015; Sienz et al., 2012; Lloyd-Hughes and Saunders, 2002) which selects different PDFs depending on the location and accumulation period of interest. The emergence of this approach stems from a phase transition in the relative performance of 2-parameter PDFs, which we also identify in this study. While WD2 performs better for an accumulation period of 1-month, GD2 is better suited for longer accumulation periods. However, any multi-PDF approach would partly sacrifice the aforementioned index's pivotal advantage of comparability across space and time. Our results suggest that such a multi-PDF approach does not improve the normality of calculated SPI time-series relative to a calculation algorithm that uses EWD3 as PDF everywhere. Furthermore, the use of an empirical cumulative distribution function has been proposed (Sienz et al., 2012). We also checked this approach which proved to be too coarse as a result of its discretized description (not shown).

Yet, in agreement with those other studies (Guenang et al., 2019; Blain and Meschiatti, 2015; Touma et al., 2015; Sienz et al., 2012; Lloyd-Hughes and Saunders, 2002), our results also suggest that 2-parameter PDFs are not able to produce sufficiently normally distributed SPI time-series for all accumulation periods, locations, and realizations. Yet, EWD3 competed against 2-parameter PDFs in our analysis. This competition unnecessarily (given the inadequacy of 2-parameter PDFs) exacerbates EWD3's performance assessed with AIC-D because AIC punishes complexity. As a consequence of EWD3's increased complexity, AIC imposes a larger penalty on EWD3 than on the 2-parameter candidate PDFs which are anyhow ill-suited to solve the outlined problem (because they are most likely too simple). Still, EWD3 conclusively out-performs any other candidate PDF without performing ideally. However, accepting the need for a 3-parameter PDF in SPI's calculation algorithm a priori levels the playing field in our AIC-D analysis and leads to an ideal performance of EWD3 globally.

The findings sketched above stay valid on every continent in both realizations with a few exceptions. It seems noteworthy, that Australia's observed DJF- and modeled JJA-precipitation totals are generally poorly described by any of our candidate distribution functions. Since the performance of all investigated distribution functions deteriorate to a similar level, it is difficult, however, to discern any new ranking. Even more troublesome is the proper description of simulated 12-months precipitation totals. Here, our candidate PDFs perform only sufficiently. Yet, despite its increased AIC-penalty, EWD3 performs still best along the 2-parameter gamma distribution.

In contrast to Blain et al. (2018), who investigated the influence of different parameter estimation methods on the normality of the resulting SPI time-series and only found minuscule effects, our results show a substantial impact. Despite using the same

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

parameter estimation methods and the same candidate PDF, the baseline investigated here enlarges deviations from $\mathcal{N}_{0,1}$ by roughly half a magnitude compared to GD2 in DJF. This result is concerning because it indicates that main differences do not only emerge when using different parameter estimation methods but rather manifest already in the applied procedure by which these methods are optimized. In our analysis, not different PDFs but different optimizations of the same parameter estimation method impact normality most profoundly.

Other consequences of this finding are apparent major season-dependent differences in the performance of the investigated baseline. This finding contradicts the results of Stagge et al. (2015) in which no seasonal differences in the performance of candidate PDFs emerged. While the baseline performs similar to GD2 during JJA, its performance severely deteriorates during DJF in our analysis. While this deterioration is overall more apparent in observations than in simulations, its most obvious instance occurs in simulations. The investigated baseline over-estimates modeled extreme droughts in Australia during DJF by more than 240 %. Therefore we urge to exercise substantial caution while analyzing SPI_{DJF} time-series with the investigated baseline's R-package irrespective of the heritage of input data. In our analysis, we encounter the largest mean deviations in the baseline. These deviations occur during DJF in Australia, but the baseline performs particularly poor during DJF in general. During DJF, the examined baseline displays larger deviations from $\mathcal{N}_{0,1}$ than any other of the here analyzed 6 SPI calculations (GD2, WD2, GGD3, EWD3, baseline, and AIC_{min}-analysis) in 63 out of 98 different analyses, which range along all seven SPI categories, all seven regions, and along observations as well as simulations. Aside from the investigated baseline and in agreement with (Stagge et al., 2015), we find no seasonal differences in the performance of our candidate PDFs.

To aggregate our AIC-D-analysis over the globe and visualize this aggregation in tables, we need to evaluate the aggregated performance of candidate PDFs for certain AIC-D categories (Burnham and Anderson, 2002). Their aggregation over all land grid-points of the globe demands the introduction of two further performance criteria which require interpretation. These criteria inform whether the candidate PDFs conform the respective AIC-D categories in sufficient grid-points globally and, therefore, need to interpret which fraction of the global grid-points can be considered sufficient. For this fraction of global land grid-points, we select 85 % and 95 % as thresholds. In consequence, we categorize our candidate PDFs for each AIC-D category into three different classes of possible performances. We consider the confirmation of the respective AIC-D category in 95 % or more grid-points globally as an indicator of substantial confidence in the candidate PDF to perform according to the respective AIC-D category globally. Confirmation of the respective AIC-D category in less than 85 % of grid-points globally is considered as an indicator of insufficient confidence in the candidate PDF. Finally, we consider it to be an indicator of average confidence in candidate PDFs when they conform to the respective AIC-D category in between 85 % and 95 % of grid-points globally. One might criticize that these thresholds lack a scientific foundation or that they are to some extent arbitrary. However, they seem adequately reasonable and agree with analog evaluations of such fractions derived by rejection frequencies from goodness-of-fit tests in previous studies (Blain et al., 2018; Blain and Meschiatti, 2015; Stagge et al., 2015; Lloyd-Hughes and Saunders, 2002). Moreover, these thresholds show a robust statistical basis in terms of being equally represented over all 160 analyzed evaluations in this study (all entries of Table 3 and Table 4). Across all 40 analyses (all rows of Table 3 and Table 4), the four candidate PDFs perform insufficiently 65 times, while they perform with substantial (average) confidence 64 (31) times.

