INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **36**: 2132–2138 (2016)
Published online 22 December 2015 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.4564

RMetS
Royal Meteorological Society

Short Communication

Response to comment on 'Candidate Distributions for Climatological Drought Indices (SPI and SPEI)'

James H. Stagge,[a]* Lena M. Tallaksen,[a] Lukas Gudmundsson,[b] Anne F. Van Loon[c] and Kerstin Stahl[d]

[a] *Department of Geosciences, University of Oslo, Norway*
[b] *Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland*
[c] *School of Geography, Earth and Environmental Sciences, University of Birmingham, UK*
[d] *Institute of Hydrology, University of Freiburg, Germany*

ABSTRACT: This response expands on the analysis performed in 'Candidate Distributions for Climatological Drought Indices (SPI and SPEI)' by explaining several topics in greater detail and by testing the conclusions of our original article against the claims made in the comment by Drs Vicente-Serrano and Begueria. Tests using the same 11 climate time series confirm the original findings from Stagge *et al.* (2015) that the Generalized Extreme Value (GEV) distribution produces consistently better fits. Claims that the GEV distribution exaggerates extreme SPEI values were found to be false by comparing Log-Logistic and GEV-generated SPEI values directly to the baseline normal distribution, rather than to one another. Once compared with the theoretical normal distribution, the GEV distribution was shown to better model the extreme tails, while the Log-Logistic distribution consistently underestimated these values. Analysis of the tails was shown to introduce significant uncertainty due to extrapolation regardless of the distribution. We thus strongly disagree with claims made in the comment by Vicente-Serrano and Begueria that their results clearly recommend the Log-Logistic distribution. Instead, we prove that differences tend to be small, but consistently support the use of the GEV distribution for SPEI analysis across multiple data sources and goodness of fit metrics.

KEY WORDS drought; drought index; standardized precipitation evapotranspiration index; generalized extreme value distribution; log-logistic distribution; generalized logistic distribution

Received 17 September 2015; Revised 14 October 2015; Accepted 15 October 2015

1. Introduction

We appreciate the work of Drs Vicente-Serrano and Beguería in both developing the Standardized Precipitation-Evapotranspiration Index (SPEI) and in providing this detailed comment on our article, 'Candidate Distributions for Climatological Drought Indices (SPI and SPEI)'. Their comment (Vicente-Serrano and Beguería, 2015) represents a significant amount of research and we thank them for expanding on our analysis.

It is important to clarify that the original article by Stagge *et al.* (2015) was not intended to invalidate previous SPEI studies that used the generalized logistic (Log-Logistic) distribution or to imply that this distribution is incorrect. The goal of our original article was to highlight the importance of distribution testing and to show a more generalized distribution test across a large number of locations. Stagge *et al.* (2015) recommends the Generalized Extreme Value (GEV) distribution based on the Watch Forcing Dataset (WFD) within the European domain, and while the improvement over the Log-Logistic distribution was consistent throughout all methods of analysis, this difference was minor and not likely to make a significant change in SPEI within the typical range of values. Therefore, there is no reason to negate the use of the Log-Logistic distribution or to disregard analysis or data generated from it.

Nevertheless, we strongly disagree with the claims made in the comment by Drs Vicente-Serrano and Beguería (Vicente-Serrano and Beguería, 2015, referred to here as VS&B-Comment) that their results 'clearly recommend' the use of the Log-Logistic distribution and that the results from the original article (Stagge *et al.*, 2015) 'are not robust and depend on the data used'. To test the conclusions of our original article against the claims of VS&B-Comment, this response attempts to replicate the analysis of VS&B-Comment using the same data. The results confirm the original findings from Stagge *et al.* (2015), that the GEV distribution produced consistently better SPEI fits than the Log-Logistic distribution across multiple metrics. The subsequent analysis included in VS&B-Comment is hampered by making a direct comparison between the GEV and Log-Logistic

* Correspondence to: J. H. Stagge, Department of Geosciences, University of Oslo, P. O. Box 1047, Blindern, N-0316 Oslo, Norway.
E-mail: j.h.stagge@geo.uio.no

distributions, never comparing the results to the theoretical normal distribution. When this is done, the GEV distribution is much nearer to the theoretical across all metrics used in VS&B-Comment and Stagge *et al.* (2015). In conclusion, we support the use of the Log-Logistic distribution for SPEI analysis, but believe that confirmation of the superiority of the GEV distribution using a previously untested dataset (the 11 climate stations from VS&B-Comment) supports its use as a slightly better alternative specifically for the European domain. Although not conclusive, the superiority of the GEV in non-European stations and the accuracy of reproducing extreme SPEI values in the global Climatic Research Unit (CRU) dataset suggests that the GEV distribution is superior globally as well.

