Response to Reviewer 1

Patrick Pieper, André Düsterhus, and Johanna Baehr

June 15, 2020

We thank the reviewer for the effort of reviewing our work. His/Her comments have been very helpful in improving our manuscript. Below we answer point-by-point to each of the reviewer's comments and explain how the respective comment helped us to improve the manuscript. Reviewer's comments are printed in black and our responses are printed in blue. Line numbers in our response refer to the initially submitted manuscript.

One comment of the reviewer concerning the sample size in simulations caused us to perform a deeper sensitivity analysis on the ensemble size. In this process, a caveat to the drawn conclusions emerged. Therefore we include this sensitivity analysis to the results section and slightly adapted the drawn conclusions.

General comments

The SPI (Standardized Precipitation Index) is a commonly and widely used index to detect droughts based on precipitation data. It can be applied to several aggregation periods of precipitation, e.g.1 month, 3 months, 6 months etc., tailored to the different drought impacts (meteorological drought, agricultural drought, hydrological drought, . . .). In doing so, a distribution function is fitted on the precipitation data and transformed to a standard distribution. This gives the possibility to detect and compare droughts over time and space. The curtail point is the reproduction of the standard distribution by the transformed original distribution. Here, the paper investigate the suitability of four distribution functions with observed and forecasted precipitation data for the SPI. The goal of this paper is to propose one distribution function applicable to observed and forecasted precipitation totals globally for all useful aggregation periods. The paper is well and clear written and addresses the scientific question well.

Thank you for these kind comments and the effort of acquiring an in-depth understanding of our work.

Specific comments

You wrote in lines 164 to 167 that you use three different procedures to estimate the parameters of the distribution function. Therefor I expect to get analyses of three procedures times four distributions equals to twelve analyses per observations and simulations. You showed only one per distribution. Which of the procedures did you used finally to fit the parameters of the distribution functions? This is also relevant as you wrote in section 3.1.3 that the procedure of estimation the distribution function parameters could have an impact on the usability of the derived parameters.

Thank you for pointing out this unclear description of our methods. The three optimization methods referred to in lines 165 to 167 are used one after another. The goal is to find the most suitable parameters of the fit. To achieve this goal all available tools (all three optimization methods) are employed.

To avoid misunderstandings we performed the following changes to lines 161 to 170 in the manuscript: "(...) and dealt with later specifically. We estimate the parameters of our candidate PDFs in SPI's calculation algorithm with the maximum likelihood method [Nocedal and Wright, 1999] 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. The starting values identified from that scan are optimized with the simulated annealing method (SANN) [Bélisle, 1992]. Subsequently, these by SANN optimized starting values are again further optimized by a limited-memory modification of the Broyden-Fletcher-Goldfarb-Shanno (also known as BFGS) quasi-Newton method [Byrd et al., 1995]. If the BFGS quasi-Newton method leads to a convergence of the parameters of our candidate PDF, we achieve our goal and end the optimization here. If the BFGS quasi-Newton method does not lead to a convergence of the parameters of our candidate PDF, then we circle back to the starting values optimized by SANN and optimize them again further but this time with the Nelder-Mead method [Nelder and Mead, 1965]. After identifying converging parameters, the probabilities of encountering the given precipitation totals are computed and transformed into cumulative probabilities ($G(x)$).

If neither the BFGS quasi-Newton nor the Nelder-Mead method leads to any convergence of the most suitable parameters of our candidate PDFs, then we omit these grid-points where convergence is not achieved. For the gamma, Weibull, and exponentiated Weibull distribution, non-converging parameters are rare exceptions and only occur in a few negligible grid-points. For the generalized gamma distribution, however, non-convergence appears to be a more common issue and occurs in observations as well as in simulations in roughly every fifth grid-point of the global land area. This shortcoming of the generalized gamma distribution needs to be kept in mind when concluding its adequacy in SPI's calculation algorithm.

Since PDFs that describe the frequency distribution of precipitation totals are required to be only defined for the positive real axis, (...)"

Do you exclude grids without converging parameter fits from the further analysis or to you use another procedure to estimate the parameters? Line 167/168

We excluded from our analysis those grid-points where we do not achieve any convergence. We also excluded grid-points where zero-precipitation events occurred more than one-third of the times in our time-period (see lines 188 to 189). Grid-points excluded through both of these reasons are mainly located in the Sahara. In the process of checking grid-points excluded from the analysis, we realized a misleading description in the manuscript concerning the excess of zero-precipitation events. While the simulated precipitation time-series of all ensemble members (n=310) exhibits in 3.68% of the global land grid-points too often (more than 103 times) zero-precipitation events, only a single grid-point (located in the Sahara) exhibits zero-precipitation events too often (more than 10 times) in observations (n=31). Barring one exception, all of the grid-points which exhibit zero-precipitation events too often in simulations are located in the Sahara and the Arabian Peninsula (9°N – 44°N; 16°E – 69°W). The only exception is one grid-point which is is located in the Nevada desert.

We clarified this asymmetry between observations and simulations in lines 189 to 191: "This limitation restricts the SPI calculation in simulations over the Sahara and the Arabian Peninsula for accumulation periods of 1- and 3-months, (...)"

Your sample sizes differ by a factor of ten between observations and forecasts (e.g. lines 198 or 277). In line 277, you wrote that the reliability of the parameters depends on the sample size and is therefore better for the modelled than for the observed data. Nevertheless, if you analyse the usability of distribution functions for the SPI, you should have parameter estimations with the same reliability. I propose to repeat the analysis with only one ensemble member and add that to the paper and add a short analysis on the impact of the available amount of data to the reliability of the SPI.

Thank you for this excellent idea. As a consequence of our focus on seasonal predictions (which heavily rely on the entire ensemble space), we did not recognize the possibility to potentially widen our conclusions through a sensitivity analysis of the sample size. As it turns out, differences between observations and simulations mostly evaporate while their main distinction results from the sample size. In contrast to observations, the sample size can easily be expanded or condensed in simulations through the employment of additional/fewer ensemble realizations.

EWD3 outperforms GD2 for a sample size of 31 years in simulations and observations (Table I). The better performance of EWD3 relative to GD2 is particularly important in those grid-points where GD2 does not perform well (AIC-D $\geq$ 4). EWD3 displays such an erroneous performance in virtually no

Table I. As in Table 3, but the evaluation of simulations bases on a single ensemble member. Observations are identical to Table 3.

SPI Period	Realization	AIC-D category	GD2	WD2	GGD3	EWD3
3-Months	Observations	Ideal (AIC-D $\leq$ 2)	**84**	**76**	**22**	**31**
		Well (AIC-D $\leq$ 4)	**94**	**91**	**98**	**100**
		Sufficient (AIC-D $\leq$ 7)	**98**	**98**	**100**	**100**
		No Skill (AIC-D $>$ 10)	**1**	**0**	**0**	**0**
	Single Ensemble Member	Ideal (AIC-D $\leq$ 2)	**83**	**76**	**19**	**28**
		Well (AIC-D $\leq$ 4)	**93**	**91**	**98**	**100**
		Sufficient (AIC-D $\leq$ 7)	**98**	**98**	**100**	**100**
		No Skill (AIC-D $>$ 10)	**1**	**0**	**0**	**0**

grid-point. While these results still support our overall conclusions, it is evident that 2-parameter distribution functions can perform distinctly better in simulation than initially expected. The 2-parameter PDFs perform equally between observations and simulations. However, the 2-parameter PDFs also perform still worse than the 3-parameter PDFs. Yet, the insights gained from Table I also expose the question concerning the sensitivity of candidate PDFs' performances to the sample size.

Table II. As in Table 3, but with a focus on the sensitivity of the ensemble/sample size in simulations.

SPI Period	Ensemble Size	AIC-D category	GD2	WD2	GGD3	EWD3
3-Months	2	Ideal (AIC-D ≤ 2)	**78**	**56**	**43**	**57**
		Well (AIC-D ≤ 4)	**87**	**74**	**96**	**99**
		Sufficient (AIC-D ≤ 7)	**94**	**90**	**98**	**100**
		No Skill (AIC-D > 10)	**3**	**4**	**1**	**0**
	3	Ideal (AIC-D ≤ 2)	**77**	**45**	**53**	**69**
		Well (AIC-D ≤ 4)	**86**	**61**	**96**	**99**
		Sufficient (AIC-D ≤ 7)	**93**	**79**	**99**	**100**
		No Skill (AIC-D > 10)	**4**	**10**	**1**	**0**
	4	Ideal (AIC-D ≤ 2)	**75**	**38**	**59**	**74**
		Well (AIC-D ≤ 4)	**84**	**50**	**95**	**99**
		Sufficient (AIC-D ≤ 7)	**90**	**67**	**98**	**100**
		No Skill (AIC-D > 10)	**7**	**19**	**2**	**0**
	5	Ideal (AIC-D ≤ 2)	**74**	**31**	**63**	**79**
		Well (AIC-D ≤ 4)	**82**	**42**	**94**	**99**
		Sufficient (AIC-D ≤ 7)	**89**	**57**	**97**	**99**
		No Skill (AIC-D > 10)	**7**	**30**	**2**	**0**
	6	Ideal (AIC-D ≤ 2)	**73**	**27**	**64**	**80**
		Well (AIC-D ≤ 4)	**81**	**36**	**93**	**99**
		Sufficient (AIC-D ≤ 7)	**88**	**50**	**96**	**99**
		No Skill (AIC-D > 10)	**9**	**37**	**2**	**0**
	7	Ideal (AIC-D ≤ 2)	**70**	**25**	**66**	**81**
		Well (AIC-D ≤ 4)	**78**	**33**	**92**	**98**
		Sufficient (AIC-D ≤ 7)	**86**	**45**	**96**	**99**
		No Skill (AIC-D > 10)	**10**	**43**	**2**	**1**
	8	Ideal (AIC-D ≤ 2)	**69**	**21**	**67**	**83**
		Well (AIC-D ≤ 4)	**77**	**29**	**91**	**98**
		Sufficient (AIC-D ≤ 7)	**85**	**39**	**95**	**99**
		No Skill (AIC-D > 10)	**11**	**49**	**3**	**1**
	9	Ideal (AIC-D ≤ 2)	**66**	**20**	**67**	**85**
		Well (AIC-D ≤ 4)	**76**	**27**	**90**	**99**
		Sufficient (AIC-D ≤ 7)	**84**	**36**	**95**	**99**
		No Skill (AIC-D > 10)	**12**	**53**	**3**	**1**

3-parameter PDFs benefit because of their increased complexity more than 2-parameter PDFs from an increased sample size which is realized by additional ensemble members (Table II). Consequently, reducing the ensemble size levels the playing field between 2- and 3-parameter PDFs. While a sample size of 31 years suffices EWD3 to outperform GD2, the margin by which EWD3 outperforms GD2 increases with a further increase in sample size.

Because of these insights, we rectified several statements in the manuscript which imply that 2-parameter PDFs are unable to sufficiently describe simulated precipitation. Instead, we emphasize that – despite the increased need of samples to fit 3 parameters – the 3-parameter distribution functions perform better than the 2-parameter PDFs among our candidate PDFs. This improved performance is already apparent for roughly 30 events and logically becomes more distinct with increasing sample size.

In view of these insights, we created subsection 3.1.4 (in between lines 562 and 563) in which we discuss Table I and Table II:

"3.1.4 Sensitivity to Ensemble Size

So far, we used all ensemble members at once to fit our candidate PDFs onto simulated precipitation. That improves the quality of the fit. In this section, we first analyze a single ensemble member and investigate subsequently the sensitivity of our candidate PDFs' performance on the ensemble size. In doing so, we properly disentangle the difference between observations and simulations from the impact of the sample size.

As before, 3-parameter candidate distribution functions also perform for a single ensemble simulation better than 2-parameter PDFs (Table I). For a single ensemble member, the difference by which 3-parameter PDFs out-perform 2-parameter PDFs reduces considerably relative to the entire ensemble simulations (compare Table I against Table 3), though. In contrast to Table 3, all of our candidate distribution functions perform similarly between a single ensemble simulation and observations. In contrast to our previous results (e.g. when analyzing weighted sums of deviations from $\mathcal{N}_{0,1}$), modeled and observed precipitation distributions now seem sufficiently similar. Reducing the sample size for the fit by a factor of ten leads to more homogeneous performances of all candidate PDFs in simulations. As a reminder, AIC-D frequencies as depicted in Table I measure only relative performance differences. Consequently, our 2-parameter candidate PDFs do not actually perform better with fewer data. Instead, limiting the input data to a single ensemble member impairs our 3-parameter candidate PDFs stronger than our 2-parameter candidate PDFs. Irrespective of the realization, GD2 performs erroneously for 31 samples (apparent in grid-points which display AIC-D values beyond 4). Despite the need for more information, 31 samples suffice EWD3 to fix GD2's erroneous performances in both analyzed realizations.

In a next step, we isolate and investigate the improvement of the fit by an increasing sample/ensemble size. As a consequence of limited observed global precipitation data, we neglect observations and their differences to simulations in this remaining section. During this investigation, we reanalyze Table I while iteratively increasing the ensemble (sample) size for the fit (and the AIC-D calculation). Irrespective of the ensemble size, EWD3 performs robustly with high proficiency (Table II). Further, the fraction of grid-points in which EWD3 performs ideal increases constantly. This is a consequence of EWD3's better performance relative to our 2-parameter candidate PDFs. Unfortunately, AIC-Ds can only compare models that base on an equal sample size without adhering to additional undesired assumptions. Thus, any direct analysis of each candidate PDF's improvement relative to its own performance for a single ensemble member is with AIC-D frequencies not feasible. Despite this caveat, Table II still indicates strongly that EWD3 benefits stronger from the increased sample size than any of our 2-parameter candidate distribution functions. The larger the sample size, the larger is the margin by which EWD3 outperforms GD2.

Despite requiring more data, our 3-parameter candidate PDFs perform already better for 31 samples. For 31 samples, we identify this better performance of 3-parameter candidate PDFs in observations and simulations. Further, since our 3-parameter candidate PDFs require more data to estimate optimal parameters, they benefit in simulations stronger from additional samples than our 2-parameter candidate PDFs. That benefit becomes apparent in a distinctly improved relative performance after multiplying the sample size through the use of additional ensemble members."

Moreover, we rewrote parts of section 3.1.1. In this process, we substituted lines 360 to 375 by: "In simulations, the fit onto 3-months precipitation totals is performed on all ten ensemble members at once. This 10-folds the sample size in simulations relative to observations. Presuming an imperfect fit for the 31 samples in observations, deviations from $\mathcal{N}_{0,1}$ are expected to reduce along our four candidate distribution functions as a result of 10-folding the sample size of their fit. Yet, GD2 does not benefit from 10-folding the sample size. GD2 performs similarly in observations and simulations (Fig. 4 (a) and (e)). In contrast, our 3-parameter PDFs display considerably smaller deviations from $\mathcal{N}_{0,1}$ in ensemble simulations than in observations (compare Fig. 4 (c) and (d) against (g) and (h)). Consequently, both 3-parameter candidate PDFs excel during both seasons in ensemble simulations (Fig. 4, (g) and (h)), while any distinction between both 3-parameter candidate distribution functions is still difficult. On the one side, different frequency distributions between observed and modeled precipitation totals might be one reason for this difference. On the other side, the fit of three parameters also requires more data than the fit of two. It is therefore sensible to expect that 3-parameter PDFs benefit stronger than 2-parameter PDFs from an increase in sample size. Are our 3-parameter candidate PDFs are better suited than our 2-parameter PDFs to describe modeled precipitation distributions? Or benefit our 3-parameter PDFs just stronger than 2-parameter PDFs from an increasing sample size?

We attempt to disentangle both effects (analyzing modeled, instead of observed, precipitation distributions, and increasing the sample size) for our 2-parameter candidate PDFs, next. If the 2-parameter PDFs are suited to be applied to modeled precipitation data, they should at least benefit to some extent from this multiplication of sample size. Despite expecting irregularities in the magnitude of these reductions, they should be notable for (...)"