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020


Hydrology and Earth System Sciences
Discussions

There is scope to further test the robustness of our derived conclusions in different models with different time horizons and accumulation periods other than 3-months (e.g. 12-months). Of additional interest would be insights about the distribution of precipitation. Such insights would enable SPI's calculation algorithm to physically base its key decision.

The results presented here further imply that the evaluated predictive skill of drought predictions assessed with SPI should be treated with caution because it is likely biased by SPI's current calculation algorithms. This bias in SPI's common calculation algorithms obscures the evaluation of predictive skill of simulations by inducing a blurred representation of the frequency distribution of modeled precipitation totals. That blurred representation translates to the simulated drought index which impedes the evaluation process. Drought predictions often try to correctly predict the drought intensity. The evaluation process usually considers this to be successfully achieved if the same SPI category as the observed one is predicted. This evaluation is quite sensitive to the thresholds used when classifying SPI categories. The bias identified here blurs these categories for the model but not for observations against which the model's predictability is customarily evaluated. As a consequence of these sensitive thresholds, such a one-sided bias potentially undermines current evaluation processes.

5 Summary and Conclusions

We investigate different candidate distribution functions (gamma (GD2), Weibull (WD2), generalized gamma (GGD3), and exponentiated Weibull distribution (EWD3)) in SPI's calculation algorithm concerning their adequacy in meeting SPI's normality requirement. We conduct this investigation for observations and simulations during summer (JJA) and winter (DJF). Our analysis evaluates globally and over each continent individually the resulting SPI_{3M} time-series based on their normality while focusing on an accumulation period of 3-months and testing the conclusions drawn from that focus for the most common other accumulation periods (1-, 6-, 9-, and 12-months). Normality of SPI is assessed by comparing actual occurrence probabilities of SPI categories (as defined by WMO's *SPI User Guide* (Svoboda et al., 2012)) against well-known theoretical occurrence probabilities of $\mathcal{N}_{0,1}$. To penalize unnecessary complexity we employ Akaike's Information Criterion (AIC).

Our results show that GD2 is sufficiently suited to calculate SPI derived from observations for all accumulation periods analyzed. WD2 performs in observations better for an accumulation period of 1-months but worse for longer accumulation periods. Based on our analysis of AIC-D values and deviations from $\mathcal{N}_{0,1}$, EWD3 performs exceptionally well and better than any 2-parameter candidate PDF in observations for all accumulation periods. Further, we identify considerable differences between observations and simulations. For all accumulation periods analyzed in simulations, both 2-parameter candidate PDFs perform inadequately (WD2) or sufficiently but only with average confidence around the globe (GD2). In contrast, EWD3 performs particularly well with substantial confidence around the entire globe in simulations and for every accumulation period analyzed. The accumulation period of 12-months poses in simulations the only exception. Here, EWD3 still performs well but only with average confidence around the globe. We find that 3-parameter PDFs are generally better suited in SPI's calculation algorithm than 2-parameter PDFs. Our results show that the risk of overfitting 3-parameter PDFs is overcompensated by the risk of underfitting 2-parameter PDFs. We strongly advocate to adapt and use 3-parameter distribution functions instead of 2-parameter PDFs for the calculation algorithm of SPI. Such an adaptation is particularly important for the proper evaluation

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

and interpretation of drought predictions and simulations. For this adaptation, we propose the employment of EWD3 as new standard PDF for SPI's calculation algorithm, irrespective of the heritage of input data or the length of scrutinized accumulation periods. Despite the issues discussed here, SPI remains a valuable tool for analyzing droughts. This study might contribute to the value of this tool by illuminating and resolving the discussed long-standing issue concerning the proper calculation of the index.

Data availability. The model simulation will be made available. The process of archiving this data at the German Climate Computing Centre and providing online access is underway. A download-link will be supplied at a later stage of the review-process.

Author contributions. PP, AD, and JB designed the study. PP led the analysis and prepared the manuscript with support from all co-authors. All co-authors contributed to the discussion of the results.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The work of P.P. is supported by the Stiftung der deutschen Wirtschaft (SDW, German Economy Foundation). A.D. and J.B. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC 2037 "Climate, Climatic Change, and Society"–Project: 390683824, contribution to the Center for Earth System Research and Sustainability (CEN) of Universität Hamburg. A.D. is also supported by A4 (Aigéin, Aeráid, agus athrú Atlantaigh), funded by the Marine Institute and the European Regional Development fund (grant: PBA/CC/18/01). The model simulations were performed at the German Climate Computing Centre. The authors also thank Frank Sienz for providing the software to compute AIC and SPI with different candidate distribution functions.

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

Hydrology and Earth System Sciences
Discussions

References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., et al.: The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present), Journal of hydrometeorology, 4, 1147–1167, 2003.

Akaike, H.: A new look at the statistical model identification, in: Selected Papers of Hirotugu Akaike, pp. 215–222, Springer, 1974.

Baehr, J., Fröhlich, K., Botzet, M., Domeisen, D. I., Kornblueh, L., Notz, D., Piontek, R., Pohlmann, H., Tietsche, S., and Mueller, W. A.: The prediction of surface temperature in the new seasonal prediction system based on the MPI-ESM coupled climate model, Climate Dynamics, 44, 2723–2735, 2015.

Beguería, S. and Vicente-Serrano, S. M.: Calculation of the Standardised Precipitation-Evapotranspiration Index, 2017.

Bélisle, C. J.: Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$, Journal of Applied Probability, 29, 885–895, 1992.

Blain, G. C. and Meschiatti, M. C.: Inadequacy of the gamma distribution to calculate the Standardized Precipitation Index, Revista Brasileira de Engenharia Agrícola e Ambiental, 19, 1129–1135, 2015.