2. Goodness of fit tests

The first argument of VS&B-Comment is that the statistical tests used in Stagge *et al.* (2015) are inadequate to determine a preferred SPEI distribution. In particular, Shapiro–Wilk *p*-values are cited as having poor discriminating power and the use of the Kolmogorov–Smirnov (K–S) and Anderson–Darling (A–D) tests is deemed inappropriate. VS&B-Comment instead states that distribution decisions should be made based on the tails of the distribution. This argument is further discussed in Section 3 of this response. Section 2 of VS&B-Comment makes an important point about the Shapiro–Wilk test, that goodness of fit tests cannot be used to distinguish between distributions above the significance level, typically $\alpha = 0.05$. This is a point that Stagge *et al.* (2015) never disagrees with. In fact, Stagge *et al.* (2015) never shows the distribution of *p*-values graphically or uses this distribution to select a preferred distribution. Instead the rejection rate is used, calculated as the proportion of times the null hypothesis is rejected. This is an appropriate use of the Shapiro–Wilk *p*-value, and this method is reproduced correctly in Table 1 of VS&B-Comment for the global CRU and WFD datasets. Unfortunately, the remainder of the Shapiro–Wilk analysis in VS&B-Comment is based on analysing the distribution of *p*-values, which should never be done. VS&B-Comment Figure 1 and its companion VS&B-Comment Supporting Information Figure 1 both show the *p*-value distribution graphically and use it to make claims regarding the goodness of fit for the Log-Logistic and GEV distributions. This is incorrect. Similarly, Supporting Information Figure 2 of VS&B-Comment presents the mean *p*-value for each grid cell globally, which suffers from the same error in statistical methodology and should be disregarded as well.

To replicate the findings from VS&B-Comment, we fit the GEV and Log-Logistic distribution to the same 11 time series used in this comment and ran the suite of goodness of fit tests used in Stagge *et al.* (2015). Station data were downloaded from the SPEI package in R (Begueria and Vicente-Serrano, 2015) and include the Lahore, Pakistan observatory, which provides the basis for much of Section 3 in the VS&B-Comment. As in our original article, we use several goodness of fit tests to provide a more robust comparison, pairing a log-likelihood test for the distribution fit Akaike Information Criterion (AIC) with the Shapiro–Wilk test for normality in the final SPEI values. Figure 1(a) shows the ratio of months where the GEV distribution had a better (lower) AIC than the Log-Logistic distribution for each station, while Figure 1(b) shows the Shapiro–Wilk rejection frequency for the GEV (top) and Log-Logistic (bottom) distributions.

Goodness of fit tests provided here strongly support the use of the GEV distribution, based on better AIC scores (Figure 1(a)) and fewer Shapiro–Wilk rejections of normality (Figure 1(b)) across nearly all locations and accumulation periods. For the GEV distribution, 7 of the 11 stations produced no rejections of normality (Figure 1(b)), including 2 of the 3 stations within the original European domain. The sole exception to improved fit with the GEV distribution is the Lahore, Pakistan station (Figure 1(a)), suggesting that it is not representative in terms of goodness of fit and has a unique, semi-arid climate with a distinct monsoon season. This particular station receives attention in the text and Figures 2 and 3 of VS&B-Comment. In addition to critiquing the Shapiro–Wilk test, VS&B-Comment claims that it is a mistake to use the K–S and the A–D tests to estimate goodness of fit for SPEI distributions because their distribution parameters must be pre-specified and not estimated from the data itself. This restriction is partially true, though it requires some clarification. These tests cannot use published critical values, as published critical values assume pre-specified distributions. However, both tests are applicable if the critical values are determined for the specific case (Crutcher, 1975; Steinskog *et al.,* 2007). We accounted for this by instead using bootstrapping to estimate K–S and A–D critical values for each distribution and set of parameters (see Stagge *et al.*, 2015, Section 2.7).