Further, we also changed parts of section 5. Here, we substituted lines 681 to 692 by: "Irrespective of the accumulation period or the data-set, GD2 seems sufficiently suited to be employed in SPI's calculation algorithm in many grid-points of the globe. Yet, GD2 also performs erroneous in a non-negligible fraction of grid-points. These erroneous performances are apparent in observations and simulations for each accumulation period. More severely, GD2's erroneous performances decline further in ensemble simulations. Here, GD2 performs in a non-negligible fraction of grid-points also insufficient or even without any skill. In contrast, EWD3 performs for all accumulation periods without any defects, irrespective of the data-set. Despite requiring more data than 2-parameter PDFs, we identify EWD3's proficient performance for a sample size of 31 years in observations as well as in simulations. Further, ensemble simulations allow us to artificially increase the sample size for the fitting procedure by including additional ensemble members. Exploiting this possibility has a major impact on the performance of candidate PDFs. The margin, by which EWD3 outperforms GD2, further increases with additional ensemble members. Furthermore, EWD3 demonstrates proficiency also for every analyzed accumulation period around the globe. The accumulation period of 12-months poses in simulations the only exception. Here, EWD3 and GD2 both perform similarly well around the globe. Still, we find that 3-parameter PDFs are generally better suited in SPI's calculation algorithm than 2-parameter PDFs.

Given all the dimensions (locations, realizations, accumulation periods) of the task, our results suggest that the risk of underfitting by using 2-parameter PDFs is larger than the risk of overfitting by employing 3-parameter PDFs. We strongly advocate adapting the calculation algorithm of SPI and the therein use of 2-parameter distribution functions in favor of 3-parameter PDFs. Such an adaptation is (...)"

Aside, we clarified the following statements of the manuscript:

We changed the wording from "simulations" to "ensemble simulations" in the following lines: 13, 362, 432, 450, 458, 495, 513, 569, 590, 665, 669, 693

We substituted the sentence in lines 462 to 464 by: "(...) However, the results justify the necessity for this increased complexity – GD2 performs erroneously in 26% (6%), insufficiently in 18% (2%), and without any skill in 12% (1%) of the global land area in ensemble simulations (observations). The risk of underfitting (...)"

We included the following paragraph in between lines 622 and 623: "Overall our 3-parameter candidate PDFs perform better than investigated 2-parameter

candidate PDFs. Despite requiring more data, a sample size of 31 years suffices our 3-parameter candidate PDFs to outperform our 2-parameter candidate PDFs in simulations and observations. Further, our 3-parameter candidate PDFs greatly benefit from an increase in the sample size in simulations. In simulations, such a sample size sensitivity analysis is feasible by exploiting different counts of ensemble members. Whether 3-parameter PDFs would benefit similarly from an increased sample size in observations is likely but ultimately remains speculative because trustworthy global observations of precipitation are temporally too constrained for such a sensitivity analysis."

We recalculated the counts in lines 656 to 659. They now read as follows: "Moreover, these thresholds show a robust statistical basis in terms of being equally represented over all 320 analyzed evaluations in this study (all entries of Table 3, Table 4. Table 5, and Table 6). Across all 80 analyses (all rows of Table 3, Table 4, Table 5, and Table 6), the four candidate PDFs perform insufficiently 132 times, while they perform with substantial (average) confidence 130 (58) times."

Lines 282 to 285: In this paragraph is no transition from absolute to relative AIC, which need to be improved. In addition, the index i is not well described.
Thank you for revealing this unclear description.

We changed lines 280 to 287 to: "(...) penalizes candidate PDFs based on their parameter-count. The best-performing distribution function attains the smallest AIC value 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. Thus, we use relative AIC differences (AIC-D) in our analysis. We calculate these AIC-D values for each PDF by computing the difference between its AIC value to the lowest AIC value of all four distribution functions. AIC-D values inform us about superiority in the optimal trade-off between bias and variance and are calculated as follows:

$$AIC\text{-}D_i = AIC_i - AIC_{min} \tag{1}$$

The index i indicates different distribution functions. AIC_{min} denotes the AIC value of the best-performing distribution function.
For our analysis, AIC-D values are well suited (...)"

Lines 224 to 226: Do you avoid parameters in the GGD3 to become GD2 or WB2?
We estimate the parameters of all PDFs independently by fitting the respective PDF to the precipitation data. Consequently, the two parameters that GGD3 share with GD2 (WD2) can differ. This is important because the third parameter of GGD3 (and EWD3) extends the phase-space spanned by the 2 parameters of GD2 (and WD2) into a third dimension. This third dimension provides opportunities for further optimizations – also for the first two parameters.

Thus, the new optimum for GGD3 in the three-dimensional phase-space does not need to be located along the normal above the optimal parameter-values of GD2 (or WD2) in the two-dimensional phase-space. The same is true in the other direction. The optimum location of parameters in the three-dimensional phase-space cannot simply be projected onto any two-dimensional phase-space. Instead, the location in the two-dimensional phase-space needs to be identified by properly optimizing the estimated fitting parameters independently.

To avoid misunderstandings, we clarified this point by inserting the following description at the end of the second paragraph of that section at line 211: "The optimization of this second shape parameter also requires the re-optimization of the first two parameters. The fitting procedure of 3-parameter PDFs needs therefore considerable more computational resources than the fitting procedure of 2-parameter distribution functions."

Section 2.7: Your region are large enough to cover several precipitation regimes in one region. I propose to reduce the size of the regions and select regions with known good/bad model performance and different precipitation regimes.

As of yet, the analysis is condensed enough to display the regional results in Figures 6-8 in single plots. Such a visualization helps to convey the results of our analysis. Until now, we presumed our results to be sufficiently robust so that the exact borders of our regions would neither distinctly alter our results nor our conclusions. Aside, the analyzed regions need to encompass several grid-points as explained in lines 322 to 327. Adhering to the *law of large numbers* is crucial for the statistical analysis performed for reach region. That being said, one can still argue for smaller regions. However, such a dispute is subjective as described in lines 330 to 324. Resolving this dispute would lead to an entirely new analysis which is beyond the scope of this investigation.

Irrespective of resolving this dispute in general, your proposal also triggered our curiosity concerning our presumption about the spatial robustness of our results and conclusions. Therefore, we tested the analysis for a region with exceptionally good performance of MPI-ESM-LR in predicting precipitation and SPI: the *North Region* of Brazil (0°– 8°S; 40°W – 60°W). As a side note, examples of poor model performance are already included in the results (e.g. the entire European continent). For the *North Region* of Brazil, we repeated Figure 4 and Table 3 of our analysis and display these results in Figure I and Table III.

Table III. As in Table 3, but solely for the *North Region* of Brazil (0°– 8°S; 40°W – 60°W).

SPI Period	Realization	AIC-D category	GD2	WD2	GGD3	EWD3
3-Months	Observations	Ideal (AIC-D $\leq$ 2)	**69**	**76**	**12**	**35**
		Well (AIC-D $\leq$ 4)	**84**	**89**	**92**	**100**
		Sufficient (AIC-D $\leq$ 7)	**100**	**97**	**100**	**100**
		No Skill (AIC-D $>$ 10)	**0**	**0**	**0**	**0**
	Simulations	Ideal (AIC-D $\leq$ 2)	**13**	**50**	**70**	**93**
		Well (AIC-D $\leq$ 4)	**13**	**53**	**84**	**100**
		Sufficient (AIC-D $\leq$ 7)	**16**	**77**	**87**	**100**
		No Skill (AIC-D $>$ 10)	**78**	**21**	**8**	**0**

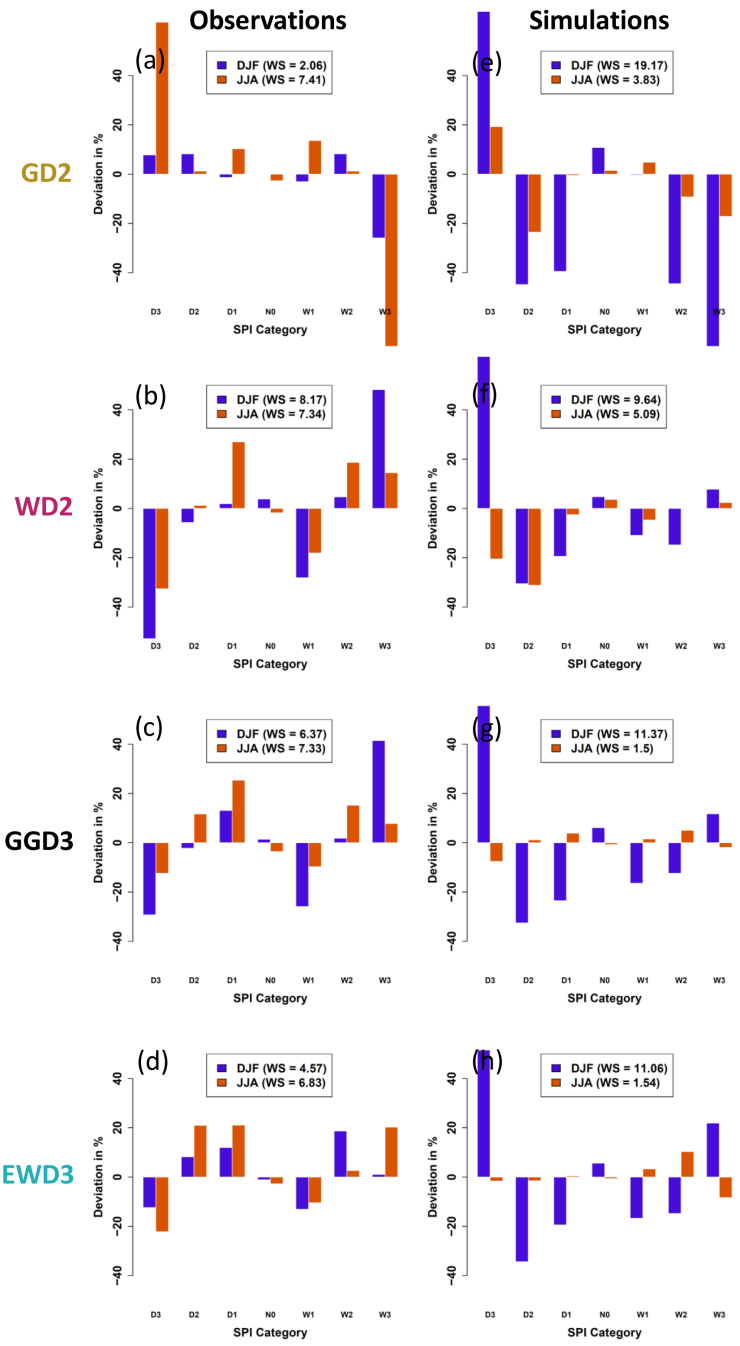

Figure I. As in Figure 4, but solely for the *North Region* of Brazil (0°– 8°S; 40°W – 60°W).

On the one hand, these results further corroborate our conclusions. EWD3 is distinctly better suited than the other candidate PDFs to describe precipitation; also when analyzed over such a small region (see Table III). On the other hand, the results also exemplify the importance of adhering to the *law of large numbers* in our analysis and its sensibility in terms of the extend of analyzed regions; specifically when evaluating deviations from $\mathcal{N}_{0,1}$ (see Figure I).

Line 355: How do you calculate the "weighted sum"? Please add a description.

Thank you for pointing out this lack of clarity.

We changed the sentence to: "Therefore, the weighted sum (weighted by the theoretical occurrence probability of the respective SPI class (Table 2)) over the absolute values of deviations from $\mathcal{N}_{0,1}$ along all SPI categories is lowest for GD2 in both analyzed seasons (see legend in Fig. 4, (a)–(d))."

We also added another description in line 377: "Therefore, we weigh each class' deviation from $\mathcal{N}_{0,1}$ by the theoretical occurrence probability (see Table 2) of the respective class and analyze weighted deviations from $\mathcal{N}_{0,1}$."

Line 574: You stated a phase transition of the SPI at 3 months precipitation accumulation. However, I cannot see it in Figure 4. What do you mean with phase transition?

Thank you for calling the misleading phrasing to our attention. In Table 4, WD2 performs better than GD2 in observation for an accumulation period of 1-month. For accumulation periods of 6-months and longer GD2 performs better than WD2 in observations.

We see how referring to this behavior as phase transition might be misleading and changed the paragraph to: "In agreement with prior studies [Stagge et al., 2015, Sienz et al., 2012], we also identify the apparent performance shift 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, neither 3-parameter candidate PDF displays such a shift in its performance. Both 3-parameter PDFs perform for accumulation periods shorter and longer than 3-months similarly well."

We also changed the sentence from line 600 to 602 in which we also used the wording *phase transition*. The reworded sentence reads as follows: "The emergence of this proposal stems from a focus on 2-parameter PDFs that exhibit a shift in their performance which depends on the scrutinized accumulation period."

Section 4: Do you compare the same number of grid cells for observations and forecasts? In addition, do you compare the same grid cells? I assume

Table IV. Percent of covered global land grid-points for each PDF in each realization and for each investigated season. Main differences between observations and simulations result from the Sahara and the Arabian Peninsula not being covered in simulations.

		GD2	WD2	GGD3	EWD3
DJF	**Simulations**	96.27	96.27	82.69	95.23
	Observations	100.00	100.00	82.33	97.16
	Ratio: Sim./Obs.	0.9627	0.9627	1.0043	0.9801
JJA	**Simulations**	96.33	96.33	77.9	95.55
	Observations	99.97	99.97	84.79	96.87
	Ratio: Sim./Obs.	0.9636	0.9636	0.9187	0.9864

different sets of selected grid cells for your analyses can have an impact on the results.

Thank you for this well-founded remark. Because of this comment and an earlier comment of yours, we double-checked the omitted grid-points again. We omit grid-points because of excessive zero-precipitation events and as a result of not achieved convergences. Consequently, the analyzed grid-points differ. They differ between simulations and observations because both realizations exhibit a different count of grid-points which exhibited too many (more than one-third) zero-precipitation events. Additionally, the analyzed grid-points also differ across the analyzed PDFs because the count of grid-points in which convergence is not achieved varies PDF-dependent. It is noteworthy, that (for GD2, WD2, and EWD3) the variations in analyzed grid-points are dominated by excessive zero-precipitation events; rather than being caused by non-converging parameters. Averaged over both seasons, 3.68% (0%) of land grid-points are PDF-independently excluded through an excessive count of zero-precipitation events in simulations (observations). In contrast, the total percentage of omitted grid-points per PDF (as a result of non-convergence and excessive zero-precipitation events) are displayed in Table IV.

We excluded non-converging grid-points only for the specific PDF, the specific season, and only in the specific realization (observation or simulation). This results in slightly different coverages for each PDF and each realization (see Table IV). Admittedly, GGD3's coverage can be described as inferior compared to the other candidate PDFs. However, this inferior performance does not impact our conclusions, but rather affirms the conclusion that EWD3 is better suited than GGD3. Additionally, the similar coverages of the other three candidate PDFs support the claim of a leveled playing field in our analysis. Thus, repeating the analysis for those grid-points where the fits of GD2, WD2, and EWD3 mutually converge is highly unlikely to change the result. Moreover, limiting the analyzed grid-points to those grid-points in which GGD3's calculation algorithm finds converging parameters would artificially reduce the reliability of the comparison between GD2, WD2, and EWD3. This impact would be similarly undesirable.

Yet, we do agree that different sets of grid-points can principally impact our analysis. Therefore, we analyzed Table 3 again to ascertain our assumption of a negligible impact on our analysis:

Table V. As in Table 3, but only for those grid-points which are mutually covered in simulations and observations by each PDF. Note: Grid-point coverage still differs between DJF and JJA. Depicted is the mean over both seasons.

SPI Period	Realization	AIC-D category	GD2	WD2	GGD3	EWD3
3-Months	Observations	Ideal (AIC-D ≤ 2)	**84**	**74**	**19**	**30**
		Well (AIC-D ≤ 4)	**94**	**90**	**98**	**100**
		Sufficient (AIC-D ≤ 7)	**98**	**98**	**100**	**100**
		No Skill (AIC-D > 10)	**0**	**0**	**0**	**0**
	Simulations	Ideal (AIC-D ≤ 2)	**64**	**18**	**68**	**86**
		Well (AIC-D ≤ 4)	**73**	**24**	**89**	**99**
		Sufficient (AIC-D ≤ 7)	**82**	**34**	**94**	**99**
		No Skill (AIC-D > 10)	**12**	**56**	**4**	**1**

Table VI. As in Table 3, but only for those grid-points which are mutually covered in simulations and observations by GD2, WD2, and EWD3. Note: Grid-point coverage still differs between DJF and JJA. Depicted is the mean over both seasons. Remark: Grid-points analyzed for GGD3 are the ones from Table 3 minus those grid-points which are not mutually covered by GD2, WD2, and EWD3.