Blain, G. C., de Avila, A. M. H., and Pereira, V. R.: Using the normality assumption to calculate probability-based standardized drought indices: selection criteria with emphases on typical events, International Journal of Climatology, 38, e418–e436, 2018.

Bunzel, F., Müller, W. A., Dobrynin, M., Fröhlich, K., Hagemann, S., Pohlmann, H., Stacke, T., and Baehr, J.: Improved Seasonal Prediction of European Summer Temperatures With New Five-Layer Soil-Hydrology Scheme, Geophysical Research Letters, 45, 346–353, 2018.

Burnham, K. P. and Anderson, D. R.: Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach, 2002.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C.: A limited memory algorithm for bound constrained optimization, SIAM Journal on Scientific Computing, 16, 1190–1208, 1995.

Field, C. B., Barros, V., Stocker, T. F., and Dahe, Q.: Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change, Cambridge University Press, 2012.

Giddings, L., SOTO, M., Rutherford, B., and Maarouf, A.: Standardized precipitation index zones for Mexico, Atmósfera, 18, 33–56, 2005.

Giorgi, F. and Francisco, R.: Evaluating uncertainties in the prediction of regional climate change, Geophysical Research Letters, 27, 1295–1298, 2000.

Guenang, G., Komkoua, M., Pokam, M., Tanessong, R., Tchakoutio, S., Vondou, A., Tamoffo, A., Djiotang, L., Yepdo, Z., and Mkankam, K.: Sensitivity of SPI to Distribution Functions and Correlation Between its Values at Different Time Scales in Central Africa, Earth Systems and Environment, pp. 1–12, 2019.

Guttman, N. B.: Accepting the standardized precipitation index: a calculation algorithm, JAWRA Journal of the American Water Resources Association, 35, 311–322, 1999.

Hayes, M., Svoboda, M., Wall, N., and Widhalm, M.: The Lincoln declaration on drought indices: universal meteorological drought index recommended, Bulletin of the American Meteorological Society, 92, 485–488, 2011.

Jungclaus, J., Fischer, N., Haak, H., Lohmann, K., Marotzke, J., Matei, D., Mikolajewicz, U., Notz, D., and Storch, J.: Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI-Earth system model, Journal of Advances in Modeling Earth Systems, 5, 422–446, 2013.

Kullback, S. and Leibler, R. A.: On information and sufficiency, The annals of mathematical statistics, 22, 79–86, 1951.

Lloyd-Hughes, B. and Saunders, M. A.: A drought climatology for Europe, International Journal of Climatology: A Journal of the Royal Meteorological Society, 22, 1571–1592, 2002.

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

Ma, F., Yuan, X., and Ye, A.: Seasonal drought predictability and forecast skill over China, Journal of Geophysical Research: Atmospheres, 120, 8264–8275, 2015.

McKee, T. B. et al.: The relationship of drought frequency and duration to time scales, in: Proceedings of the 8th Conference on Applied Climatology, vol. 17, pp. 179–183, American Meteorological Society Boston, MA, 1993.

Mo, K. C. and Lyon, B.: Global meteorological drought prediction using the North American multi-model ensemble, journal of Hydrometeorology, 16, 1409–1424, 2015.

Naresh Kumar, M., Murthy, C., Sesha Sai, M., and Roy, P.: On the use of Standardized Precipitation Index (SPI) for drought intensity assessment, Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling, 16, 381–389, 2009.

Nelder, J. A. and Mead, R.: A simplex method for function minimization, The computer journal, 7, 308–313, 1965.

Nocedal, J. and Wright, S. J.: Springer series in operations research. Numerical optimization, 1999.

Pietzsch, S. and Bissolli, P.: A modified drought index for WMO RA VI, Advances in Science and Research, 6, 275–279, 2011.

Quan, X.-W., Hoerling, M. P., Lyon, B., Kumar, A., Bell, M. A., Tippett, M. K., and Wang, H.: Prospects for dynamical prediction of meteorological drought, Journal of Applied Meteorology and Climatology, 51, 1238–1252, 2012.

Ribeiro, A. and Pires, C.: Seasonal drought predictability in Portugal using statistical–dynamical techniques, Physics and Chemistry of the Earth, Parts A/B/C, 94, 155–166, 2016.

Sienz, F., Bothe, O., and Fraedrich, K.: Monitoring and quantifying future climate projections of dryness and wetness extremes: SPI bias, Hydrology and Earth System Sciences, 16, 2143, 2012.

Stagge, J. H., Tallaksen, L. M., Gudmundsson, L., Van Loon, A. F., and Stahl, K.: Candidate distributions for climatological drought indices (SPI and SPEI), International Journal of Climatology, 35, 4027–4040, 2015.

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K., et al.: Atmospheric component of the MPI-M Earth system model: ECHAM6, Journal of Advances in Modeling Earth Systems, 5, 146–172, 2013.

Svoboda, M., Hayes, M., and Wood, D.: Standardized precipitation index user guide, World Meteorological Organization Geneva, Switzerland, 2012.

Touma, D., Ashfaq, M., Nayak, M. A., Kao, S.-C., and Diffenbaugh, N. S.: A multi-model and multi-index evaluation of drought characteristics in the 21st century, Journal of Hydrology, 526, 196–207, 2015.

Wu, H., Svoboda, M. D., Hayes, M. J., Wilhite, D. A., and Wen, F.: Appropriate application of the standardized precipitation index in arid locations and dry seasons, International Journal of Climatology, 27, 65–79, 2007.

Yoon, J.-H., Mo, K., and Wood, E. F.: Dynamic-model-based seasonal prediction of meteorological drought over the contiguous United States, Journal of Hydrometeorology, 13, 463–482, 2012.

Yuan, X. and Wood, E. F.: Multimodel seasonal forecasting of global drought onset, Geophysical Research Letters, 40, 4900–4905, 2013.