3. Testing distribution tails

VS&B-Comment provides a detailed analysis of the differences in the tails of the Log-Logistic and GEV distributions, claiming that the GEV distribution has a more marked decrease in the tails, which results in 'an overrepresentation of extreme SPEI values' and thus 'unrealistically high return periods.' While the GEV distribution does generally have thinner tails, relative comparisons of this type have no bearing without relating to the theoretical SPEI baseline: the Gaussian distribution (McKee *et al.*, 1993). When transformed SPEI values are compared with the standard normal, it is clear that the GEV distribution better models the extreme tails, whereas the thicker tails of the Log-Logistic distribution underestimate the severity of extreme values (see below).

The lack of a true baseline for tail comparisons is a fundamental problem in Figures 2, 3 and 4 of VS&B-Comment, which all directly compare the Log-Logistic distribution to the GEV distribution. In order

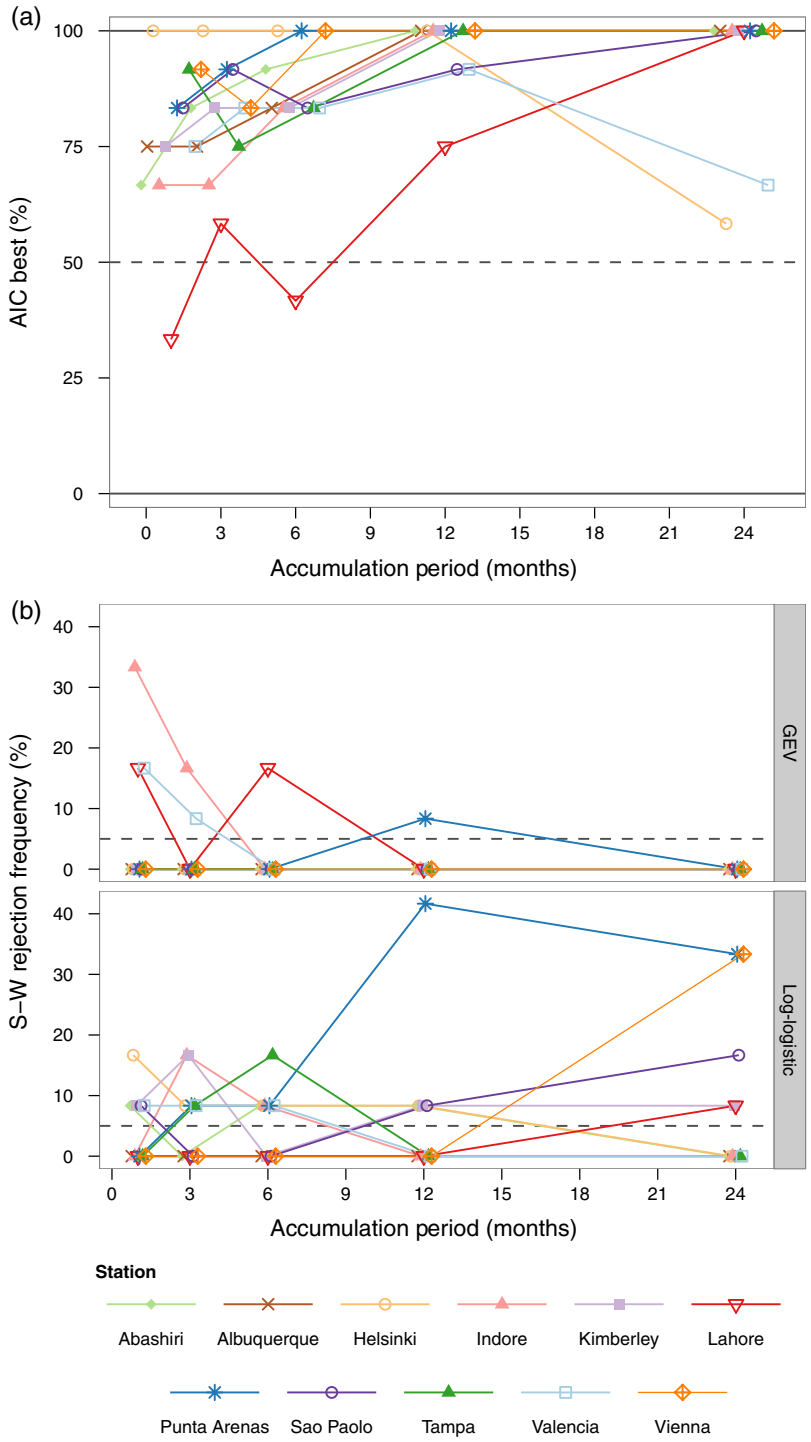

Figure 1. Goodness of fit test for the 11 stations in VS&B-Comment. Tests include (a) the fraction of months where the GEV distribution produces a better fit according to AIC and (b) the Shapiro–Wilk (S–W) rejection frequency.