SPI Period	Realization	AIC-D category	GD2	WD2	GGD3	EWD3
3-Months	Observations	Ideal (AIC-D ≤ 2)	**84**	**75**	**20**	**30**
		Well (AIC-D ≤ 4)	**94**	**91**	**98**	**100**
		Sufficient (AIC-D ≤ 7)	**98**	**98**	**100**	**100**
		No Skill (AIC-D > 10)	**0**	**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**

Averaged over all 32 entries, Table VI (Table V) differs on average by just 0.16 (0.34) percentage points from Table 3. The largest difference emerges in observations for GGD3 in the ideal category which deviates in Table VI (Table V) by 2 (3) percentage points from Table 3. In conclusion, we consider our assumption of a negligible impact on our analysis ascertained.

Lines 604/605: I think the investigations to the empirical cumulative distribution functions are very relevant for this topic and should be added to the paper or, at least, add a reference to the paper where you want to describe it.

We tried the empirical cumulative density function (ECDF) but quickly realized its shortcoming: Its discrete nature is too coarse for the task at hand which results in a massive dependence of possible SPI-values on the sample size. As explained in lines 323 to 328, the crucial performance requirement demands that deviations from $\mathcal{N}_{0,1}$ spatially balance each other sufficiently quickly. For SPI time-series derived with an ECDF, however, these deviations will never balance each other but aggregate with each additional grid-point. In the example from line 325, SPI time-series derived with an ECDF would not lead in a single grid-point to an extremely dry/wet event and would lead in each grid-point to exactly one severely dry/wet event during a 31-year time-series. Thus, for each grid-point over which we aggregate, we would add 0.7 missing extreme events and 0.4 missing severe events on both tails of the distribution.

To prevent any confusion, we adjusted the ending of the sentence in line 607 and included another explanation: "(...) We checked this approach which proved to be too coarse because of its discretized nature (not shown). As a result of its discretized nature, the analyzed sample size prescribes the magnitude of deviations from $\mathcal{N}_{0,1}$. Consequently, these deviations are spatially invariant and aggregate with each additional grid-point. Thus, deviations from $\mathcal{N}_{0,1}$ will not spatially balance each other."

Section 5: The base problem, from my point of view is, that the models are not able to reproduce the observed precipitation distribution function and procedures developed on observed data need to be adapted to be applied to model data (the GD2 performs well on the observed data). That is the base of your research and you should comment on this here or in the introduction.

Thank you for pin-pointing this motivation. This is exactly the motivation we had in mind which triggered us to conduct this analysis. We thought that we sufficiently pointed that out. However, after re-reading the respective paragraphs, we also realized that it comes a bit short. Therefore, we adjusted the Introduction and Section 5 and address this motivation in separate, stand-alone paragraphs:

To adjust the Introduction, we split the paragraph from lines 118 to 134. The changes read as follows: "SPI calculation procedures were developed for observed precipitation data. Since models do not exactly reproduce the observed precipitation distribution, these procedures need to be tested and eventually

adapted before being applied to modeled data. Here, we aspire to identify an SPI calculation algorithm that coherently describes modeled and observed precipitation (i.e. describes both modeled and observed precipitation distributions individually and concurrently). While testing SPI's calculation algorithm on modeled precipitation data is usually neglected, such a test demands nowadays a similarly prominent role as the one for 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 close that gap, this study evaluates the performance of candidate distribution functions in an output of 10 ensemble members of initialized seasonal hindcast simulations.

In this study, we test the adequacy of the gamma, Weibull, generalized gamma, and exponentiated Weibull distribution in SPI's calculation algorithm. 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 complementary evaluations of their normality: (i) evaluating their normality in absolute terms by comparing 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) evaluating their normality relative to each other with Akaike's information criterion (AIC) which analytically assesses of 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. Observed and simulated precipitation is obtained from the monthly precipitation data-set of the Global Precipitation Climatology Project (GPCP) and the above mentioned initialized seasonal hindcast simulations, respectively. 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)."

To adjust Section 5, we inserted in between Lines 672 and 673 (at the start of the section) the following paragraph: "Current SPI calculation algorithms are tailored to describe observed precipitation distributions. Consequently, current SPI calculation algorithms are ineptly suited to describe precipitation distributions obtained from ensemble simulations. Also in observations, erroneous performances are apparent and well-known, but less conspicuous than in ensemble simulations. We propose a solution that rectifies these issues and improves the description of modeled and observed precipitation distributions individually as well as concurrently. The performance of 2-parameter candidate distribution functions is inadequate for this task. By increasing the parameter count of the candidate distribution function (and thereby also its complexity) a distinctly

better description of precipitation distributions can be achieved. In simulations and observation, the here identified best-performing candidate distribution function – the exponentiated Weibull distribution (EWD3) – performs proficiently for every common accumulation period (1-, 3-, 6-, 9-, and 12-months) virtually everywhere around the globe. Additionally, EWD3 excels when analyzing ensemble simulations. Its increased complexity (relative to GD2) leads to an outstanding performance of EWD3 when an available ensemble multiplies the sample size."

Figure 6: Can you add the global average, as for Figure 4, as an additional domain to this figure?

We agree that the global average belongs in this Figure. To avoid any confusion, we decided to prominently label the global average in the caption of the figure.

The caption now reads as follows: "Mean deviations from $\mathcal{N}_{0,1}$ per SPI category for the entire global land area and each investigated region. Results are depicted for observations (**left**) and simulations (**right**) during DJF (**top**) and JJA (**bottom**)."

Technical corrections

Lines 379/380: It was not clear what was set in relation to what. Please reword this part.

Corrected.

Reworded to: "Relative to observations, GD2's weighted deviations increase in simulations by more than 120% in JJA, while WD2's increase by more than 25% in JJA and 80% in DJF."

Line 527: I think you to refer to Figure 8 and not to Figure 7.

We do mean Figure 7.

To clarify this misunderstanding, we reworded the sentence to: "The comparison between the performance of our baseline against GD2's performance (compare Fig. 8 against Fig. 7) thus also indicates the impact of the meticulousness applied to the optimization of the same parameter estimation method."

Line 583: I think you want to refer to "GD2" instead of "GGD2" (typo).

We want to refer to GGD3. We corrected that typo and changed "GGD2" to "GGD3".

Figure 4: Add to the caption that it is for global average.

Added.

References

[Bélisle, 1992] Bélisle, C. J. (1992). Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^{\mathrm{d}}$. *Journal of Applied Probability*, 29(4):885–895.

[Byrd et al., 1995] Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

[Nelder and Mead, 1965] Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4):308–313.

[Nocedal and Wright, 1999] Nocedal, J. and Wright, S. J. (1999). Springer series in operations research. numerical optimization.

[Sienz et al., 2012] Sienz, F., Bothe, O., and Fraedrich, K. (2012). Monitoring and quantifying future climate projections of dryness and wetness extremes: Spi bias. *Hydrology and Earth System Sciences*, 16(7):2143.

[Stagge et al., 2015] Stagge, J. H., Tallaksen, L. M., Gudmundsson, L., Van Loon, A. F., and Stahl, K. (2015). Candidate distributions for climatological drought indices (spi and spei). *International Journal of Climatology*, 35(13):4027–4040.

[Svoboda et al., 2012] Svoboda, M., Hayes, M., and Wood, D. (2012). Standardized precipitation index user guide. *World Meteorological Organization Geneva, Switzerland*.

Response to Reviewer 2

Patrick Pieper, André Düsterhus, and Johanna Baehr

June 15, 2020

We thank Gabriel Blain for the effort of reviewing our work. Your comments have been very helpful in improving our manuscript. Below we answer point-by-point to each of your comments and explain how the respective comment helped us to improve the manuscript. Your comments are printed in black and our responses are printed in blue. Line numbers in our response refer to the initially submitted manuscript.

One of your comments concerning the complexity punishment by our employed information criterion caused us to reconsider our storyline. This reconsideration does not alter our conclusions. Yet, it simplifies for us to conclude which helps readers to follow our conclusions.

General comments

The manuscript "Global and regional performances of SPI candidate distribution functions in observations and simulations" proposes and new methodology to select candidates distributions for calculating the SPI; a widely used standardized drought index. The study is interesting and adds important information to the SPI literature because it evaluates the advantages and shortcomings of previous methodologies designed for the same purpose. It is also well written. So, it should be considered for publication.

Thank you for these kind comments and your endorsement.

Specific comments

L.105 The Shapiro-Wilk [...] "is unreliable to evaluate SPI normality (Naresh Kumar et al., 2009)". This is a very important statement, which now I tend to agree with. Please, provide further information regarding it.

First, we are pleased that we were able to convince you. Second, we thank you for pointing out this lack of depth in our introduction.

Goodness-of-Fit (GoF) tests are ill-suited to assess the normality of SPI time-series, because of their spatial aggregation in combination with their binary convention. To fully understand this interplay we start with SPI's calculation

procedure: (i) fit a candidate PDF onto precipitation, then (ii) Z-transform the fitted probabilities to SPI values. Because the choice of an appropriate candidate PDF is the key decision in this process, the initial fit of the candidate PDF onto precipitation should be scrutinized. GoF tests, however, measure the normality of the resulting SPI values. In theory, this switch of focus in the analysis only complicates its structure but should not impact its outcome: if the candidate PDF's fit is appropriate, then its estimated probabilities are appropriate. Thus, their exact equiprobability transformations to the standard normal variable Z are also appropriate.

Anyhow, this complicated structure blurs the view on the measure of interest: the fit of the candidate PDF (onto precipitation). Therefore, the following caveat easily arises unnoticed and is, thus, not properly dealt with. After losing sight of the actual measure of interest (the fit of the candidate PDF), the focus lays on the normality of SPI time-series. The intuitive tool to assess normality leads to GoF tests. The drawback of GoF test is the biased discrimination between the tails and the center of the distribution. GoF tests equally evaluate each value that contributes to the distribution. Such an evaluation assigns more weight to the center and almost no weight to the tails of the distribution. Yet, appropriately fitting the tails of precipitation distributions should logically be of paramount importance to any sensible candidate PDF employed in SPI's calculation algorithm (see also our argument against weighting deviations from $\mathcal{N}_{0,1}$ by the theoretical occurrence probability of the respective class in lines 244 to 251). But the complicated structure blurs the view from this consideration. Instead, GoF tests conveniently present an allegedly easy solution.

As seen in Fig. 1, deviations from $\mathcal{N}_{0,1}$ are smallest in the center and largest in the tails of the distribution. Candidate PDFs typically fit precipitation better for the center than for the tails of the distribution: its center counts more samples which translates to more weight in the optimization (e.g. by the maximum likelihood estimation). This behavior deludes GoF tests in the analysis of SPI normality. That delusion obscures the tails of the distribution from GoF tests. Nevertheless this delusion, despite this obscurity surfacing skepticism about the proper depiction of the tails of the distribution can still aggregate over many grid-points. This aggregated skepticism can still lead to a robust analysis if evaluated relative to the similarly obscured performance of other candidate PDFs (as shown by metrics such as AIC-D, and BIC-D). Anyhow, additionally aggravating for GoF tests is their convention to be interpreted binarily. As a consequence of this convention, SPI literature typically aggregates results of GoF tests over domains by counting rejections. This typical aggregation prevents surfaced skepticism to fully aggregate over many grid-points. The interplay of both caveats, the blurred tails of the distribution and the prevention of remaining skepticism to fully aggregate, leads to the conclusion that GoF tests are ill-suited to assess SPI normality. I.e. it is (admittedly more obvious but) similarly inept to round normally distributed ($\mathcal{N}_{0.1\pm\epsilon,0.1}$) variables to their nearest integer before calculating their mean to estimate ϵ.

This full explanation is too extensive for the scope of the introduction of

our publication. However, we do admit that only indicating problems with the binary nature of GoF tests and hinting at issues with their spatial aggregation might cut the story too short. To rectify this shortcoming, we split the paragraph (lines 106 -117). This allows us to elaborate on GoF tests (in)ability to evaluate SPI candidate distribution functions: "(...) which in turn is unreliable to evaluate SPI normality [Naresh Kumar et al., 2009].

The above-mentioned goodness-of-fit tests equally evaluate each value of SPI's distribution. Such an evaluation focuses on the center of the distribution because the center of any distribution contains per definition more samples than the tails. In contrast, SPI usually analyzes (and thus depends on a proper depiction of) the distribution's tails. Therefore, a blurred focus manifests in these goodness-of-fit tests. Moreover, the convention to binarily interpret the above-mentioned goodness-of-fit tests aggravates this blurred focus. Because of this convention, 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). This inability prevents any reasonable aggregation of limitations that surface despite the blurred focus. Thus, they are ill-suited to discriminate the best performing PDF out of a set of PDFs [Blain et al., 2018]. For SPI distributions 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. Because of the ill-fitting focus and the ill-suited convention of these goodness-of-fit tests, they are inept to identify SPI's best-performing candidate distribution function out of a set of PDFs.

In agreement with this insight, those studies, that rigorously analyzed candidate distribution functions, or investigate an appropriate test methodology for evaluating SPI candidate PDFs, consequently advocate the use of relative assessments: (...)"

While elaborating on the methodology to test the normality of SPI time-series, we realized a missing differentiation between the analysis of AIC-D frequencies and the analysis of deviations from $\mathcal{N}_{0,1}$ in the initial submission. The fact that both analyses complement each other comes a bit too short. Thus, we also rectified this shortcoming through the following changes to the manuscript:

We substituted a sentence from the abstract in lines 6 to 7 by: "Our normality comparison bases on a complementary evaluation. Actual against theoretical occurrence probabilities of SPI categories evaluate the absolute performance of candidate distribution functions. In contrast, Akaike's information criterion evaluates candidate distribution functions relative to each other while analytically punishing complexity. SPI time-series (...)"

We added another paragraph at the end of section 2.5 in between lines 293 and 294 which reads as follows: "The analysis of deviations from $\mathcal{N}_{0,1}$ assesses performances of candidate PDFs in absolute terms irrespective of the candidate PDF's complexity. In contrast, the AIC-D analysis evaluates the performance of candidate PDFs relative to each other while analytically punishing complexity.

Consequently, the AIC-D analysis cannot evaluate whether the best-performing candidate distribution function also performs adequately in absolute terms. In opposition, deviations from $\mathcal{N}_{0,1}$ encounter difficulties when evaluating whether an increased complexity from one PDF to another justifies any given improvement. Both analyses together, however, augment each other complementary. This enables us to conclusively investigate: (i) which candidate PDF performs best while (ii) ensuring adequate absolute performance and while (iii) constraining the risk of over-fitting."

We substituted three sentences in a paragraph of section 3.1.1 (lines 401 to 405) by: "It is noteworthy, that investigating deviations from $\mathcal{N}_{0,1}$ over the entire globe contains the risk of encountering deviations that balance each other in different grid-points with unrelated climatic characteristics. Until dealing with this risk, our analysis of deviations from $\mathcal{N}_{0,1}$ only indicates that three candidate PDFs (GD2, GGD3, and EWD3) display an adequate absolute performance. On the one hand, we can reduce that risk by analyzing deviations from $\mathcal{N}_{0,1}$ only over specific regions. This analysis safeguards our investigation by ensuring (rather than just indicating) an adequate absolute performance around the globe and is performed later. On the other hand, we first completely eliminate this risk 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. We investigate AIC-D frequencies first to evaluate whether GGD3 and/or EWD3 perform sufficiently better than GD2 to justify their increased complexities."

We added another paragraph at the end of section 3.1.1 (in between lines 475 and 476): "Among our candidate PDFs, EWD3 is obviously the best-suited PDF for SPI. Yet, we still need to confirm whether also EWD3's absolute performance is adequate. While the global analysis indicated EWD3's adequateness, the ultimate validation of this claim is incumbent upon the regional analysis."

We added another paragraph at the end of section 3.1.2 (in between lines 514 and 515): "The analysis of AIC-D frequencies proves that EWD3 is SPI's best distribution function among our candidate PDFs. Additionally, the regional investigation confirms the global analysis: the absolute performance of EWD3 is at minimum adequate in observations and ensemble simulations."

The Bayesian information criterion (BIC) is similar to the AIC. However, the BIC uses a different penalty for the number of parameters [ln(n) k]. Can the authors verify if the BIC leads to similar results as those of the AIC.

We thank you for this exciting idea. Whether we use AIC or BIC to punish candidate PDFs for their complexity does not change our conclusions. Most of our drawn conclusions from AIC-D frequencies bases on the behavior of candidate PDFs' coverages for AIC-D_{max} values larger than 10 (right edge of Figure 5). These conclusions are then substantiated by candidate PDFs' coverages for AIC-D_{max} values larger than 7. These coverages (for AIC-D_{max}/BIC-D_{max}

Table I. Complexity penalty of candidate PDFs assessed with AIC and BIC.