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Table 1. Abbreviations used for candidate distribution functions.

Distribution function	Parameter count	Abbreviation
Gamma distribution	2	GD2
Weibull distribution	2	WD2
Generalized gamma distribution	3	GGD3
Exponentiated Weibull distribution	3	EWD3

Table 2. Standardized Precipitation Index (SPI) classes with their corresponding definition and occurrence probabilities (according to WMO's *SPI User Guide* (Svoboda et al., 2012)).

SPI interval	SPI class	Probability [%]
$\text{SPI} \geq 2$	W3: extremely wet	2.3
$2 > \text{SPI} \geq 1.5$	W2: severely wet	4.4
$1.5 > \text{SPI} \geq 1$	W1: moderately wet	9.2
$1 > \text{SPI} > -1$	N0: normal	68.2
$-1 \geq \text{SPI} > -1.5$	D1: moderately dry	9.2
$-1.5 \geq \text{SPI} > -2$	D2: severely dry	4.4
$\text{SPI} \leq -2$	D3: extremely dry	2.3

Table 3. Percent of grid-points which are classified according to Burnham and Anderson (2002) depending on whether they display AIC-D values lower than specific thresholds or higher than 10 for each candidate PDF over both seasons. Percentages of grid-points indicate the confidence in candidate PDFs to overall perform according to the respective AIC-D category. We consider percentages that exceed (subceed in case of AIC-D values beyond 10) 95 % (5 %) as sign of substantial confidence in the candidate PDF (green) to overall perform according to the respective AIC-D category. In contrast, we consider those candidate PDFs which exceed/subceed in 85/15 % of the grid-points as sign of average confidence in the candidate PDF (yellow) to overall perform according to the respective AIC-D category. Percentages which fall short of 85 % (or which show no skill in more than 15 %) are considered as overall sign of insufficient confidence in the candidate PDF (red).

SPI Period	Realization	AIC-D category	GD2	WD2	GGD3	EWD3
3-Months	Observations	Ideal (AIC-D ≤ 2)	**84**	**76**	**22**	**31**
		Well (AIC-D ≤ 4)	**94**	**91**	**98**	**100**
		Sufficient (AIC-D ≤ 7)	**98**	**98**	**100**	**100**
		No Skill (AIC-D > 10)	**1**	**0**	**0**	**0**
	Simulations	Ideal (AIC-D ≤ 2)	**65**	**18**	**68**	**86**
		Well (AIC-D ≤ 4)	**74**	**24**	**89**	**99**
		Sufficient (AIC-D ≤ 7)	**82**	**34**	**94**	**99**
		No Skill (AIC-D > 10)	**12**	**57**	**4**	**1**

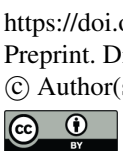
https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Table 4. Percent of grid-points which are classified according to Burnham and Anderson (2002) depending on whether they display AIC-D values lower than specific thresholds or higher than 10 for each candidate PDF over both seasons. Percentages of grid-points indicate the confidence in candidate PDFs to overall perform according to the respective AIC-D category. We consider percentages that exceed (subceed in case of AIC-D values beyond 10) 95 % (5 %) as sign of substantial confidence in the candidate PDF (green) to overall perform according to the respective AIC-D category. In contrast, we consider those candidate PDFs which exceed/subceed in 85/15 % of the grid-points as sign of average confidence in the candidate PDF (yellow) to overall perform according to the respective AIC-D category. Percentages which fall short of 85 % (or which show no skill in more than 15 %) are considered as overall sign of insufficient confidence in the candidate PDF (red).

SPI Period	Realization	AIC-D category	GD2	WD2	GGD3	EWD3
1-Month	Observations	Ideal (AIC-D $\leq$ 2)	84	86	30	33
		Well (AIC-D $\leq$ 4)	94	97	100	100
		Sufficient (AIC-D $\leq$ 7)	98	99	100	100
		No Skill (AIC-D $>$ 10)	0	0	0	0
	Simulations	Ideal (AIC-D $\leq$ 2)	55	43	81	87
		Well (AIC-D $\leq$ 4)	64	54	96	100
		Sufficient (AIC-D $\leq$ 7)	73	66	98	100
		No Skill (AIC-D $>$ 10)	21	26	1	0
6-Months	Observations	Ideal (AIC-D $\leq$ 2)	82	67	16	30
		Well (AIC-D $\leq$ 4)	93	86	96	99
		Sufficient (AIC-D $\leq$ 7)	99	98	99	100
		No Skill (AIC-D $>$ 10)	0	0	0	0
	Simulations	Ideal (AIC-D $\leq$ 2)	75	11	49	77
		Well (AIC-D $\leq$ 4)	82	15	82	95
		Sufficient (AIC-D $\leq$ 7)	88	22	90	97
		No Skill (AIC-D $>$ 10)	8	71	7	2
9-Months	Observations	Ideal (AIC-D $\leq$ 2)	83	64	13	28
		Well (AIC-D $\leq$ 4)	93	84	93	98
		Sufficient (AIC-D $\leq$ 7)	99	97	98	99
		No Skill (AIC-D $>$ 10)	0	1	1	0
	Simulations	Ideal (AIC-D $\leq$ 2)	75	10	40	76
		Well (AIC-D $\leq$ 4)	82	13	76	93
		Sufficient (AIC-D $\leq$ 7)	89	18	85	95
		No Skill (AIC-D $>$ 10)	7	76	12	3
12-Month	Observations	Ideal (AIC-D $\leq$ 2)	82	61	13	29
		Well (AIC-D $\leq$ 4)	92	81	91	96
		Sufficient (AIC-D $\leq$ 7)	98	96	97	98
		No Skill (AIC-D $>$ 10)	1	1	1	1
	Simulations	Ideal (AIC-D $\leq$ 2)	79	9	34	69
		Well (AIC-D $\leq$ 4)	86	11	75	87
		Sufficient (AIC-D $\leq$ 7)	91	15	83	90
		No Skill (AIC-D $>$ 10)	6	80	14	7