to address this problem, Figure 2 of our response compares the SPEI values calculated using the Log-Logistic and GEV distributions for the 11 stations discussed previously against the theoretical SPEI values calculated from the normal distribution. The median for all 12 months and 11 stations is shown as a solid line, while the 10th–90th percentiles are shown as shaded regions. The left figure plots the fitted SPEI values directly against the theoretical SPEI, as in Figure 3 of VS&B-Comment, while the right figure rotates this figure and plots the difference between fitted and theoretical along the 1:1 axis to make more detailed comparisons.

These results confirm the finding in VS&B-Comment, that the GEV distribution produces more extreme SPEI values than the Log-Logistic distribution, i.e. in excess of SPEI ±2. However, by comparing these distributions to the theoretical normal distribution, it is clear that the GEV distribution's overestimation at the extremes is mirrored

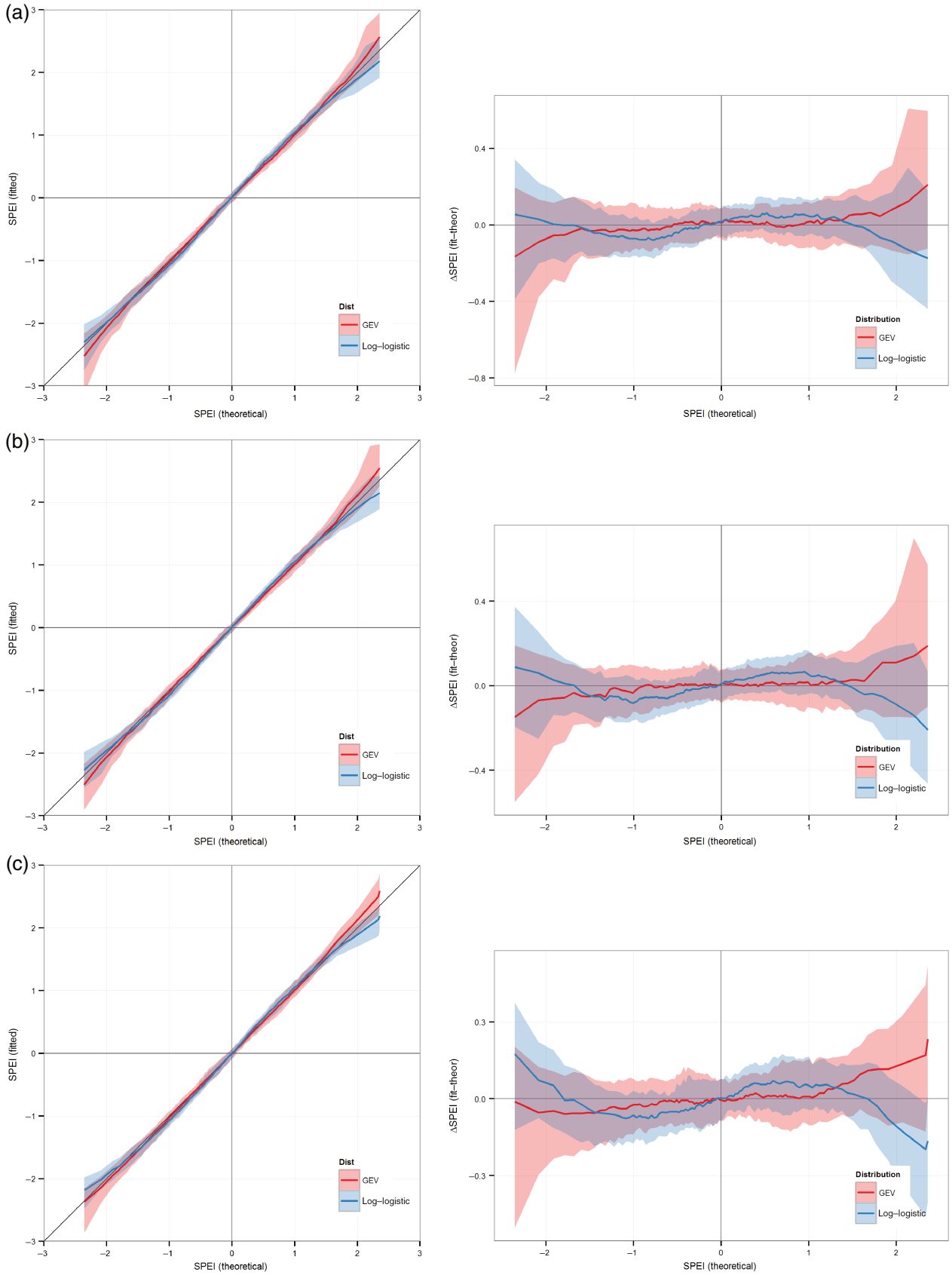

Figure 2. SPEI values for all 11 stations and 12 months plotted against the theoretical SPEI values for (a) 1, (b) 6, and (c) 12 month accumulation periods. GEV distribution is plotted in red, while the Log-Logistic distribution is plotted in blue. The line represents the 50th percentile, while the shaded regions show the 10th–90th percentiles. The figures on the right are rotated to show the difference between measured and theoretical in greater detail.

by the Log-Logistic's underestimation. This contradicts VS&B-Comment's assertion that the Log-Logistic distribution is more accurate at the extremes. Instead, both distributions are inaccurate for the most extreme values, with the GEV overestimating severity and the Log-Logistic underestimating severity. The accuracy of extreme SPEI values will be discussed in greater detail in Section 4.