Information Criterion	**AIC**		**BIC**		**Difference** BIC-AIC	
Realization	Obs. (N=31)	Sim. (N=310)	Obs. (N=31)	Sim. (N=310)	Obs. (N=31)	Sim. (N=310)
2-param. PDFs	4.43	4.04	6.87	11.47	2.44	7.43
3-param. PDFs	6.89	6.08	10.3	17.21	3.41	11.13
Difference 3-2 param.	**2.46**	**2.04**	**3.43**	**5.74**	**0.97**	**3.7**

values ≤ 7) are insensitive to the magnitude of changes caused by altered complexity penalties (Table I)

What impacts our analysis is not the absolute, parameter- and sample size-dependent punishment of candidate PDFs (values in the center of Table I). Instead, only the penalty difference between 2- and 3-parameter PDFs that base on the same sample size matters (evaluate observations and simulations isolated in the last row of Table I).

Similar, altering the information criterion (from AIC to BIC) impacts our analysis through the penalty difference between BIC and AIC (last column of Table I). Here, the difference between 2- and 3-parameter PDFs that base on the same sample size matters again (evaluate observations and simulations isolated in the bottom- and rightmost cell of Table I). I.e. the additional margin by which 3-parameter PDFs need to further outperform 2-parameter PDFs in order to still be considered as better by the new information criterion. This margin (bottom- and rightmost cell in Table I) increases in observations (simulations) by 0.97 (3.7) when using BIC instead of AIC.

The robustness of our conclusions stems from the robustness of the candidate's coverages for large AIC-D_{max}/BIC-D_{max} values (≥ 7). In this AIC-D_{max} regime, the candidate PDFs' coverages are sufficiently robust concerning changes caused by altered complexity penalties (Fig. I). Comparing 2- against 3-parameter in Fig. 5 with AIC-D or BIC-D does not substantially change the evaluation of large AIC-D_{max}/BIC-D_{max} values (≥ 7). As a first-order approximation, we can compare in observations the coverages of 2-parameter PDFs at the AIC-D_{max} value of 7 against the coverages of 3-parameter PDFs at the AIC-D_{max} value 7.97 (we shift the line indicating the coverages of 3-parameter PDFs by 0.97 units to the right). Since the slope of that line is sufficiently flat, this shift does not impact the conclusions for large AIC-D_{max} values (≥ 7).

In observations (simulations), coverages of 3-parameter PDFs are highly sensitive to the change of the information criterion at AIC-D_{max}/BIC-D_{max} values smaller than approximately 4 (6) (compare in Fig I the top row against the bottom row). The first-order approximation outlined before (shifting the coverages of 3-parameter PDFs by 0.97 (3.7) units to the right in observations (simulations)), describes the changes caused by using BIC (instead of AIC) quite well. The shifted coverages of 3-parameter PDFs exhibit slope-dependent

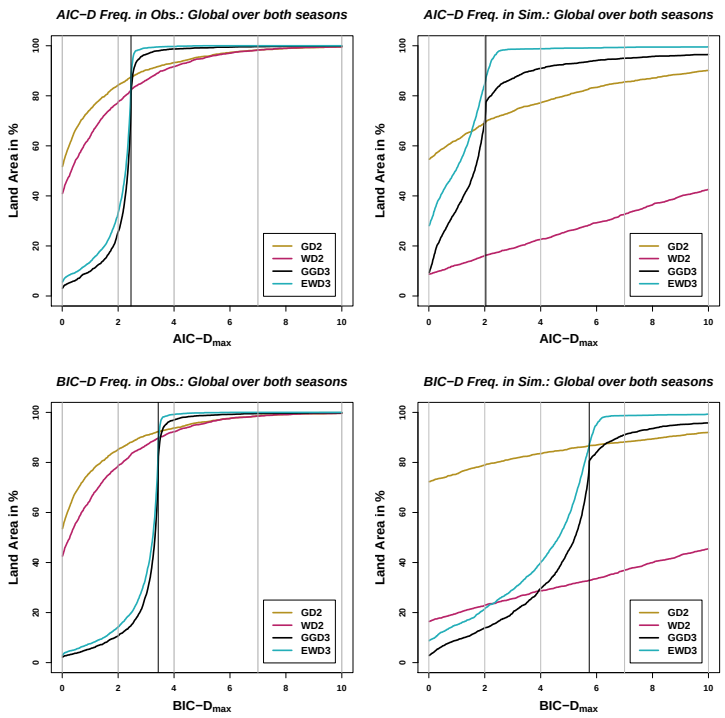

Figure I. AIC-D (**top**) and BIC-D (**bottom**) frequencies: percentage of global grid-points during both seasons in which each PDF yields AIC-D/BIC-D values that are smaller than or equal to a given AIC-D$_{max}$/BIC-D$_{max}$ value. The vertical black line indicates the different complexity penalties between 3- and 2-parameter PDFs (see bottom row of Table I). AIC-D/BIC-D frequencies are displayed for each candidate PDF for observations (**left**) and simulations (**right**).

changes at all AIC-D$_{max}$/BIC-D$_{max}$ values. That causes 3-parameter PDFs to be best-suited (AIC-D$_{max}$/BIC-D$_{max}$ value of 0) in fewer grid-points. In each grid-point, a single PDF must still be best-suited. In a second-order approximation, the coverages of 2-parameter PDFs, thus, slightly adjust for small AIC-D$_{max}$/BIC-D$_{max}$ values to the changes of 3-parameter PDFs' coverages at the AIC-D$_{max}$/BIC-D$_{max}$ value of 0. Consequently, the coverages of 2-parameter PDFs are overall fairly insensitive to the change of the information criterion because they only adjust slightly. The coverages of 3-parameter PDFs are more sensitive to the changed information criterion because they universally exhibit a horizontal shift.

This shift, however, does not result in a universally uniform sensitivity. The sensitivity of the coverages of 3-parameter PDFs depends on their slope. Because their slope is in both realizations flat for AIC-D$_{max}$ values beyond 2.5, the coverages of 3-parameter PDFs are insensitive beyond AIC-D$_{max}$/BIC-

D_{max} values of 2.5 plus 0.97 (3.7) in observations (simulations). Therefore, the coverages of the AIC-D/BIC-D category "no skill" (AIC-D/BIC-D > 10) and "sufficient" (AIC-D/BIC-D ≤ 7) are robust concerning a change of the information criterion from AIC to BIC in both realizations. In observations, the AIC-D/BIC-D category "well" (AIC-D/BIC-D ≤ 4) is also robust to the change of the information criterion (because $2.5 + 0.97 \leq 4$). Further, the slope of coverages of both 3-parameter PDFs is rather flat between AIC-D_{max} values of 1 and 2, in observations. In observations, the AIC-D/BIC-D category "ideal" (AIC-D/BIC-D ≤ 2) is, therefore, also rather robust to the change of the information criterion. Ergo, all AIC-D/BIC-D categories are in observations sufficiently robust to the change of the information criterion. We identify sensitive performances to the change of the information criterion only in simulations for the AIC-D/BIC-D categories "ideal" and "well". This sensitivity does not affect the main argument against GD2 in simulations. GD2 displays a worthless (insufficient) performance in 12% (18%) of grid-points. Also for BIC-D frequencies, GD2 displays a worthless (insufficient) performance in more than 10% (14%) of grid-points in simulations. In contrast, EWD3 displays, irrespective of the employed information criterion, a worthless or insufficient performance only in 1% of grid-points – EWD3 reduces the count of grid-point characterized by this highly undesirable performance by over one magnitude.

We extensively draw our conclusion from erroneous performances of our candidate PDFs. Irrespective of the information criterion, erroneous performances are for EWD3 virtually non-existent, but manifest for GD2 in a non-negligible percentage of grid-points in both realizations. Thus, as also discussed in the initial submission (e.g. when introducing AIC-D in the results, and when elaborating on them in the discussion), the risk of underfitting by using 2-parameter PDFs seems larger than the risk of overfitting by using 3-parameter PDFs. Consequently, once the need for 3-parameter candidate PDFs is established, their remaining punishment relative to 2-parameter PDFs biases the analysis; particularly for small AIC-D values. Because of the complexity penalty in the information criterion, our 3-parameter candidate PDFs outperform our 2-parameter candidate PDF only for AIC-D_{max} values beyond their increased complexity penalty (black vertical line in Fig I). We argue that maintaining the complexity penalty (beyond the proven inability of 2-parameter distributions) causes an artificial disadvantage for 3-parameter PDFs for small AIC-D values. Therefore, the complexity penalty biases and obscures our analysis for small AIC-D_{max} values. We interpret the results from this BIC-D analysis as another confirmation of our line of argumentation. Anyhow, this discussion (and our interpretation of a confirmation of our line of argumentation) only underlines our conclusion that EWD3 is better suited than GD2. In contrast, we draw that conclusion from erroneous performances of GD2 that manifest irrespective of the employed information criterion.

The above-conducted analysis helped us to streamline our reasoning. In consequence, we slightly altered several lines of the manuscript to simplify our

line of argumentation. This helps us to convey, and readers to intuitively understand our conclusions. In this process, we conducted two different types of changes. Firstly, changes concerning the proper communication of AIC's punishment (including the above-mentioned bias). Secondly, changes that focus our analysis on GD2 and EWD3, instead of highlighting all four candidate PDFs almost equally prominent.

In the thorough analysis of AIC's and BIC's complexity penalties, we identified an intuitive way to visualize the penalty difference between 2- and 3-parameter PDFs. The black vertical line in Fig. I. Including this black line also in Fig. 5 enables us to elaborate more precise on the impact of that penalty difference. Therefore, we adapted Fig. 5 and discuss the adaptation in the text. This simplifies our line of argumentation.

We changed a paragraph in Section 3.1.1 (lines 458 to 470) to: "It seems worth elaborating on the insufficient (only average) confidence in EWD3 to perform ideally in observations (ensemble simulations) around the globe. The complexity penalty of AIC correctly punishes EWD3 stronger than GD2 because AIC evaluates whether EWD3's increased complexity (relative to GD2) is necessary. However, the results justify the necessity for this increased complexity – GD2 performs erroneously in 26% (6%), insufficiently in 18% (2%), and without any skill in 12% (1%) of the global land area in ensemble simulations (observations). The risk of underfitting by using 2-parameter PDFs seems larger than the risk of overfitting by using 3-parameter PDFs. Once the need for 3-parameter candidate PDFs is established, their remaining punishment relative to 2-parameter PDFs biases the analysis; particularly for the ideal AIC-D category. EWD3's increased complexity penalty relative to 2-parameter candidate PDFs depends on the sample size and amounts to 2.46 in observations and 2.04 in ensemble simulations (see black vertical lines in Fig. 5 (a)–(d)). The AIC-D_{max} value beyond which EWD3 reaches coverages close to 100% approximately amounts to EWD3's increased penalty (see Fig. 5 (a)–(d)). Correcting EWD3's coverages for this bias would affect our evaluation of EWD3's performance only for the ideal AIC-D category. To illustrate this effect, we only consider AIC's estimated likelihood (without its penalty). Such a consideration corrects this complexity bias in EWD3's performance. While we analytically analyzed this consideration, a first-order approximation suffices for the scope of this publication. In that first-order approximation of this consideration, we simply shift the curve of EWD3 by 2.46 units leftwards in observations (Fig. 5 (a) and (b))) and by 2.04 units leftwards in ensemble simulations (Fig. 5 (c) and (d)). After this shift, EWD3 would also perform ideal with substantial confidence."

We substituted a sentence in Section 3.3 (lines 579 to 580) by the following elaboration: "(...) higher AIC-penalty compared to GD2. As a reminder, AIC punishes EWD3 stronger than GD2. Nevertheless this complexity punishment, it is obvious by now that our 2-parameter PDFs are inept to universally deliver normal distributed SPI time-series; particularly if one considers all depicted dimensions of the task at hand. As it turns out, this punishment is the sole

reason for both performance limitations that EWD3 displays in Table 6: (i) for the ideal AIC-D category and (ii) EWD3's tied performance with GD2 for an accumulation period of 12-months in ensemble simulations. As shown before, AIC's punishment is particularly noticeable in the ideal category. Further, this punishment also affects the tied performance ranking for the accumulation period of 12-months. To illustrate this effect, we again consider AIC's estimated likelihood (without its penalty) to correct EWD3's performance for the complexity punishment. While we again analytically analyzed this consideration, for the scope of this publication a first-order approximation suffices also here. In that first-order approximation of this consideration, EWD3's coverages of Table 6 shift again by 2.46 (2.04) AIC units in observations (ensemble simulations). Since neighboring AIC-D categories differ by 2-3 AIC units, this approximation shifts EWD3's coverages of Table 6 by roughly one category. Such a shift would solve EWD3's limitation in the ideal AIC-D category. Further, EWD3 would also perform best across all AIC-D categories in ensemble simulations; including the accumulation period of 12-months.

Despite the inclusion of the complexity penalty, EWD3 performs (...)"

Answering this question helped us to further streamline the conclusions we would like to convey. We realized that the manuscript highlights all four candidate PDFs almost equally for too long. Dismissing WD2 and GGD3 earlier helps us in telling the story. To focus our story on GD2 and EWD3, we conducted the following changes to the manuscript:

We substituted lines 342 to 345 in Section 3.1.1 by: "(...) during both seasons (Fig. 4, (b)). Aside from GD2, GGD3 and EWD3 also perform adequately in absolute terms for observations. Discriminating their deviations from $\mathcal{N}_{0,1}$ is difficult. On the one hand, GD2 represents the especially important left-hand tail of SPI_{3M} time-series' frequency distribution (D3) in JJA worse than our 3-parameter candidate PDFs (compare Fig. 4, (a) against (c) and (d)). On the other hand, GD2 displays smaller deviations from $\mathcal{N}_{0,1}$ than our 3-parameter candidate PDFs in the center of the SPI's distribution. Despite these minor differences (...)"

We substituted lines 411 to 431 in Section 3.1.1 by: "(...) considerably faster than GD2. EWD3 quickly compensates for AIC's complexity punishment (which is 2.46 units larger for EWD3 than for GD2 (indicated by the vertical black line in Fig. 5)). Beyond this vertical black line, EWD3 conclusively outperforms GD2 (the only intersection of the yellowish, and the bluish lines coincide with the intersection of that vertical black line in Fig. 5, (a) and (b)). EWD3 performs well (AIC-D_{max} ¡ 4) in virtually every global land grid-point. During DJF (JJA), EWD3 displays globally (in all land grid-points) AIC-D values of less than 5.03 (7.03). In contrast, GD2 performs erroneously (apparent by AIC-D_{max} values in excess of 4) in approximately 7% (6%) of the global land grid-points during DJF (JJA). Further, GD2 performs during both seasons insufficiently (AIC-D_{max} values beyond 7) in 2% and without skill (AIC-D_{max} values beyond 10) in 1% of the global land area. While EWD3 strictly

outperforms GGD3, GGD3 still performs similarly to EWD3 in observations. Thus, our focus on EWD3 becomes only plausible during the investigation of AIC-D frequencies in ensemble simulations."

We substituted lines 436 to 448 in Section 3.1.1 by: "We interpret EWD3's performance in ensemble simulations as ideal in approximately 85% (86%) of the global land area during DJF (JJA). For AIC-D$_{max}$ values beyond 2, EWD3 quickly approaches 100 % coverage, again, and performs erroneously or insufficiently only in 1% of the global land area during both seasons. In contrast, GD2 performs erroneously in 23% (30%) and insufficient in 14% (21%) of the global land grid-points during DJF (JJA). Yet, most telling might be the fraction (...)"

We included the following transition in between lines 453 and 454 in Section 3.1.1: "(...) over one magnitude (by a factor of roughly 20). EWD3 also universally outperforms GGD3. In view of their equal parameter-count, it seems rational to rather employ EWD3 than GGD3.

Analyzing AIC-D frequencies for both seasons (DJF and JJA) discloses no distinct season-dependent differences, similar to before in the investigation of deviations from $\mathcal{N}_{0,1}$. Therefore, we average identified land area coverages over both seasons in the summary of AIC-D frequencies. Table 3 summarizes (...)"