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020



Hydrology and Earth System Sciences
Discussions
Open Access
EGU

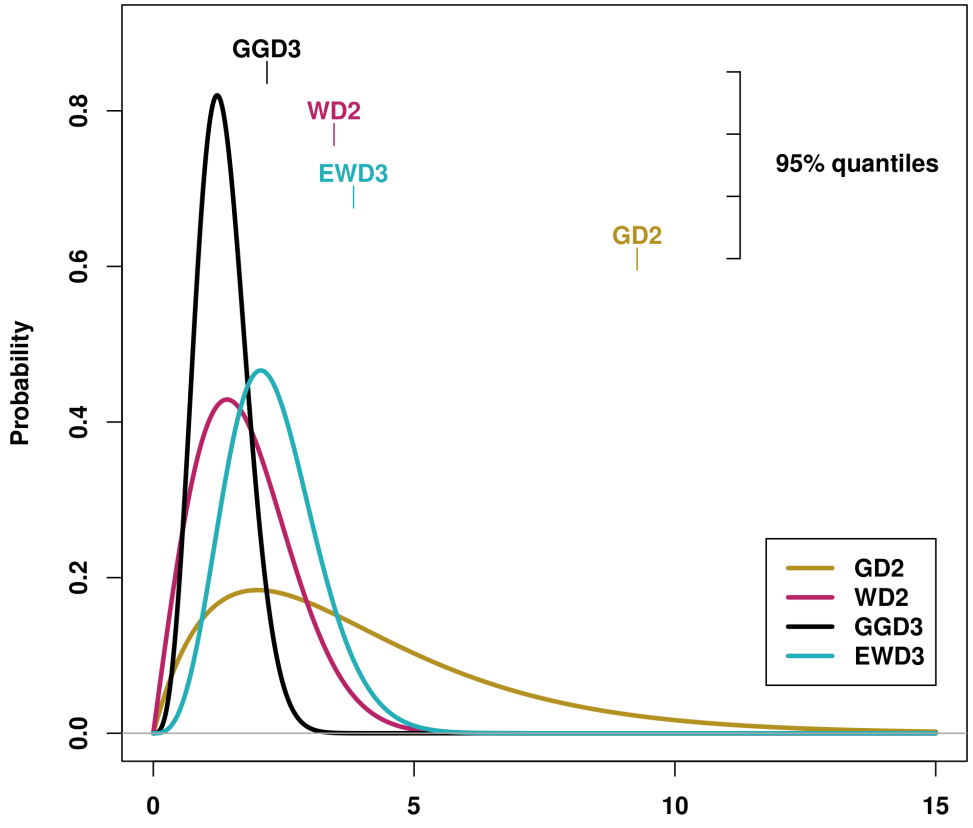


Figure 1. Candidate Distribution functions whose performance is investigated in this study: the 2-parameter gamma distribution (GD2), the 2-parameter Weibull distribution (WD2), the 3-parameter generalized gamma distribution (GGD3) and the 3-parameter exponentiated Weibull distribution (EWD3). Displayed are examples of those PDFs for $\sigma = \gamma(= \alpha) = 2$ and their corresponding 95 % quantiles.

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

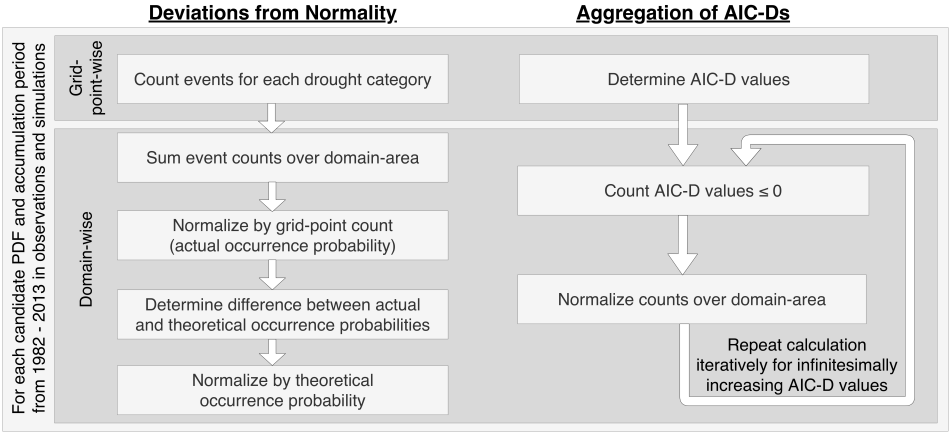


Figure 2. Flow chart of methods to aggregate deviations from $\mathcal{N}_{0,1}$ (**left**) and AIC-D frequencies (**right**) over domains.

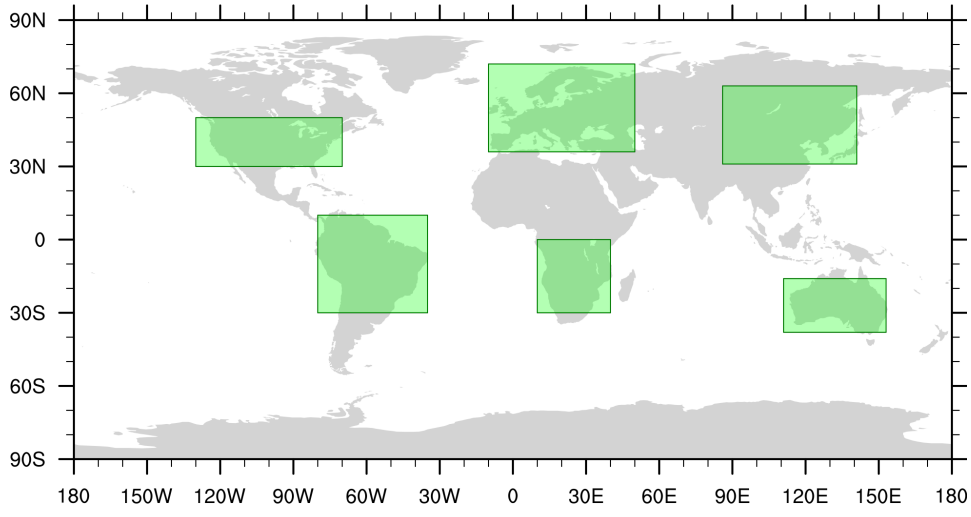


Figure 3. Borders of regions examined in this study.