It is more important to note that the GEV distribution is significantly more accurate across the range of typical SPEI values [−2, 2]. The Log-Logistic distribution tends to overestimate positive SPEI values (0 to 2) and to underestimate negative SPEI values (−2 to 0), with this effect becoming more pronounced at longer accumulation periods of 6 and 12 months. McKee *et al.* (1993) established drought classes of [0, −1), [−1, −1.5), [−1.5, −2) and [−2, −∞), which have become *de facto* definitions, often repeated in drought classification schemes. We therefore argue that producing accurate SPEI values with no apparent exaggeration across the range (−2, 2) is most important for drought monitoring and management. Based on this criterion, the GEV distribution is superior. Not surprisingly, uncertainty increases in the extreme tails, as will be discussed in the next section.

While comparing the tails of the SPEI distribution, VS&B-Comment states that *the frequencies of high and low SPEI events using the GEV are unrealistically high*, citing the number of times the SPEI exceeds the 0.5% values (SPEI = −2.58). To support this claim, VS&B-Comment plots the fraction of global cells that exceed this limit a given number of times during the 64 year CRU TS3.2 time series (VS&B-Comment Figure 6). There is a clear distinction between the GEV and Log-Logistic distributions, but as in the discussion above, results are not compared with a theoretical distribution making it unclear which is most correct.

In order to determine which distribution is more accurate, we simulated 64 year time series of monthly SPEI values randomly generated from the standard normal distribution. This method of generating synthetic SPEI time series with perfectly normal distributions was repeated for 10 000 grid cells to determine the theoretical fraction of grid cells with a given number of exceedances. This method is only valid for the 1-month SPEI because longer accumulation periods aggregate the same values, leading to temporal autocorrelation. The theoretical distribution is plotted in Figure 3(b), alongside the original Figure 6 from VS&B-Comment (Figure 3(a)). The GEV distribution from VS&B-Comment is extremely similar to the theoretical distribution generated here. This analysis is a partially blind experiment, as the data fitting and analysis was performed without any knowledge of the theoretical distribution. This further confirms that the GEV distribution better reproduces extreme statistics in the tails.