Aside, we inserted a sentence in the discussion. This sentence states that we also analyzed BIC-D frequencies and that they deliver similar results as shown for AIC-D frequencies. We insert this sentence at the beginning of the paragraph that starts in line 617: "We also repeated our AIC-D analysis with the Bayesian information criterion [Schwarz et al., 1978] which delivered similar results. Irrespective of the employed information criterion, the findings sketched above stay valid (...)"

References

[Blain et al., 2018] Blain, G. C., de Avila, A. M. H., and Pereira, V. R. (2018). 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.

[Naresh Kumar et al., 2009] Naresh Kumar, M., Murthy, C., Sesha Sai, M., and Roy, P. (2009). 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(3):381–389.

[Schwarz et al., 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

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 the literature. This study illuminates the long-standing dispute and proposes a solution ~~which~~ that 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~~ bases on a complementary evaluation. Actual against theoretical occurrence probabilities of SPI categories ~~, and the quality of the fit~~ evaluate the absolute performance of candidate distribution functions~~against their complexity with~~. In contrast, Akaike's ~~Information Criterion~~information criterion evaluates candidate distribution functions relative to each other while analytically punishing complexity. SPI time-series, spanning 1983–2013, are calculated from Global Precipitation Climatology Project's monthly precipitation data-set 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~~ exercising caution when using the gamma distribution to calculate SPI; especially in ensemble 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~~ deriving 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 ser-

vices for classifying droughts (Hayes et al., 2011). Advantages of SPI are its standardization (Sienz et al., 2012), its simplicity, 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~~ Neglecting such an investigation 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~~PDFs.

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~~ that tested candidate PDFs globally. However, they solely ~~provided~~ provide highly aggregated results ~~which~~ that are globally averaged for accumulation periods between 3- and 12-months and ~~concluded~~ conclude 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~~ that 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~~ has 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 one-sided 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~~ a set of PDFs. The best-suited PDF ~~for each accumulation periodand in each location~~of this set is then employed. Thus, the employed PDF might differ depending on the accumulation period, the location, or the data-set. 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.

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~~ because 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~~The above-mentioned goodness-of-fit tests equally evaluate each value of SPI's distribution. Such an evaluation focuses on the center of the distribution because the center of any distribution contains per definition more samples than the tails. In contrast, SPI usually analyzes (and thus depends on a proper depiction of) the distribution's tails. Therefore, a blurred focus manifests in these goodness-of-fit tests. Moreover, the convention to binarily interpret the above-mentioned goodness-of-fit tests aggravates this blurred focus. Because of this convention, 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~~). This inability prevents any reasonable aggregation of limitations that surface despite the blurred focus. Thus, 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~~ ought to 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~~Because of the ill-fitting focus and the ill-suited convention of these goodness-of-fit tests, they are inept to identify SPI's best-performing candidate distribution function out of a set of PDFs.

In agreement with this insight, those studies, that rigorously analyzed candidate distribution functions, or investigate an appropriate test methodology for evaluating SPI candidate PDFs, consequently 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~~ use AIC to discriminate the performance of different goodness-of-fit tests.

SPI calculation procedures were developed for observed precipitation data. Since models do not exactly reproduce the observed precipitation distribution, these procedures need to be tested and eventually adapted before being applied to modeled data. Here, we aspire to identify an SPI calculation algorithm that coherently describes modeled and observed precipitation (i.e. describes both modeled and observed precipitation distributions individually and concurrently). While testing SPI's calculation algorithm on modeled precipitation data is usually neglected, such a test demands nowadays a similarly prominent role as the one for 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 close that gap, this study evaluates the performance of candidate distribution functions in an output of 10 ensemble members of initialized seasonal hindcast simulations.

In this study, we test the adequacy of the gamma, Weibull, generalized gamma, and exponentiated Weibull distribution in SPI's calculation ~~algorithms~~algorithm. 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~~ complementary evaluations of their normality: (i) ~~it compares~~ evaluating their normality in absolute terms by comparing 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~~ evaluating their normality relative to each other with Akaike's ~~Information Criterion~~ information criterion (AIC) which analytically assesses of 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 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~~ . Observed and simulated precipitation is obtained from the monthly precipitation data-set of the Global Precipitation Climatology Project (GPCP) ~~serves as an observational product~~and the above mentioned initialized seasonal hindcast simulations, respectively. 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 GR15 (maximum 1.5°x1.5°) with 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). Here, we analyze the sum of convective and large-scale precipitation from these simulations (Pieper et al., 2020) .

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 events are excluded from the precipitation time-series before fitting the PDF and dealt with ~~specifically later . We optimize~~ later specifically. We estimate the parameters of our candidate PDFs in SPI's calculation algorithm with the maximum likelihood method (Nocedal and Wright, 1999) 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~~ The starting values identified from that scan are optimized with the simulated annealing method ~~(Bélisle, 1992) , (ii)~~ (SANN) (Bélisle, 1992) . Subsequently, these by SANN optimized starting values are again further optimized by 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~~ . If the BFGS quasi-Newton method leads to a convergence of the parameters of our candidate PDF, we achieve our goal and end the optimization here. If the BFGS quasi-Newton method does not lead to a convergence of the ~~most suitable~~ parameters of our candidate ~~PDFs and omitting~~

~~cases where convergence is not achieved~~PDF, then we circle back to the starting values optimized by SANN and optimize them again further but this time with the Nelder-Mead method (Nelder and Mead, 1965) . After identifying converging parameters, the probabilities of encountering the given precipitation totals are computed and transformed ~~to~~ into cumulative probabilities $(G(x))$.

~~Since PDFs which~~ If neither the BFGS quasi-Newton nor the Nelder-Mead method leads to any convergence of the most suitable parameters of our candidate PDFs, then we omit these grid-points where convergence is not achieved. For the gamma, Weibull, and exponentiated Weibull distribution, non-converging parameters are rare exceptions and only occur in a few negligible grid-points. For the generalized gamma distribution, however, non-convergence appears to be a more common issue and occurs in observations as well as in simulations in roughly every fifth grid-point of the global land area. This shortcoming of the generalized gamma distribution needs to be kept in mind when concluding its potential adequacy in SPI's calculation algorithm.

Since PDFs that describe the frequency distribution of precipitation totals are required to be only defined for the positive real axis, ~~that~~ the 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 in simulations over the Sahara and the Arabian Peninsula 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 precipita-

tion amounts. Those spurious precipitation amounts prevent us from directly identifying the probability of zero-precipitation events in modeled precipitation time-series. 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. We checked and ascertained the underlying assumption of this procedure – ~~assuming~~ that all ensemble members show in ~~the long-term~~ each grid-point 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~~ these 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~~ that 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 of a location parameter. The optimization of this second shape parameter also requires the re-optimization of the first two parameters. The fitting procedure of 3-parameter PDFs needs therefore considerable more computational resources than the fitting procedure of 2-parameter distribution functions.

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

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

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~~ weigh 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 ~~which~~ that 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~~. Avoiding unnecessary complexity 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~~ PDF is too simple) and variance (~~model~~ PDF is 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~~ best-performing distribution function attains ~~a minimum AIC value (AIC_{min})~~ the smallest AIC value 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~~. Thus, we use relative AIC differences (AIC-D) in our analysis. We calculate these AIC-D values for each PDF by computing the difference between its AIC value to the lowest AIC value of all four distribution functions. AIC-D

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~~and are calculated as follows:

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

The index i indicates different distribution functions. AIC_{min} denotes the AIC value of the best-performing distribution function.

For our analysis, AIC-D values 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 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~~ , which exhibit sufficiently similar AIC-D values, is unfeasible as a consequence of our rather small sample size (particularly in observations, but also in simulations). AIC-D values below two ~~should~~ ought to 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~~ considerably less confidence and values beyond ten essentially none (Burnham and Anderson, 2002).

The analysis of deviations from $\mathcal{N}_{0,1}$ assesses performances of candidate PDFs in absolute terms irrespective of the candidate PDF's complexity. In contrast, the AIC-D analysis evaluates the performance of candidate PDFs relative to each other while analytically punishing complexity. Consequently, the AIC-D analysis cannot evaluate whether the best-performing candidate distribution function also performs adequately in absolute terms. In opposition, deviations from $\mathcal{N}_{0,1}$ encounter difficulties when evaluating whether an increased complexity from one PDF to another justifies any given improvement. Both analyses together, however, augment each other complementary. This enables us to conclusively investigate: (i) which candidate PDF performs best while (ii) ensuring adequate absolute performance and while (iii) constraining the risk of over-fitting.

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~~ that 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~~ a percentage of the theoretically expected occurrence probability.

Again for each candidate distribution function, accumulation period, domain, and both seasons, we aggregate ~~AIC-Ds~~ AIC-D 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 $AIC\text{-}D_{max}$ 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~~ that balance each other in different grid-points with unrelated climatic characteristics. This statement is based on either one of the following two assumptions. First, the sum over ~~less~~ fewer 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 around~~ accumulated over 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~~ over 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~~ that identify differing borders for regions ~~which~~ that 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 the 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,~~ Aside from GD2~~leads to a better agreement between the frequency distribution of SPI_{3M} time-series and~~ , GGD3 and EWD3 also perform adequately in absolute terms for observations. Discriminating their deviations from $\mathcal{N}_{0,1}$ ~~than any of the here investigated 3-parameter PDFs over both seasons of interest. Still~~is difficult. On the one hand, GD2 represents the especially important left-hand tail of SPI_{3M} time-series' frequency distribution (D3) in JJA ~~relatively poor. Here, the investigated~~ worse than our 3-parameter ~~distributions, GGD3 and the exponentiated Weibull distribution (EWD3), perform better (~~candidate PDFs (compare Fig. 4, (a) against (c) and (d)). On the other hand, GD2 displays smaller deviations from $\mathcal{N}_{0,1}$ than our 3-parameter candidate PDFs in the center of the SPI's distribution. 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 always 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 of parameters 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 fitting methods. This results in missed optimizations opportunities ~~which~~ that impact GGD3 stronger than GD2 because of GGD3's ~~complexity. As a result~~increased complexity which leads to GGD3 requiring more data than GD2. Therefore, the weighted sum (weighted by the theoretical occurrence probability of the respective SPI class (Table 2)) 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~~ display similar deviations from $\mathcal{N}_{0,1}$ in observations (Fig. 4 (a), (c), and (d)), a noticeable difference emerges in ensemble simulations (Fig. 4 (e), (g), and (h)). GD2's ~~performance distinctly deteriorates in~~ performs distinctly worse than our 3-parameter PDFs in ensemble simulations.

In simulations, the fit onto 3-months precipitation totals is performed on all ten ensemble members at once. This 10-folds the sample size in simulations relative to observations. Presuming an imperfect fit for the 31 samples in observations, deviations from $\mathcal{N}_{0,1}$ are expected to reduce along our four candidate distribution functions as a result of 10-folding the sample size of their fit. Yet, GD2 does not benefit from 10-folding the sample size. GD2 performs similarly in observations and simulations (Fig. 4 ~~,~~ (a) and (e))~~relative to observations~~. In contrast, ~~both~~ our 3-parameter ~~candidate PDFs excel in describing the frequency distribution of 3-months precipitation totals in both seasons (~~PDFs display considerably smaller deviations from $\mathcal{N}_{0,1}$ in ensemble simulations than in observations (compare Fig. 4 ~~,~~ (c) and (d) against (g) and (h)). ~~Any distinction between~~ Consequently, both 3-parameter candidate ~~distribution functions is still difficult~~ PDFs excel during both seasons in ensemble simulations (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~~while any distinction between both 3-parameter candidate distribution functions is still difficult. On the one side, different frequency distributions between observed and modeled precipitation totals might be one reason for this difference. On the other side, 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 observedand 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~~ of three parameters also requires more data than the fit of two. It is therefore sensible to expect that 3-parameter PDFs benefit stronger than 2-parameter PDFs from an increase in sample size. Are our 3-parameter candidate PDFs are better suited than our 2-parameter PDFs to describe modeled precipitation distributions? Or benefit our 3-parameter PDFs just stronger than 2-parameter PDFs from an increasing sample size?

We attempt to disentangle both effects (analyzing modeled, instead of observed, precipitation distributions, and increasing the sample size) for our 2-parameter candidate PDFs, next. If the 2-parameter PDFs are suited to be applied to modeled precipitation data, they should benefit at least to some extent from this multiplication of sample size. Despite expecting irregularities in the magnitude of these reductions, they ought to be notable for candidate distribution functions ~~which~~ that are adequately suited to describe modeled 3-months precipitation totals – assuming an imperfect fit for the 31 events spanning our observational time-series. Therefore, we weigh each class' deviation from $\mathcal{N}_{0,1}$ by the theoretical occurrence probability (see Table 2) of the respective class and analyze weighted deviations from $\mathcal{N}_{0,1}$.

For the 2-parameter PDFs, the weighted deviations from $\mathcal{N}_{0,1}$ (shown in the legend of Fig. 4) either stay constant (for GD2 in DJF) or increase in simulations relative to observations (~~shown in the legend of Fig. 4, compare the~~ compare the legends in the left against the ~~right column).~~ one in the right column of Fig. 4). Relative to observations, GD2's weighted

deviations increase in simulations 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~~ Our 2-parameter candidate PDFs are better suited to describe observed than modeled 3-months precipitation totals. In contrast, ~~the~~ for our 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, the~~ The 3-parameter ~~candidate distribution functions perform distinctly better in describing modeled 3-months precipitation totals than the~~ candidate distribution functions benefit strongly from the artificial increase of our time-series and seem better suited than our 2-parameter candidate PDFs ~~in both of our investigated seasons~~to describe precipitation distributions obtained from ensemble simulations.

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 the 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,~~ It is noteworthy, that investigating deviations from $\mathcal{N}_{0,1}$ over the entire globe contains the risk of encountering deviations ~~which~~ that balance each other in different grid-points with unrelated climatic characteristics. Until dealing with this risk, our analysis of deviations from $\mathcal{N}_{0,1}$ only indicates that three candidate PDFs (GD2, GGD3, and EWD3) display an adequate absolute performance. On the one hand, we can reduce that risk by analyzing ~~these deviations~~ deviations from $\mathcal{N}_{0,1}$ only over specific regions~~, which is done~~ . This analysis safeguards our investigation by ensuring (rather than just indicating) an adequate absolute performance around the globe and is performed later. On the other hand, we first completely 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. We investigate AIC-D frequencies first to evaluate whether GGD3 and/or EWD3 perform sufficiently better than GD2 to justify their increased complexities.

In general, each ~~of the candidate distribution functions perform~~ candidate distribution function performs 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~~ ideally describes observed 3-months precipitation totals during both seasons in many 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% (83.5%) of the global land area in DJF ~~and 83.5 % in JJA. That should~~ (JJA). That ought to 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~~ EWD3 quickly compensates for AIC's complexity punishment (which is 2.46 units larger for ~~their increased penalty imposed by AICthrough 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~~ than for GD2 (indicated by the vertical black line in Fig. 5)). Beyond this vertical black line, EWD3 ~~in both seasons and 2.51 (2.58) for GGD3 in DJF (JJA))(see intersect between the yellowish~~conclusively outperforms GD2 (the only intersection of the yellowish, and the bluish ~~as well as the yellowish and black lines~~ lines coincide with the intersection of that vertical black line 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~~ performs well (AIC-D$_{max}$ ~~value of 2.65 (2.95) in~~ < 4) in virtually every global land grid-point. During 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~~ , EWD3 displays globally (in all land grid-points~~(apparent by~~) AIC-D values ~~beyond 4 in these grid-points~~of less than 5.03 (7.03). In ~~consequence, they need considerably longer to reach 98 % coverages – even allowing~~ contrast, GD2 performs erroneously (apparent by AIC-D$_{max}$ values ~~as high as~~ in excess of 4) in approximately 7% (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 JJAand 6.68~~ %) of the global land grid-points during DJF (JJA). Further, GD2 performs during both seasons insufficiently (~~6.66) in DJF). As a reminder:~~ AIC-D~~values between 4~~ $_{max}$ values beyond 7) in 2% 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~~ without skill (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)~~ values beyond 10) 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~~ 1% 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~~ . While EWD3 ~~does not fail in a single grid-point during both investigated seasons .~~ strictly outperforms GGD3, GGD3 still performs similarly to EWD3 in observations. Thus, our focus on EWD3 becomes only plausible during the investigation of AIC-D frequencies in ensemble simulations.