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

Hydrology and Earth System Sciences
Discussions

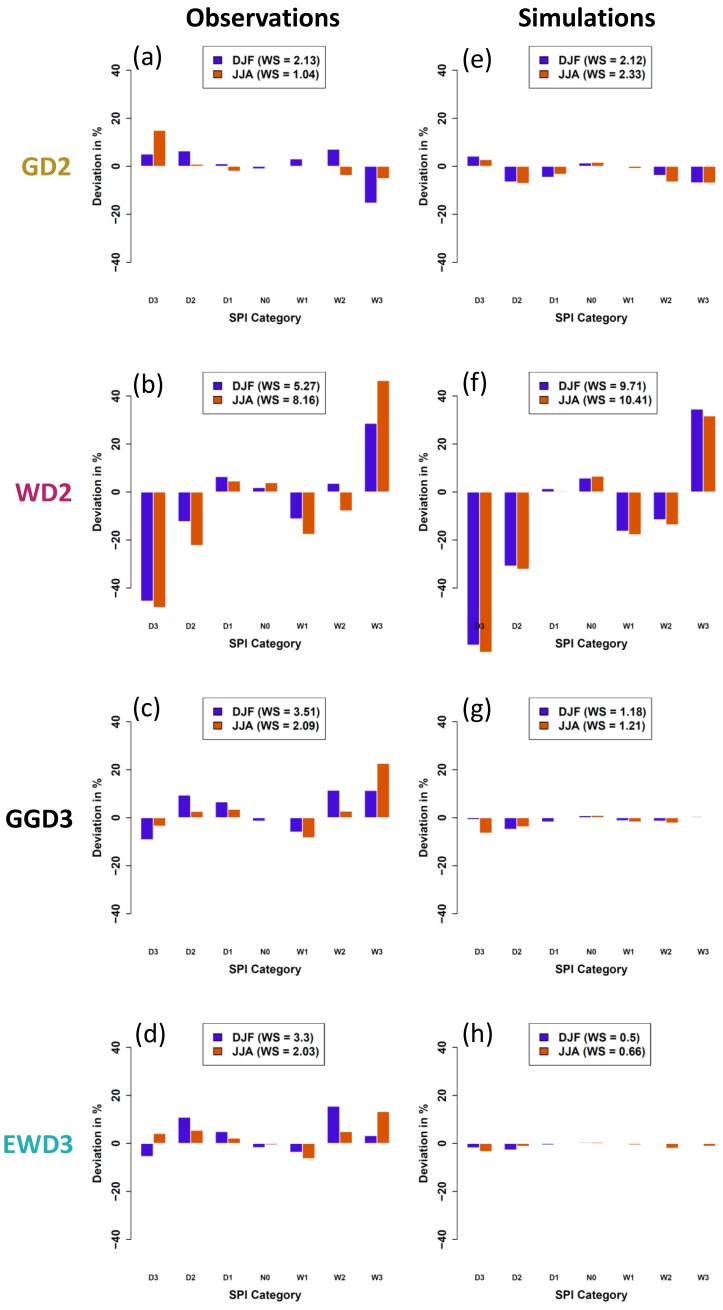
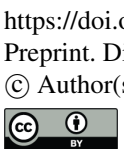


Figure 4. Deviations from $\mathcal{N}_{0,1}$ for observed (**left**) and modeled (**right**) SPI time-series. SPI time-series are derived by using the simple 2-parameter gamma distribution (GD2, **top row**), the simple 2-parameter Weibull distribution (WD2, **second row**), the 3-parameter generalized gamma distribution (GGD3, **third row**), and the 3-parameter exponentiated Weibull distribution (EWD3, **bottom row**). The legends depict the weighted sum (WS) of deviations from $\mathcal{N}_{0,1}$ over all SPI categories weighted by their respective theoretical occurrence probability.