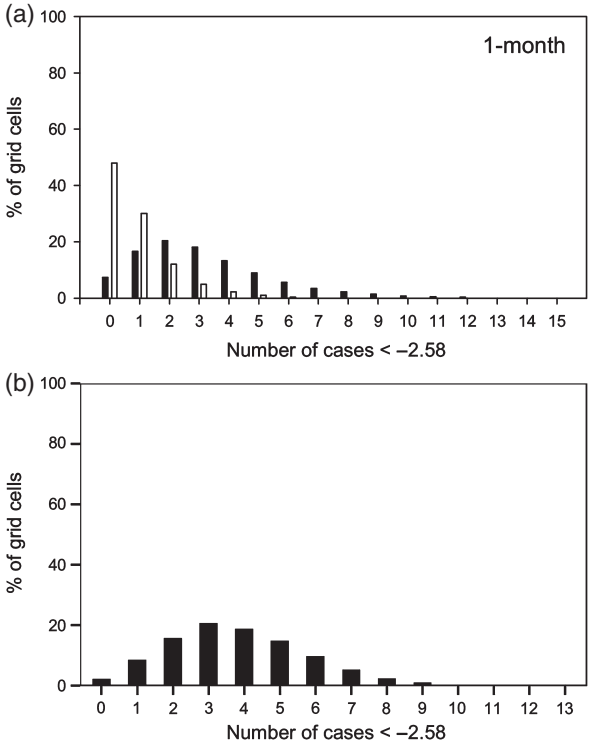

Figure 3. The global proportion of grid cells with SPEI values in excess of −2.58, reproduced from VS&B-Comment (a), with the Log−Logistic distribution shown in white and the GEV distribution shown in black. Figure (b) shows the theoretical proportion of grid cells that exceed −2.58 for a time series randomly selected from normally distributed SPEI values.

4. Confidence in distribution tails

We would like to dispel a slight misunderstanding in the rationale for introducing limits on SPEI values, proposed in Stagge *et al.* (2015). Section 2.5 of Stagge *et al.* (2015) points out that calculating extreme SPEI values based on limited historical data involves significant extrapolation and therefore significant uncertainty in extreme SPEI estimates. The article provides an example of the extreme levels of uncertainty inherent in SPEI values of −4 and goes on to suggest that the most accurate course of action would be to calculate confidence intervals on the entire time series of SPEI values. However, this is computationally expensive for large datasets. As a less desirable, but workable compromise, Stagge *et al.* (2015) suggested placing reasonable limits on the SPEI time series based on an acceptable level of uncertainty. A reasonable limit for the WFD was found to be [−3, 3], well beyond the 'extreme drought' class suggested by McKee *et al.* (1993). Values beyond this limit were not discarded, but rather retained as −3, with a note showing that this period is extremely dry, but is too uncertain to quantify accurately.

VS&B-Comment claims that this problem could be solved without the need for artificial bounds by simply using the Log-Logistic distribution, rather than the GEV distribution. In order to test this theory, we used SPEI-6 values calculated for the Lahore, Pakistan station in August, as in VS&B-Comment. Uncertainty was calculated for each point using a parametric bootstrap approach (1000 iterations), similar to the method described in Section 2.5 of Stagge *et al.* (2015). The 95% confidence intervals for various levels of SPEI are presented in Figure 4.