In ensemble simulations, 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. The confidence in GD2 drastically diminishes further when we analyze the performance of ~~the~~ our four candidate PDFs in ~~describing the frequency distribution of modeled 3-month precipitation totals~~ensemble simulations. 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 %~~ performance in ensemble simulations as ideal in approximately 85% (86%) of the global land area ~~in DJF and 86.4 % in JJA~~during DJF (JJA). For AIC-D$_{max}$ values beyond 2, EWD3 ~~again~~ quickly approaches 100 % coverage~~in both seasons. Our results are againrather 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~~ , again, and ~~69.04 % for GGD3)~~ performs erroneously or insufficiently only in 1% 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~~ during both seasons. In contrast, GD2 ~~'s performance in about one-third to one-fourth~~ performs erroneously in 23% (30%) and insufficient in 14% (21%) of the global land grid-points ~~(apparent by AIC-D values beyond 4 in these grid-points). Most~~ during DJF (JJA). Yet, 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~~skill in ensemble simulations. GD2 ~~and WD2 fail in their description~~ fails during DJF (JJA) in ~~9.87 % (14.95 %) and 57.84 % (56.57~~ 10% (15%) of the global land area~~, respectively. While GGD3 still fails in 3.61 % (4.23 %) of grid-points~~. In opposition, EWD3 only fails in ~~0.59 % (0.71~~ 0.45% (0.87%) during DJF (JJA). Ergo, employing EWD3, instead of GD2, reduces the count of grid-points ~~in which it's description of modeled 3-months precipitation totals is without any skill~~ without any skillful performance by over one magnitude (by a factor of roughly 20)~~relative to GD2~~. EWD3 also universally outperforms GGD3. Given their equal parameter-count, it seems rational to rather employ EWD3 than GGD3.

Analyzing AIC-D frequencies for both seasons (DJF and JJA) discloses no distinct season-dependent differences, similar to before in the investigation of deviations from $\mathcal{N}_{0,1}$. Therefore, we average identified land area coverages over both seasons in the summary of AIC-D frequencies. 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~~ during both seasons. EWD3 performs well (AIC-D $\leq$ 4) with substantial confidence ~~around the globe (AIC-D $\leq$ 2 in~~ (at least 95~~or more~~ % of land grid-points ~~) in either~~

~~realization~~conform performance) around the globe in both realizations. Additionally, EWD3 ~~performs well~~ also performs best in each of these analyses (each row of Table 3 in which we consider its performance with substantial confidence~~around the globe (AIC-D $\leq$ 4 in 95 or more % of land grid-points)in both realizations~~). The other analyzed candidate PDFs perform substantially worse than EWD3 in ensemble 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~~ insufficient (only average) confidence in EWD3 ~~does not perform ideally with substantial confidence~~ to perform ideally in observations (ensemble simulations) around the globe. ~~On the one side,~~ The complexity penalty of AIC correctly punishes EWD3 stronger than GD2 because AIC evaluates whether EWD3's increased complexity ~~justifies the increased penalty when evaluating whether that increased complexity~~ (relative to GD2) is necessary. However, the results justify the necessity for this increased complexity – GD2 performs erroneously in 26% (6%), insufficiently in 18% (2%), and without any skill in 12% (1%) of the global land area in ensemble simulations (observations). The risk of underfitting by using 2-parameter PDFs ~~is higher~~ seems larger 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~~ Once the need for 3-parameter candidate PDFs is established, their remaining ~~competition against~~ punishment relative to 2-parameter PDFs biases the analysis; ~~especially~~ particularly for the ideal AIC-D category. EWD3's increased complexity 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~~ ensemble simulations (see black vertical lines in Fig. 5 (a)–(d)). The AIC-D$_{max}$ value ~~where~~ beyond which EWD3 reaches ~~a coverage~~ coverages close to 100% ~~(~~approximately amounts to EWD3's increased penalty (see Fig. 5 (a)–(d)). ~~Indeed, if~~ Correcting EWD3~~solely competes with GGD3,~~'s coverages for this bias would affect our evaluation of EWD3~~performs ideally (~~'s performance only for the ideal AIC-D $\leq$ 2)~~over both seasons~~ category. To illustrate this effect, we only consider AIC's estimated likelihood (without its penalty). Such a consideration corrects this complexity bias in EWD3's performance. While we analytically analyzed this consideration, a first-order approximation suffices for the scope of this publication. In that first-order approximation of this consideration, we simply shift the curve of EWD3 by 2.46 units leftwards in observations (~~simulations) in 99 % (100 %) of the global land grid-points (not shown). Thus~~Fig. 5 (a) and (b))) and by 2.04 units leftwards in ensemble simulations (Fig. 5 (c) and (d)). After this shift, EWD3 ~~already performs at least on par with the best-performing candidate PDF in both realizations at virtually every grid-point.~~ would also perform ideal with substantial confidence.

~~These characteristics stay valid~~ The AIC-D frequencies of Table 3 are robust in all investigated regions except Australia ~~. Here~~(not shown). In Australia, GD2~~performs better than any other analyzed PDF~~ 's performance slightly improves relative to the global results during DJF in observations. In contrast~~during JJA observations~~, GD2 performs worse than any other investigated candidate PDFs (even worse than WD2) ~~. Additionally, WD2 and the other candidate PDFs also out-perform GD2~~ during JJA in observations and 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 $\mathcal{N}_{0,1}$.

Among our candidate PDFs, EWD3 is obviously the best-suited PDF for SPI. Yet, we still need to confirm whether also EWD3's absolute performance is adequate. While the global analysis indicated EWD3's adequateness, the ultimate validation of this claim is incumbent upon the regional analysis.

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~~ In this process, our results indicate an adequate absolute performance of GD2, GGD2, and EWD3. However, that investigation 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~~ that focuses on 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~~ across 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 the 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~~ delivering 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 during JJA – EWD3 outperforms GGD3 which performs better than GD2. Aside, all candidate PDFs ~~perform almost identical in their attempt to describe observed precipitation~~ display almost identical deviations from $\mathcal{N}_{0,1}$ over Australia during DJF in observation.

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~~ shown before, 2-parameter PDFs ~~are inadequately suited to properly describe modeled precipitation totals . In consequence~~ineptly describe precipitation totals obtained from ensemble simulations. Consequently, 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~~ Similarly 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~~ generally 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~~ For observations, the regional analysis confirms the ~~overall~~ insights from the global analysis in ~~observations for each region. In~~ each region: EWD3 is (same as GD2 and GGD3) an adequate PDF in SPI's calculation algorithm. For ensemble simulations, the regional analysis additionally corroborates the finding of the AIC-D analysis that ~~our 3-parameter candidate distribution functions perform in simulations~~ EWD3 performs noticeably better than ~~our 2-parameter PDFs~~GD2. The corroboration of this finding substantiates support for ~~the 3-parameter~~ EWD3.

The analysis of AIC-D frequencies proves that EWD3 is SPI's best distribution function among our candidate PDFs. Additionally, the regional investigation confirms the global analysis: the absolute performance of EWD3 is at minimum adequate in observations and ensemble simulations.

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~~ The results obtained from the ~~multi-distribution function calculation algorithm~~ SPI calculation algorithm that uses a multi-PDF approach are labeled AIC_{min}-analysis. As a baseline comparison, we choose the calculation algorithm and optimization method of the frequently used R-package from Beguería and Vicente-Serrano (2017) and refer to these results as baseline. To maximize the comparability of SPI time-series calculated with ~~our~~ this 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~~ compare Fig. 8 against Fig. 7~~(a) and (b)~~) thus also ~~serves as an indicator for~~ indicates the impact of ~~very similar parameter estimation methods which only differ by their optimization procedure~~the meticulousness applied to the optimization of the same parameter estimation method.

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 the best performing distribution function if it fulfills that expectation.

In contrast to previous results in this and other studies (Stagge et al., 2015), which showed no seasonal differences in the performance of candidate PDFs, ~~our~~ the 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~~ the 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~~procedure, at least for DJF-precipitation totals. Further, ~~our~~ the 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~~ the 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~~ the 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~~ identical and the only difference is the meticulousness of the optimization procedure. Since GD2 and ~~our~~ the 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~~ procedures by which these methods are optimized.

Unsurprisingly the same deficit as identified before for both 2-parameter candidate PDFs also emerges in ~~our~~ the 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~~ the baseline already performs especially ~~poor~~ poorly 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~~ the baseline's ~~inability~~ difficulties to sufficiently describe the frequency distribution of modeled SPI_{3M}. In ~~our~~ the 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).

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

3.1.4 Sensitivity to Ensemble Size

So far, we used all ensemble members at once to fit our candidate PDFs onto simulated precipitation. That improves the quality of the fit. In this section, we first analyze a single ensemble member and investigate subsequently the sensitivity of our candidate PDFs' performance on the ensemble size. In doing so, we properly disentangle the difference between observations and simulations from the impact of the sample size.

As before, 3-parameter candidate distribution functions also perform for a single ensemble simulation better than 2-parameter PDFs (Table 4). For a single ensemble member, the difference by which 3-parameter PDFs out-perform 2-parameter PDFs reduces considerably relative to the entire ensemble simulations (compare Table 4 against Table 3), though. In contrast to Table 3, all of our candidate distribution functions perform similarly between a single ensemble simulation and observations. In contrast to our previous results (e.g. when analyzing weighted sums of deviations from $\mathcal{N}_{0,1}$), modeled and observed precipitation distributions now seem sufficiently similar. Reducing the sample size for the fit by a factor of ten leads to more homogeneous performances of all candidate PDFs in simulations. As a reminder, AIC-D frequencies as depicted in Table 4 measure only relative performance differences. Consequently, our 2-parameter candidate PDFs do not actually perform better with fewer data. Instead, limiting the input data to a single ensemble member impairs our 3-parameter candidate PDFs stronger than our 2-parameter candidate PDFs. Optimizing 3 parameters needs more information than the optimization of 2 parameters. Irrespective of the realization, GD2 performs erroneously for 31 samples (apparent in grid-points which display AIC-D values beyond 4). Despite the need for more information, 31 samples suffice EWD3 to fix GD2's erroneous performances in both analyzed realizations.

In the next step, we isolate and investigate the improvement of the fit by an increasing sample/ensemble size. As a consequence of limited observed global precipitation data, we neglect observations and their differences to simulations in this remaining section. During this investigation, we reanalyze Table 4 while iteratively increasing the ensemble (sample) size for the fit (and the AIC-D calculation). Irrespective of the ensemble size, EWD3 performs proficiently (Table 5). Further, the fraction of grid-points in which EWD3 performs ideal increases constantly. This is a consequence of EWD3's better performance relative to our 2-parameter candidate PDFs. Unfortunately, AIC-Ds can only compare models that base on an equal sample size without adhering to additional undesired assumptions. Thus, any direct analysis of each candidate PDF's improvement relative to its own performance for a single ensemble member is with AIC-D frequencies not feasible. Despite this caveat, Table 5 still indicates strongly that EWD3 benefits stronger from the increased sample size than any of our 2-parameter candidate distribution functions. The larger the sample size, the larger is the margin by which EWD3 outperforms GD2.

Despite requiring more data, our 3-parameter candidate PDFs perform already better for 31 samples. For 31 samples, we identify this better performance of 3-parameter candidate PDFs in observations and simulations. Further, since our 3-parameter candidate PDFs require more data to estimate optimal parameters, they benefit in simulations stronger from additional samples than our 2-parameter candidate PDFs. That benefit becomes apparent in a distinctly improved relative performance after multiplying the sample size through the use of additional ensemble members.

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~~7). 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 ensemble 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~~ performance shift 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~~ neither 3-parameter candidate ~~PDFs do not display such a phase transition~~PDF displays such a shift in its performance. Both 3-parameter PDFs perform for accumulation periods shorter and longer than 3-months similarly well.

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~~6) 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~~ As a reminder, AIC punishes EWD3 ~~only competes against GGD3,~~ stronger than GD2. Nevertheless this complexity punishment, it is obvious by now that our 2-parameter PDFs are inept to universally deliver normal distributed SPI time-series; particularly if one considers all depicted dimensions of the task at hand. As it turns out, this punishment is the sole reason for both performance limitations that EWD3 ~~performs ideal (~~displays in Table 6: (i) for the ideal AIC-D ~~$\leq$ 2) in 88 % and shows no skill (~~category and (ii) EWD3's tied performance with GD2 for an accumulation period of 12-months in ensemble simulations. As shown before, AIC's punishment is particularly noticeable in the ideal category. Further, this punishment also affects the tied performance ranking for the accumulation period of 12-months. To illustrate this effect, we again consider AIC's estimated likelihood (without its penalty) to correct EWD3's performance for the complexity punishment. While we again analytically analyzed this consideration, for the scope of this publication a first-order approximation suffices also here. In that first-order approximation of this consideration, EWD3's coverages of Table 6 shift again by 2.46 (2.04) AIC units in observations (ensemble simulations). Since neighboring AIC-D ~~$>$ 10) in less than 5 % of the global land grid-points. Moreover~~categories differ by 2-3 AIC units, this approximation shifts EWD3's coverages of Table 6 by roughly one category. Such a shift would solve EWD3's limitation in the ideal AIC-D category. Further, EWD3 would also perform best across all AIC-D categories in ensemble simulations; including the accumulation period of 12-months.

Despite the inclusion of the complexity penalty, EWD3 performs still best in 32 out of all 40 analyses (all rows of Table 3 and Table ~~4~~6), 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 respective table). In contrast, GD2 ~~(WD2) only performs 2 (1~~only performs 7 (2) times best (while also performing with at least average confidence) – WD2 performs once best and ~~GGD2 never performs best~~GGD3 never.

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 than 3-months) accumulation periods (Stagge et al., 2015; Sienz et al., 2012). Introducing ensemble 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~~ around the 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~~ proposal stems from a ~~phase transition in the relative performance of~~ focus on 2-parameter PDFs ~~, which we also identify in this study~~that exhibit a shift in their performance which depends on the scrutinized accumulation period. 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~~ because of its discretized ~~description~~ nature (not shown). As a result of its discretized nature, the analyzed sample size prescribes the magnitude of deviations from $\mathcal{N}_{0,1}$. Consequently, these deviations are spatially invariant and aggregate with each additional grid-point. Thus, deviations from $\mathcal{N}_{0,1}$ will not spatially balance each other.

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~~ inept for all accumulation periods, locations, and realizations. ~~Yet~~Despite this inability of 2-parameter PDFs, EWD3 competed against 2-parameter PDFs in our analysis. This competition unnecessarily (given the inadequacy of 2-parameter PDFs, the risk of underfitting seems to outweigh the risk of overfitting) exacerbates EWD3's performance assessed with AIC-D because AIC punishes complexity (irrespective of that risk consideration). 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 outperforms any other candidate PDF~~without performing ideally~~. Yet, EWD3 does not perform ideally with substantial confidence in ensemble simulations. However, ~~accepting the need for a 3-parameter PDF in SPI's calculation algorithm a priori levels~~ leveling the playing field for candidate distribution functions with different parameter counts in our AIC-D analysis ~~and~~ leads to an ideal performance of EWD3 ~~globally~~universally.

~~The~~ We also repeated our AIC-D analysis with the Bayesian information criterion (Schwarz et al., 1978) which delivered similar results. Irrespective of the employed information criterion, 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~~ performances 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 with the 2-parameter gamma distribution.

Overall our 3-parameter candidate PDFs perform better than investigated 2-parameter candidate PDFs. Despite requiring more data, a sample size of 31 years suffices our 3-parameter candidate PDFs to outperform our 2-parameter candidate PDFs in simulations and observations. Further, our 3-parameter candidate PDFs greatly benefit from an increase in the sample size in simulations. In simulations, such a sample size sensitivity analysis is feasible by exploiting different counts of ensemble members. Whether 3-parameter PDFs would benefit similarly from an increased sample size in observations is likely but ultimately remains speculative because trustworthy global observations of precipitation are temporally too constrained for such a sensitivity analysis.