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020



Hydrology and Earth System Sciences
Discussions

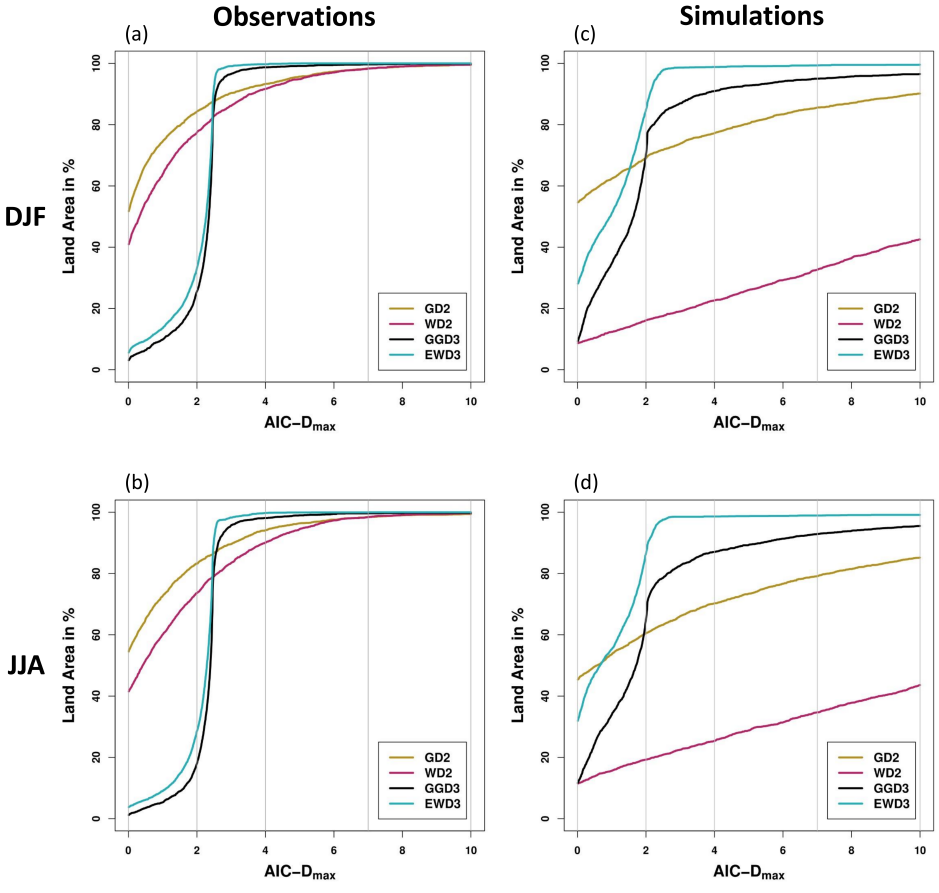


Figure 5. AIC-D frequencies: percentages of global land grid-points in which each distribution function yields AIC-D values that are smaller than or equal to a given AIC-D$_{max}$ value. AIC-D frequencies are displayed for each candidate PDF for observations (**left**) and simulations (**right**) during DJF (**top**) and JJA (**bottom**).

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

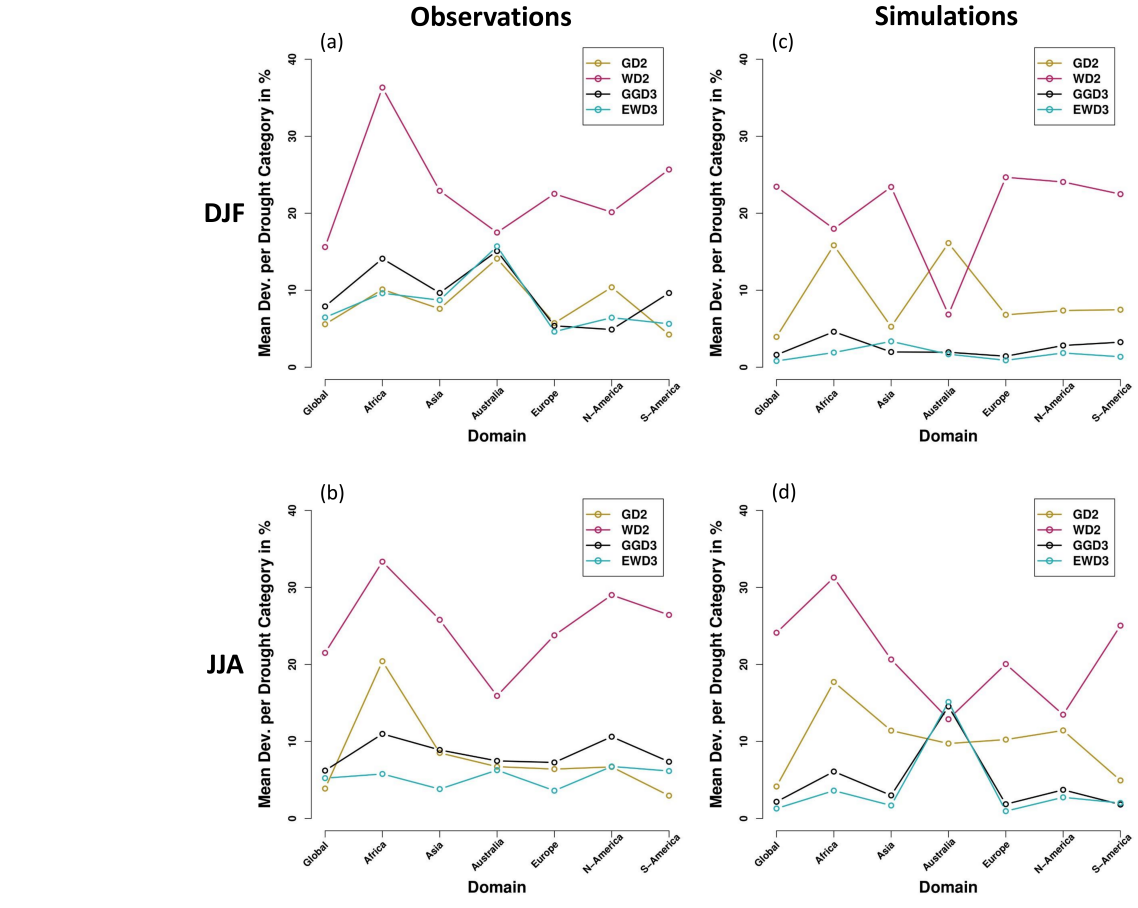


Figure 6. Mean deviations from $\mathcal{N}_{0,1}$ per SPI category for each investigated domain for observations (**left**) and simulations (**right**) during DJF (**top**) and JJA (**bottom**).