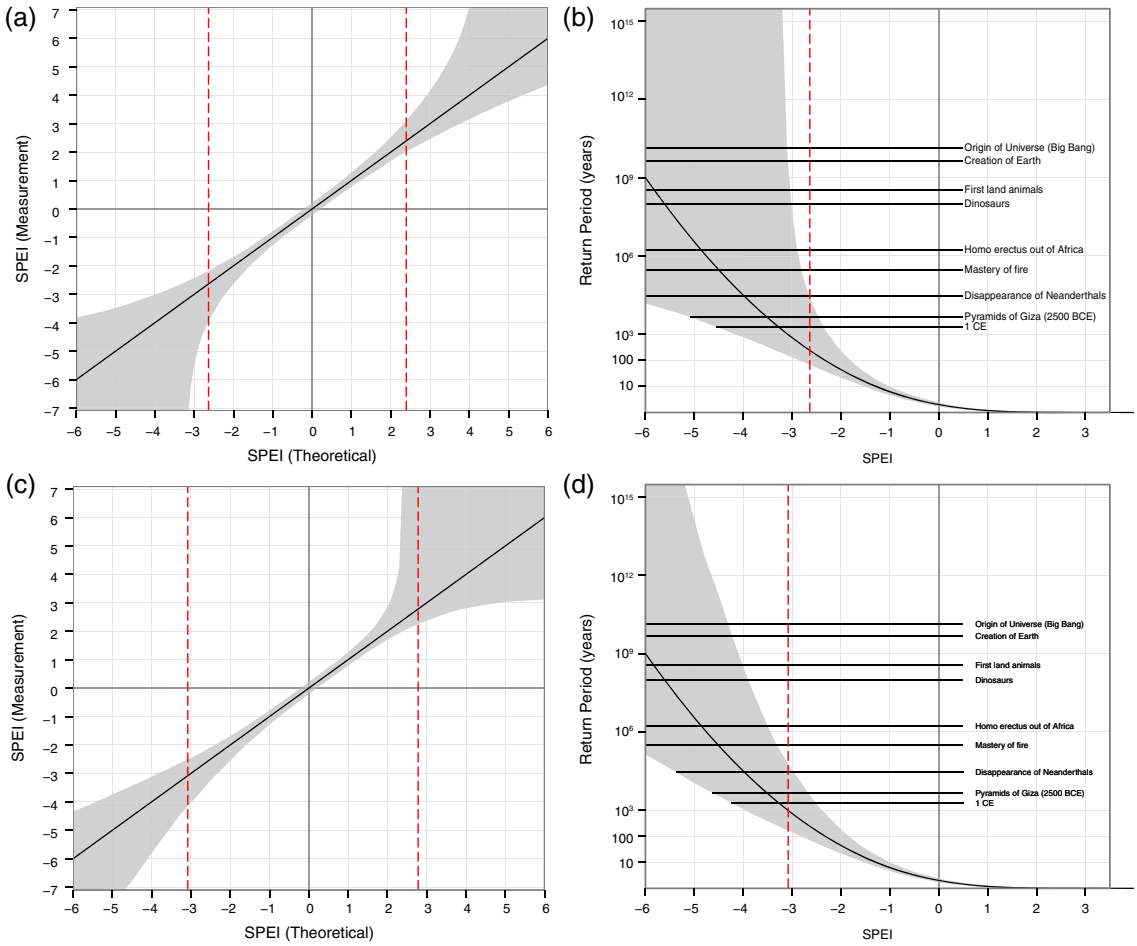

Figure 4. Confidence intervals (95%) for SPEI-6 values calculated at the Lahore station in August. Confidence intervals are shown as shaded regions for the Log-Logistic distribution as SPEI values (a) and return periods (b), and also for the GEV distribution, as SPEI values (c) and return periods (d). The red-dotted line shows the range of the maximum and minimum SPEI values calculated for this time series.

Figure 4(a) and (b) show that the 95% confidence intervals increase significantly beyond the SPEI values of ± 3 for the Log-Logistic and GEV distributions, respectively. This is reasonable, given that the minimum SPEI values for the August time series are -2.64 (Log-Logistic) and -3.08 (GEV) as shown by dotted red lines in Figure 4. These values are relatively close to the expected minimum value for 107 observations (1901–2007), which is -2.36. Figure 4 consequently highlights that estimates outside the observed SPEI range are increasingly subject to uncertainty due to extrapolation. Converting the confidence intervals from probabilities to return periods further highlights the danger of producing highly unreasonable and uncertain SPEI values by extrapolation (Figure 4(b) and (d)). In engineering practice, it is rare to see return periods longer than 100 or 500 years, so important historical dates are shown in this figure to provide context. In order for these return periods to have meaning, it is necessary to imagine that climate has remained stationary over all time periods. Confidence intervals for SPEI values and return periods are shown in the range of $[-6, 6]$, mirroring the bounds in VS&B-Comment Figure 3. The SPEI limit proposed in Stagge *et al.* (2015) is $[-3, 3]$, corresponding to a return period of approximately once in 740 years for each month. This estimate is only 95% accurate to return periods between 100 years ago and prior to the first appearance of humans outside of Africa (1.7×10^6 years ago) for the GEV distribution. It should be noted that these bounds are conservative, as they consider only uncertainty due to distribution fitting and not the uncertainty in historical climate observations. This extreme level of uncertainty, independent of distribution choice, was the original rationale behind the argument for SPEI limits in Stagge *et al.* (2015).

5. Fitting differences

Results for the Shapiro–Wilk test using the CRU and WFD in VS&B-Comment (Table 1) and using the WFD in Stagge *et al.* (2015) (VS&B-Comment, Figure 5) are in direct opposition, with VS&B concluding that the Log-Logistic distribution produces fewer rejections and Stagge *et al.* (2015) concluding that the GEV is better. This reversal may result from differences that occur due to minor differences in methodology and could be caused by comparing global results using CRU data with European results using WFD. Additionally, Stagge *et al.* (2015) used

a daily time step, resulting in 365 distributions for each grid cell, whereas VS&B-Comment used monthly data, producing 12 distributions for each cell. Use of a daily time step allowed for more detail and a larger sample population in Stagge *et al.* (2015). This daily time step was the true reason for discussing temporal autocorrelation in Stagge *et al.* (2015), not a concern for excessive extreme values, as suggested in VS&B-Comment.