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 of the meticulousness applied to optimize the same parameter estimation method. Despite using the same 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~~ procedures by which these methods are optimized. In our analysis, not different PDFs but different ~~optimizations~~ optimization procedures of the same parameter estimation method can 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~~ (and the results we obtained from the analysis of our candidate PDFs). These results suggest

that the performances of candidate PDFs are independent of the season. In contrast, the baseline performs similar to GD2 during JJA, ~~its performance~~ but the performance of the baseline 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 – that depicts the largest deviation from $\mathcal{N}_{0,1}$~~ we encountered in this study. 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~~ While the largest deviations from $\mathcal{N}_{0,1}$ occur during DJF in Australia, ~~but~~ the baseline performs particularly ~~poor~~ poorly 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~~ six here analyzed SPI calculations (GD2, WD2, GGD3, EWD3, baseline, and AIC_{min}-analysis) in 63 out of 98 different analyses, which range ~~along~~ across all seven SPI categories, all seven regions, and ~~along observations as well as simulations~~both realizations. Aside from the investigated baseline and in ~~agreement with (Stagge et al., 2015)~~ general agreement with Stagge et al. (2015) , we find ~~no~~ only in Australia minor 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~~ another performance criterion that requires interpretation. That criterion informs whether the candidate PDFs conform to the respective AIC-D categories in sufficient grid-points globally and, therefore, ~~need~~ needs to interpret which fraction of the global land grid-points can be considered sufficient. For this fraction of global land grid-points, we select 85 % and 95 % as thresholds. ~~In consequence~~Consequently, 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 at least 95% ~~or more~~ of 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~~Lastly, 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~~ 320 analyzed evaluations in this study (all entries of Table 3~~and~~ , Table 4, Table ~~4~~5, and Table 6). Across all ~~40~~ 80 analyses (all rows of Table 3~~and~~ , Table 4, Table 5, and Table 6), the four candidate PDFs perform insufficiently ~~65~~ 132 times, while they perform with substantial (average) confidence ~~64 (31~~130 (58) times.

There is scope to further test the robustness of our derived conclusions in different models with different time horizons and foci on 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. A

recent study suggests that a 4-parameter extended generalized Pareto distribution excels in describing the frequency distribution of precipitation (Tencaliec et al., 2020) . Anyhow, the inclusion of yet another distribution parameter additionally complicates the optimization of the parameter estimation method. We already exemplified the impact of the meticulousness of the applied optimization in this study. Establishing a standard for the optimization process seems currently more urgent than attempts to improve SPI through 4-parameter PDFs.

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 common bias in SPI's ~~common~~ calculation algorithms obscures the evaluation of predictive skill of ensemble simulations by inducing a blurred representation of ~~the frequency distribution of modeled precipitation totals~~their precipitation distributions. That blurred representation ~~translates to~~ emerges in 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~~ in ensemble simulations stronger than in 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

Current SPI calculation algorithms are tailored to describe observed precipitation distributions. Consequently, current SPI calculation algorithms are ineptly suited to describe precipitation distributions obtained from ensemble simulations. Also in observations, erroneous performances are apparent and well-known, but less conspicuous than in ensemble simulations. We propose a solution that rectifies these issues and improves the description of modeled and observed precipitation distributions individually as well as concurrently. The performance of 2-parameter candidate distribution functions is inadequate for this task. By increasing the parameter count of the candidate distribution function (and thereby also its complexity) a distinctly better description of precipitation distributions can be achieved. In simulations and observation, the here identified best-performing candidate distribution function – the exponentiated Weibull distribution (EWD3) – performs proficiently for every common accumulation period (1-, 3-, 6-, 9-, and 12-months) virtually everywhere around the globe. Additionally, EWD3 excels when analyzing ensemble simulations. Its increased complexity (relative to GD2) leads to an outstanding performance of EWD3 when an available ensemble multiplies the sample size.

We investigate different candidate distribution functions (gamma (GD2), Weibull (WD2), generalized gamma (GGD3), and exponentiated Weibull distribution (EWD3)) in SPI's calculation algorithm ~~concerning~~ and evaluate 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~~ . This analysis focuses on an accumulation period of 3-months and ~~testing~~ tests the conclusions drawn from that focus for the most common other accumulation periods (1-, 6-, 9-, and 12-months). ~~Normality~~ The normality of

SPI is assessed by two complementary analyses. The first analysis checks the absolute performance of candidate PDFs 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~~ The second analysis evaluates candidate PDFs relative to each other while penalizing unnecessary complexity with Akaike's Information Criterion (AIC).

~~Our results show that~~ Irrespective of the accumulation period or the data-set, GD2 ~~is~~ seems sufficiently suited to ~~calculate SPIderived from observations for all accumulation periods analyzed. WD2 performs in observations better for an accumulation periodof 1-months but worse for longer accumulation periods. Based on our analysis of AIC-D values and deviations from $\mathcal{N}_{0,1}$~~be employed in SPI's calculation algorithm in many grid-points of the globe. Yet, GD2 also performs erroneous in a non-negligible fraction of grid-points. These erroneous performances are apparent in observations and simulations for each accumulation period. More severely, GD2's erroneous performances decline further in ensemble simulations. Here, GD2 performs in a non-negligible fraction of grid-points also insufficient or even without any skill. In contrast, EWD3 performs ~~exceptionally well and better than any 2-parameter candidate PDF in observations~~ for all accumulation periods ~~. Further~~without any defects, irrespective of the data-set. Despite requiring more data than 2-parameter PDFs, we identify ~~considerable differences between observations and simulations. For all accumulation periods analyzed in simulations, both 2-parameter candidate PDFsperform inadequately (WD2) or sufficiently but only with average confidence around the globe (~~EWD3's proficient performance for a sample size of 31 years in observations as well as in simulations. Further, ensemble simulations allow us to artificially increase the sample size for the fitting procedure by including additional ensemble members. Exploiting this possibility has a major impact on the performance of candidate PDFs. The margin, by which EWD3 outperforms GD2~~). In contrast, ~~, further increases with additional ensemble members. Furthermore, EWD3 performs particularly well with substantial confidence around the entire globe in simulations and for every accumulation period analyzed~~demonstrates proficiency also for every analyzed accumulation period around the globe. The accumulation period of 12-months poses in simulations the only exception. Here, EWD3 ~~still performs well but only with average confidence~~ and GD2 both perform similarly well around the globe. ~~We~~ Still, we find that 3-parameter PDFs are generally better suited in SPI's calculation algorithm than 2-parameter PDFs. ~~Our results show~~

Given all the dimensions (locations, realizations, accumulation periods) of the task, our results suggest that the risk of ~~overfitting 3-parameter PDFs is overcompensated by~~ underfitting by using 2-parameter PDFs is larger than the risk of ~~underfitting 2-parameter~~ overfitting by employing 3-parameter PDFs. We strongly advocate ~~to adapt and use 3-parameter distribution functions instead of 2-parameter PDFsfor the calculation algorithm of SPI~~adapting the calculation algorithm of SPI and the therein use of 2-parameter distribution functions in favor of 3-parameter PDFs. Such an adaptation is particularly important for the proper evaluation and interpretation of drought predictions ~~and~~ derived from ensemble simulations. For this adaptation, we propose the employment of EWD3 as a 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 simulations are available at the World Data Center for Climate (WDCC): http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=DKRZ_LTA_1075_ds00001 maintained by the Deutsche Klimarechenzentrum (DKRZ, German Climate Computing Centre).

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~~ This work was funded by the BMBF-funded joint research projects RACE – Regional Atlantic Circulation and Global Change and RACE – Synthesis. 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.

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.

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.: Numerical optimization, Springer series in operations research, 1999.

Pieper, P., Düsterhus, A., and Baehr, J.: MPI-ESM-LR seasonal precipitation hindcasts, http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=DKRZ_LTA_1075_ds00001, 2020.

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.

Schwarz, G. et al.: Estimating the dimension of a model, The annals of statistics, 6, 461–464, 1978.

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.

Tencaliec, P., Favre, A.-C., Naveau, P., Prieur, C., and Nicolet, G.: Flexible semiparametric Generalized Pareto modeling of the entire range of rainfall amount, Environmetrics, 31, e2582, 2020.

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.

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~~ SPI intervals and theoretical occurrence probabilities (according to WMO's *SPI User Guide* (Svoboda et al., 2012)).

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

Table 3. Percent of grid-points ~~which~~ that are classified into specific AIC-D categories (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 a 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~~ that exceed/subceed in 85/15% of the grid-points as a sign of average confidence in the candidate PDF (yellow) to overall perform according to the respective AIC-D category. Percentages ~~which~~ that fall short of 85% (or ~~which~~ that show no skill in more than 15%) are considered as an 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**
	Ensemble 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**

Table 4. Percent of grid-points that are classified into specific AIC-D categories (according to Burnham and Anderson (2002)) 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 a 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 that exceed/subceed in 85/15% of the grid-points as a sign of average confidence in the candidate PDF (yellow) to overall perform according to the respective AIC-D category. Percentages that fall short of 85% (or that show no skill in more than 15%) are considered as an overall sign of insufficient confidence in the candidate PDF (red). In contrast to Table 3, the evaluation of simulations bases on a single ensemble member. Observations are identical to Table 3.

SPI Period	Realization	AIC-D category	GD2	WD2	GGD3	EWD3
3-Months	Observations	Ideal (AIC-D $\leq$ 2)	**84**	**76**	**22**	**31**
		Well (AIC-D $\leq$ 4)	**94**	**91**	**98**	**100**
		Sufficient (AIC-D $\leq$ 7)	**98**	**98**	**100**	**100**
		No Skill (AIC-D $>$ 10)	**1**	**0**	**0**	**0**
	Single Ensemble Member	Ideal (AIC-D $\leq$ 2)	**83**	**76**	**19**	**28**
		Well (AIC-D $\leq$ 4)	**93**	**92**	**98**	**100**
		Sufficient (AIC-D $\leq$ 7)	**98**	**98**	**100**	**100**
		No Skill (AIC-D $>$ 10)	**1**	**0**	**0**	**0**

Table 5. Percent of grid-points that are classified into specific AIC-D categories (according to Burnham and Anderson (2002)) 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 a 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 that exceed/subceed in 85/15% of the grid-points as a sign of average confidence in the candidate PDF (yellow) to overall perform according to the respective AIC-D category. Percentages that fall short of 85% (or that show no skill in more than 15%) are considered as an overall sign of insufficient confidence in the candidate PDF (red). In contrast to Table 3, the evaluation of simulations bases on different ensemble sizes.

SPI Period	Ensemble Size	AIC-D category	GD2	WD2	GGD3	EWD3
3-Months	2	Ideal (AIC-D $\leq$ 2)	78	56	43	57
		Well (AIC-D $\leq$ 4)	87	74	96	99
		Sufficient (AIC-D $\leq$ 7)	94	90	98	100
		No Skill (AIC-D $>$ 10)	3	4	1	0
	3	Ideal (AIC-D $\leq$ 2)	77	45	53	69
		Well (AIC-D $\leq$ 4)	86	61	96	99
		Sufficient (AIC-D $\leq$ 7)	93	79	99	100
		No Skill (AIC-D $>$ 10)	4	10	1	0
	4	Ideal (AIC-D $\leq$ 2)	75	38	59	74
		Well (AIC-D $\leq$ 4)	84	50	95	99
		Sufficient (AIC-D $\leq$ 7)	90	67	98	100
		No Skill (AIC-D $>$ 10)	7	19	2	0
	5	Ideal (AIC-D $\leq$ 2)	74	31	63	79
		Well (AIC-D $\leq$ 4)	82	42	94	99
		Sufficient (AIC-D $\leq$ 7)	89	57	97	99
		No Skill (AIC-D $>$ 10)	7	30	2	0
	6	Ideal (AIC-D $\leq$ 2)	73	27	64	80
		Well (AIC-D $\leq$ 4)	81	36	93	99
		Sufficient (AIC-D $\leq$ 7)	88	50	96	99
		No Skill (AIC-D $>$ 10)	9	37	2	0
	7	Ideal (AIC-D $\leq$ 2)	70	25	66	81
		Well (AIC-D $\leq$ 4)	78	33	92	98
		Sufficient (AIC-D $\leq$ 7)	86	45	96	99
		No Skill (AIC-D $>$ 10)	10	43	2	1
	8	Ideal (AIC-D $\leq$ 2)	69	21	67	83
		Well (AIC-D $\leq$ 4)	77	29	91	98
		Sufficient (AIC-D $\leq$ 7)	85	39	95	99
		No Skill (AIC-D $>$ 10)	11	49	3	1
	9	Ideal (AIC-D $\leq$ 2)	66	20	67	85
		Well (AIC-D $\leq$ 4)	76	27	90	99
		Sufficient (AIC-D $\leq$ 7)	84	36	95	99
		No Skill (AIC-D $>$ 10)	12	53	3	1

Table 6. Percent of grid-points ~~which~~ that are classified into specific AIC-D categories (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 a 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~~ that exceed/subceed in 85/15% of the grid-points as a sign of average confidence in the candidate PDF (yellow) to overall perform according to the respective AIC-D category. Percentages ~~which~~ that fall short of 85% (or ~~which~~ that show no skill in more than 15%) are considered as an overall sign of insufficient confidence in the candidate PDF (red). In contrast to Table 3, this table evaluates different accumulations periods of SPI.

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
	Ensemble 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
	Ensemble 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
	Ensemble 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
	Ensemble 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

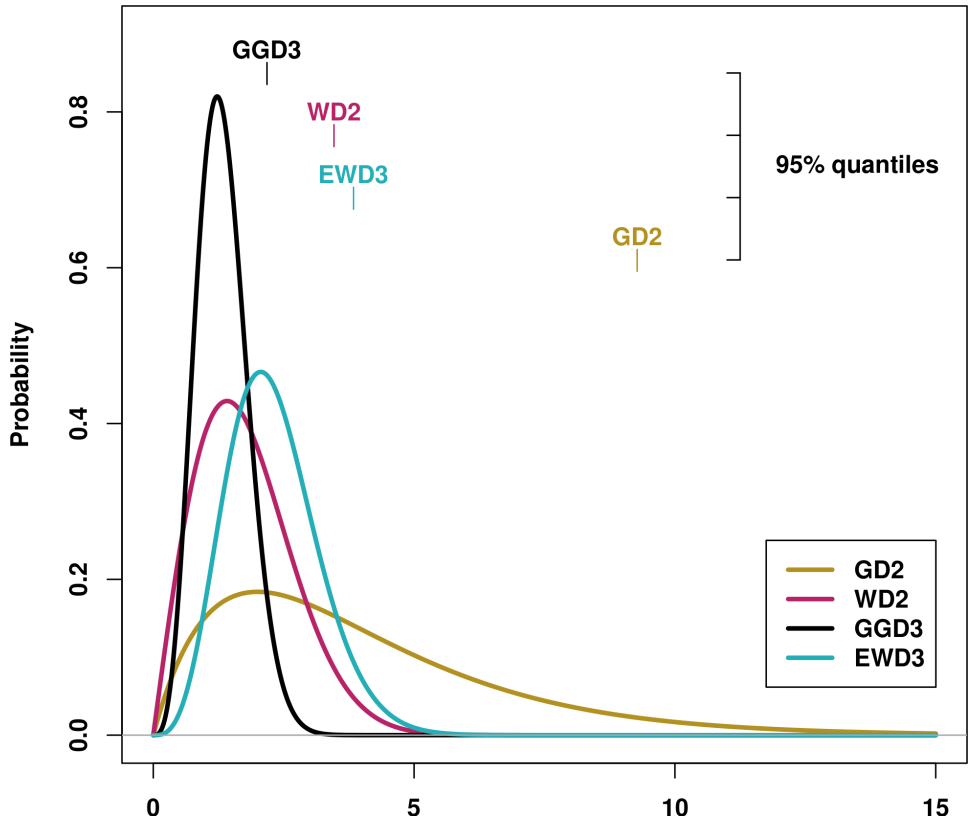

Figure 1. Candidate Distribution functions whose performance is investigated in this study: the 2-parameter gamma distribution (GD2), the 2-parameter Weibull distribution (WD2), the 3-parameter generalized gamma distribution (GGD3) and the 3-parameter exponentiated Weibull distribution (EWD3). Displayed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

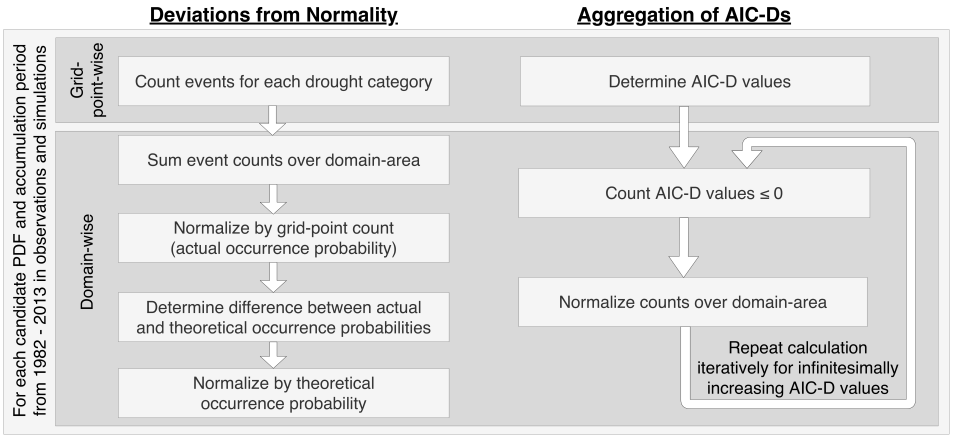

Figure 2. Flow chart of methods to aggregate deviations from $\mathcal{N}_{0,1}$ (**left**) and AIC-D frequencies (**right**) over domains.

