Response to Reviewer 1

Patrick Pieper, André Düsterhus, and Johanna Baehr

June 6, 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 ≤ 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**
	Single Ensemble Member	Ideal (AIC-D ≤ 2)	**83**	**76**	**19**	**28**
		Well (AIC-D ≤ 4)	**93**	**91**	**98**	**100**
		Sufficient (AIC-D ≤ 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**

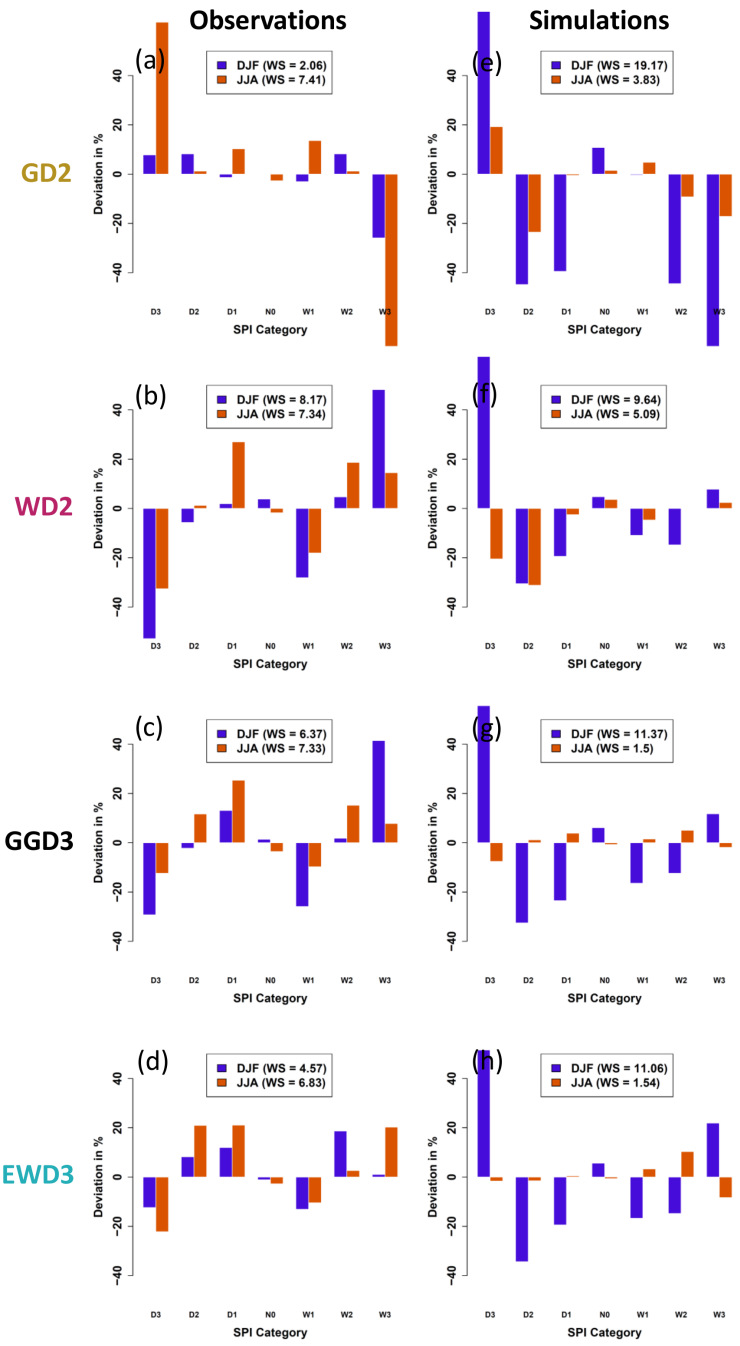

Figure I. As in Figure 4, but solely for the *North Region* of Brazil (0°– 8°S; 40°W – 60°W).

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*.