Regardless, it is unlikely that the difference in spatial domain and time step would account for all the recorded difference between the independent analysis of VS&B-Comment and Stagge *et al.* (2015). We suspect that much of this difference derives from the fitting procedure, as suggested in VS&B-Comment. VS&B-Comment and Vicente-Serrano *et al.* (2010) both use Probability Weighted Moments (PWM) for distribution fitting. Stagge *et al.* (2015) uses the same PWM method to set initial parameter values, but these are then input into a maximum-likelihood estimation (MLE) scheme. VS&B-Comment shows that use of MLE decreases the anomalously extreme SPEI values they found when using a PWM fitting method. Further evidence pointing towards the fitting procedure as the root of this difference is the relatively high proportion of distribution fitting failures for the GEV in VS&B-Comment. Our use of PWMs for initial values and MLE for the final fit affords the advantages of both methods: speed and stability from PWM and accuracy from MLE. We experienced no fitting failures, using either the WFD or the 11 stations for both the GEV and Log-Logistic distributions. Given our lack of fitting failures and overall statistical support for the GEV distribution across all tests, we believe that the proportion of cases with no solution and slower processing times highlighted in VS&B-Comment is not a compelling argument for selecting the Log-Logistic distribution, but rather an indication of fitting difficulties for the GEV that resulted in poor goodness of fit.

6. Conclusions

VS&B-Comment and Stagge *et al.* (2015) both agree that continued analysis and careful testing are important to ensure that the SPEI is a robust drought index that can be used globally to measure, monitor and forecast meteorological drought. Further, we would like to reiterate that Stagge *et al.* (2015) and this response is not intended to invalidate prior research using the Log-Logistic distribution to normalize SPEI values. Overall, the differences between the Log-Logistic and GEV distributions tend to be small. However, the results presented here strongly refute claims made in VS&B-Comment that the Log-Logistic distribution should be clearly recommended and that the selection criteria outlined in Stagge *et al.* (2015) are inadequate.

Replicating the analysis of VS&B-Comment using the same 11 time series confirmed the original conclusions of the European study, i.e. the GEV distribution produces a consistently better fit for the SPEI. Claims that the GEV distribution produces an overabundance of extreme SPEI values were found to be false by comparing each distribution directly to the baseline Gaussian distribution, rather than to one another. The GEV distribution was thus shown to be more accurate and the Log-Logistic distribution specifically found to under-represent extreme values. Uncertainty in extreme SPEI values was also shown to be a function of extrapolation that cannot be resolved by distribution selection, as suggested in VS&B-Comment. Finally, distribution fitting failures were not experienced in our analysis and thus do not represent a strong argument for choosing the Log-Logistic distribution in light of the statistical support for the GEV distribution. Thus, we support the use of the Log-Logistic distribution for SPEI analysis, but respectfully believe that the GEV represents a better alternative specifically for the European domain, but possibly the larger globe.

References

Begueria S, Vicente-Serrano SM. 2015. *"SPEI: Calculation of the Standardised Precipitation-Evapotranspiration Index" R Package Version 1.6*. https://cran.r-project.org/web/packages/SPEI/SPEI.pdf

Crutcher HL. 1975. A note on the possible misuse of the Kolmogorov-Smirnov test. *J. Appl. Meteorol.* **14**: 1600–1603.

McKee TB, Doesken NJ, Kleist J. 1993. The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, Anaheim, CA, USA, 17–22 January, American Meteorological Society, Boston, MA, 179–184.

Stagge JH, Tallaksen LM, Gudmundsson L, Van Loon AF, Stahl K. 2015. Candidate distributions for climatological drought indices (SPI and SPEI). *Int. J. Climatol.* **35**: 4027–4040, doi: 10.1002/joc.4267.

Steinskog DJ, Tjøstheim DB, Kvamstø NG. 2007. A cautionary note on the use of the Kolmogorov–Smirnov test for normality. *Mon. Weather Rev.* **135**: 1151–1157.

Vicente-Serrano SM, Beguería S. 2015. Comment on 'Candidate distributions for climatological drought indices (SPI and SPEI)' by James H. Stagge et al. *Int. J. Climatol.*, doi: 10.1002/joc.4474.

Vicente-Serrano SM, Beguería S, López-Moreno JI. 2010. A multi-scalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index – SPEI. *J. Clim.* **23**: 1696–1718.