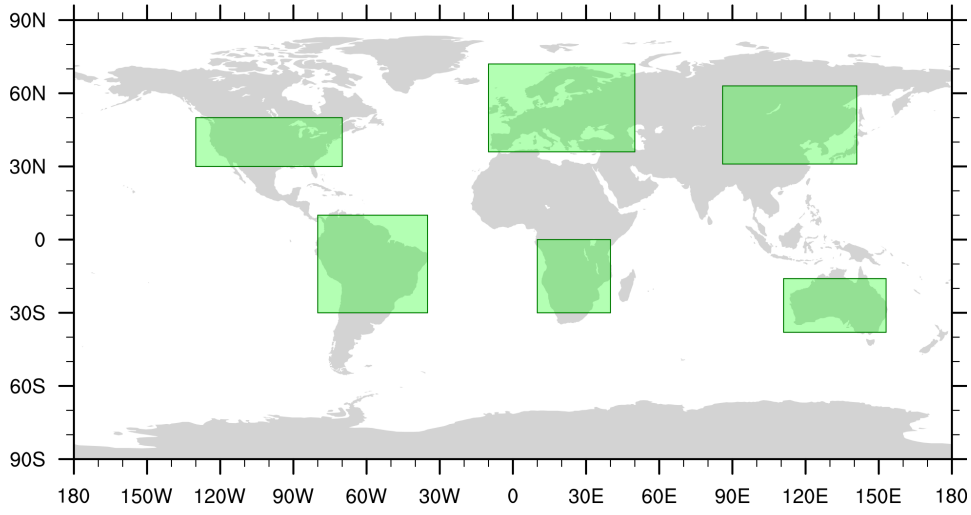

Figure 3. Borders of regions examined in this study.

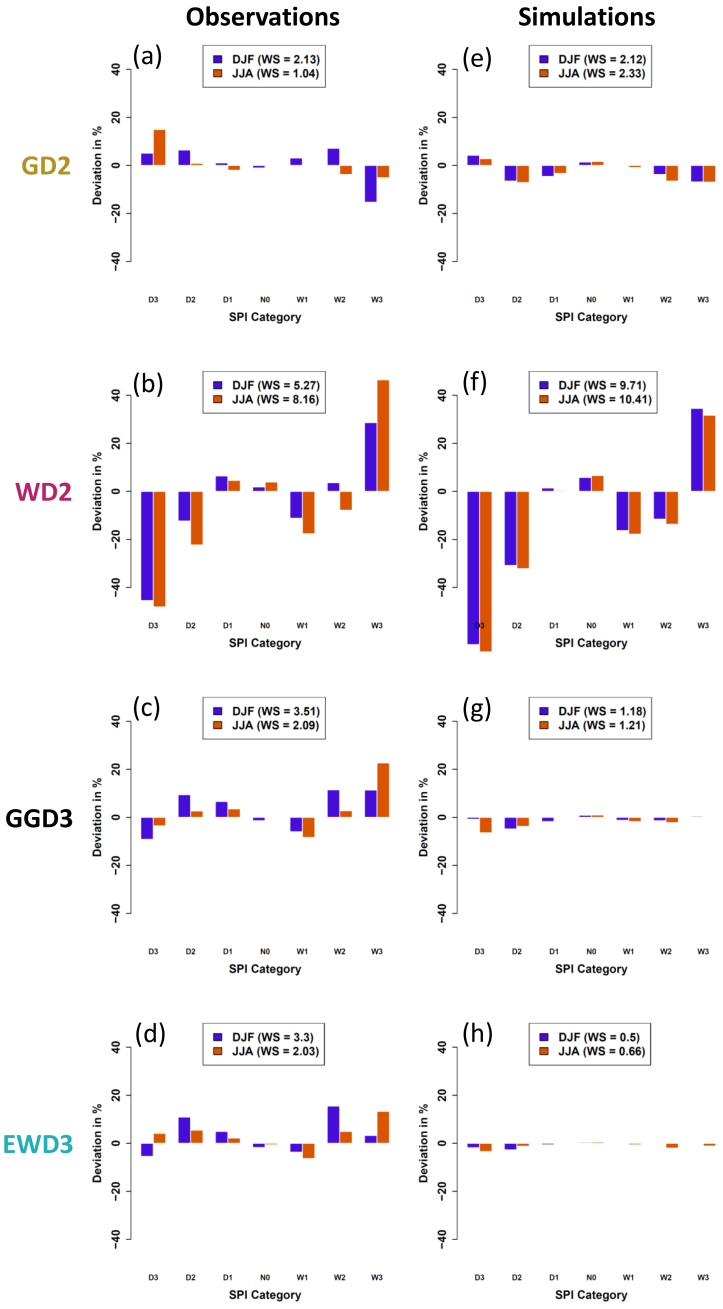

Figure 4. Deviations from $\mathcal{N}_{0,1}$ over the entire globe for observed (**left**) and modeled (**right**) SPI time-series. SPI time-series are derived by using the simple 2-parameter gamma distribution (GD2, **top row**), the simple 2-parameter Weibull distribution (WD2, **second row**), the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

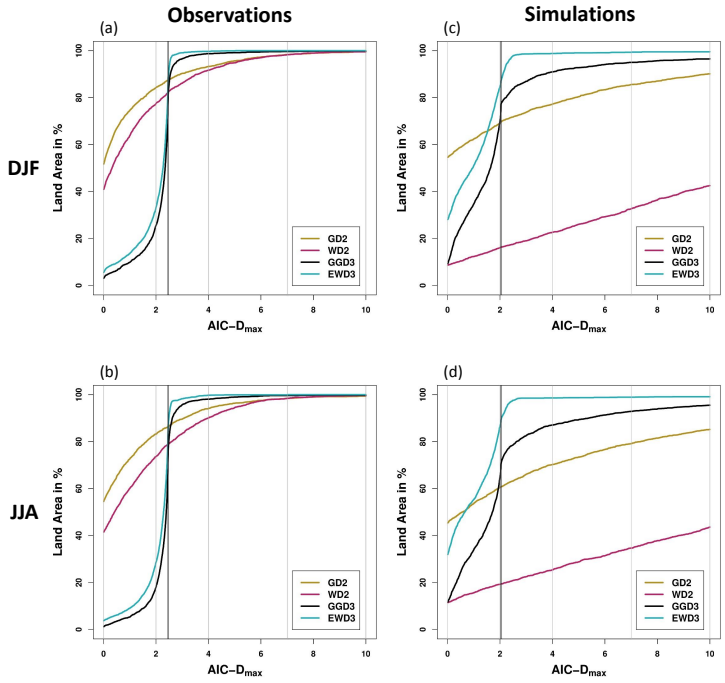

Figure 5. AIC-D frequencies: percentages of global land grid-points in which each distribution function yields AIC-D values that are smaller than or equal to a given AIC-D$_{max}$ value. The vertical black line indicates the different complexity penalties between 3- and 2-parameter PDFs. AIC-D frequencies are displayed for each candidate PDF for observations (**left**) and simulations (**right**) during DJF (**top**) and JJA (**bottom**).

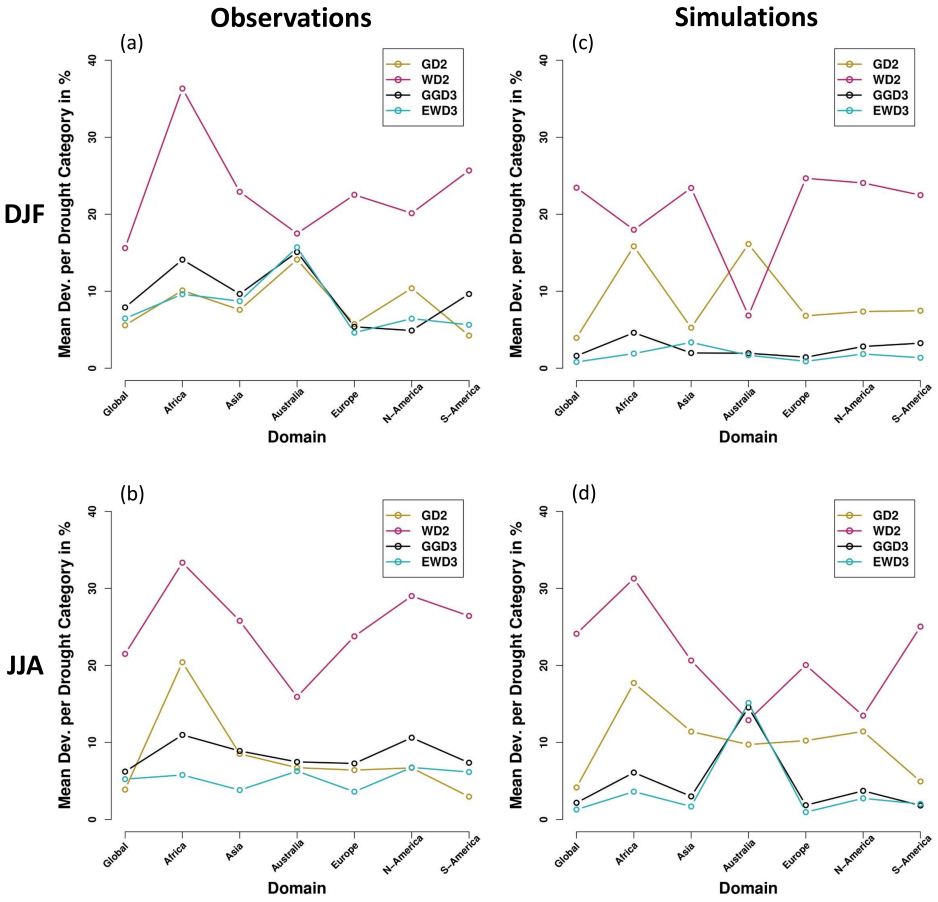

Figure 6. Mean deviations from $\mathcal{N}_{0,1}$ per SPI category for the entire global land area and each investigated ~~domain~~ region. Results are depicted for observations (**left**) and simulations (**right**) during DJF (**top**) and JJA (**bottom**).

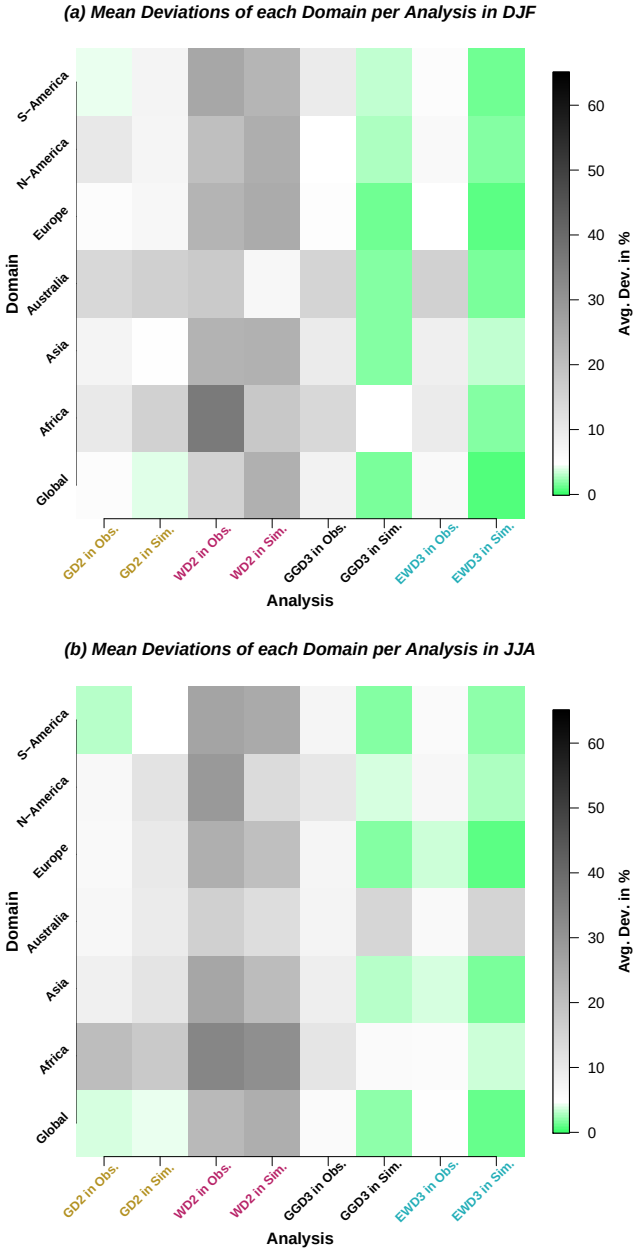

Figure 7. Mean deviations from $\mathcal{N}_{0,1}$ per SPI category during DJF (**a**) and JJA (**b**). Mean deviations are displayed for each investigated domain and each analyzed PDF for observations and simulations.

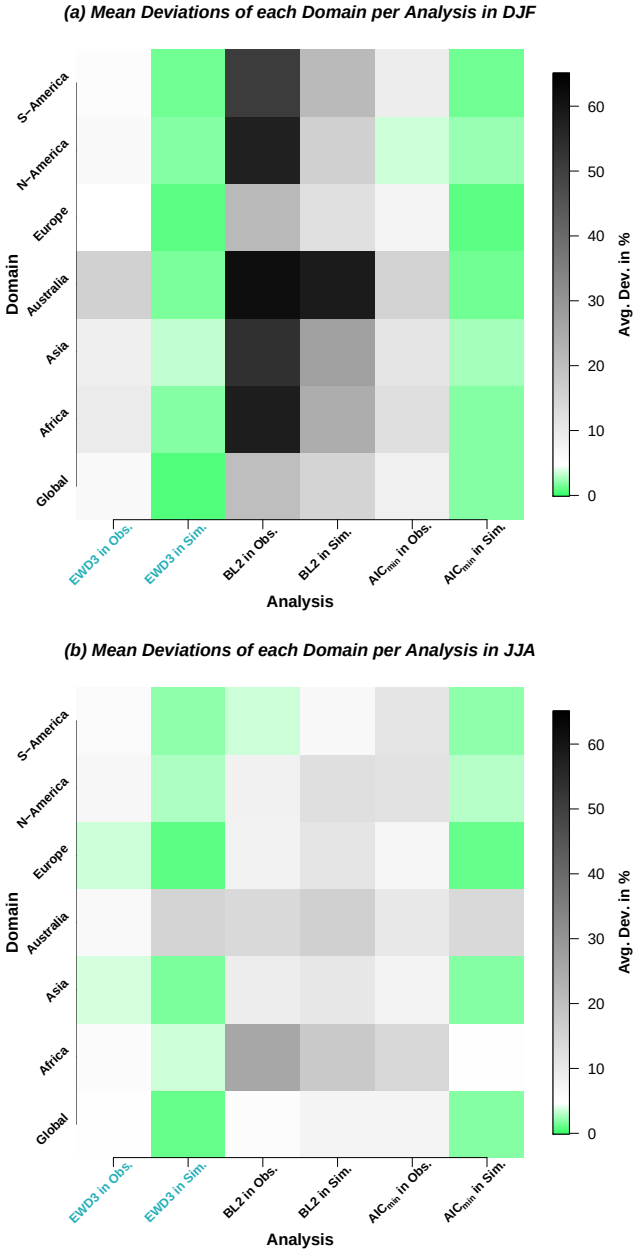

Figure 8. As in Fig. 7 but for the 3-parameter exponentiated Weibull distribution (EWD3) – the best performing candidate distribution function in this study –, a baseline which uses the 2-parameter gamma distribution (BL2)~~with a simpler parameter optimization than employed in our previous analysis~~, and a frequently proposed multi-PDF SPI calculation algorithm ~~which~~ that uses in each grid-point and season that distribution function ~~which~~ that yields in ~~this~~ the respective grid-point and during the respective season the minimum AIC value (AIC_{min}-analysis which is denoted as AIC_{min} in this figure). In contrast to GD2 in our previous analysis, BL2 employs a simpler optimization procedure of the same parameter estimation method (*maximum likelihood estimation*).