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

Hydrology and Earth System Sciences
Discussions

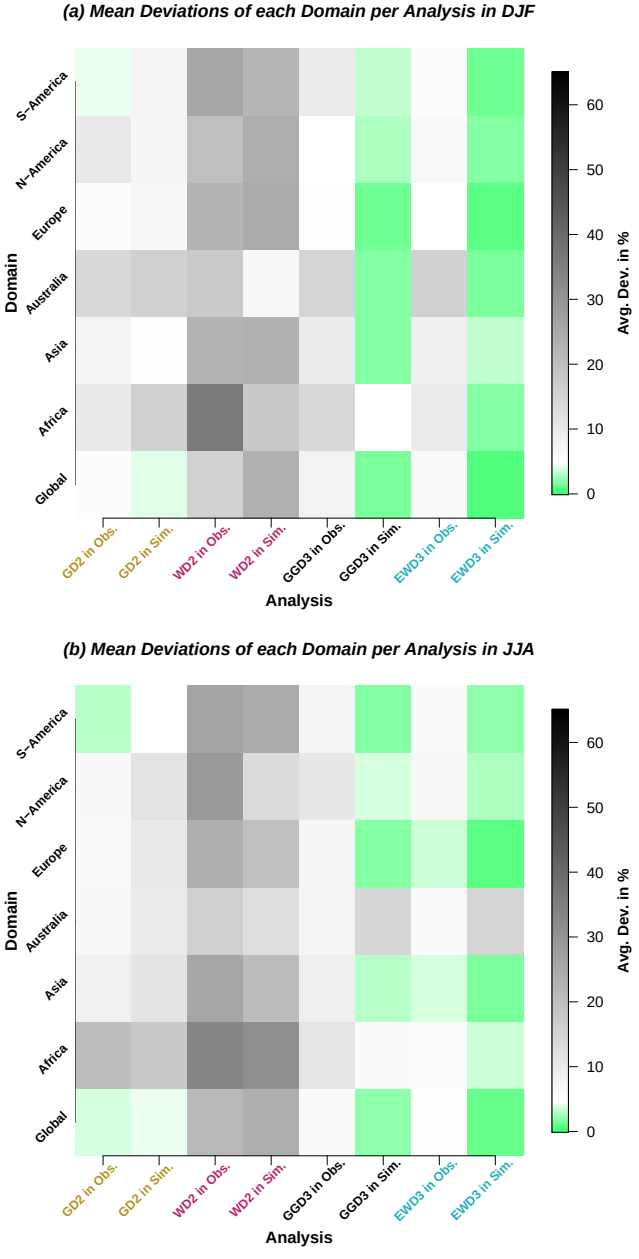


Figure 7. Mean deviations from $\mathcal{N}_{0,1}$ per SPI category during DJF (**a**) and JJA (**b**). Mean deviations are displayed for each investigated domain and each analyzed PDF for observations and simulations.

https://doi.org/10.5194/hess-2019-614
Preprint. Discussion started: 2 January 2020

Hydrology and Earth System Sciences
Discussions

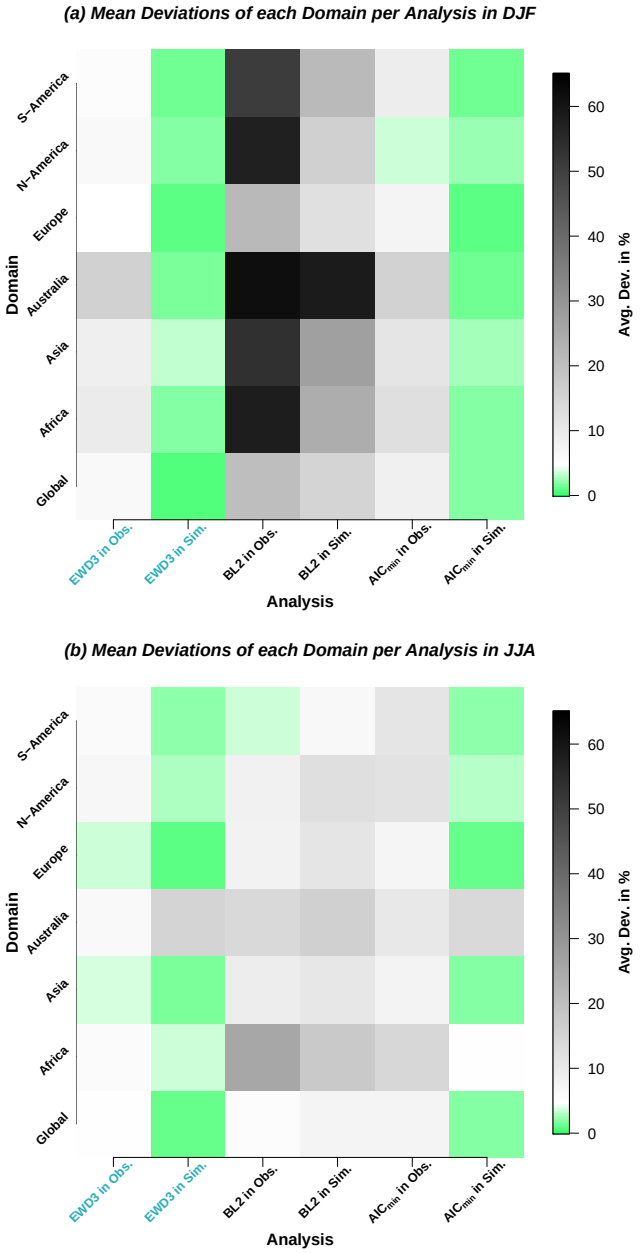


Figure 8. As in Fig. 7 but for the 3-parameter exponentiated Weibull distribution (EWD3) – the best performing candidate distribution function in this study –, a baseline which uses the 2-parameter gamma distribution (BL2) with a simpler parameter optimization than employed in our previous analysis, and a frequently proposed multi-PDF SPI calculation algorithm which uses in each grid-point that distribution function which yields in this respective grid-point the minimum AIC value (AIC$_{min}$-analysis which is denoted as AIC$_{min}$ in this figure).