Reviewer 2

Summary:

In this paper, the authors introduce a multi-site multi-variable stochastic weather generator called "PRSim.weather" to assess the (joint) occurrence probabilities, severity, and spatial patterns of compound hot-dry events in the US at various time scales (1 week, 1 month, 3 months, 6 months, 1 year). The proposed weather generator is a simple extension of a previously published version for a single variable, and they here make some necessary adjustments for its application to study high temperatures / low precipitation. The authors conclude that their model correctly replicates the distribution and dependencies in observed data, and their analysis further reveals that

(1) Northwestern/Southeastern US are more likely to experience hot-dry events

(2) the time scale influences the size of compound hot-dry events (i.e., shorter time scales imply larger spatial extents of joint extreme events)

(3) temperature mostly determines compound events for short time scales, while precipitation is the key factor for longer time scales.

Assessment:

Overall I like the paper and the data analysis. The topic tackled by the authors, namely to understand the spatio-temporal distribution of compound extreme weather events, is difficult and timely. The paper is well written, is relatively concise and the authors precisely detail the findings of their analysis. The proposed approach (PRSim.weather) has, however, some limitations that the authors should, I think, better acknowledge and discuss more openly. I discuss some of those in my comments below. Another point to mention is that although the data analysis and the findings are well supported and of practical interest, the methodological novelty is rather limited, since the proposed method is a simple adjustment of an already published approach.
Reply: *We thank the reviewer for highlighting the value of our work and for indicating the need to discuss the limitations of the stochastic simulation approach in more detail. We added methodological clarifications where suggested and expanded the discussion section by discussing model limitations in more detail.*

Comments:

1. I found the 5 steps of the method on page 4 (lines 90-109) difficult to understand. For example,
Reply: *Thank you for highlighting the need for methodological clarification.*

- how do you "fit monthly distributions to T and P"? do you first a distribution to the data within each month separately assuming that they are iid during that month?
Reply: *Yes, we specified that: '(i.e. one separate distribution is fitted to the data in each month)'.*

- what do you mean by "we combine the E-GPD with as many zero-values as in the observations"? Do you mean that you don't simulate zero observations, but keep them fixed like in the data? If so, is this not "cheating" (i.e., over-fitting)? and do you keep the zeros at the same time points?

Reply: *Yes, the E-GPD is only used to simulate the non-zero part of the distribution similar to most existing stochastic precipitation simulation approaches. However, the zero-values are not pinned to the same time points as in the observations to enable temporally varying precipitation patterns. This reordering is achieved thanks to the rank-ordering in Step 5. We rephrase the description for clarification: 'We use the E-GPD to simulate non-zero precipitation values and complement it with as many zero-values as in the observations to obtain the full P distribution with appropriate probability of precipitation occurrence.'*

- how do you apply the continuous wavelet transform? and how to interpret the amplitude and phase signals?

Reply: *The continuous wavelet transform was performed using the Morlet mother wavelet and the R-package wavScalogram (function cwt_wst). A reference and the equation of the continuous wavelet transform and the Morlet wavelet were added to the manuscript. The amplitude tells us about the strength of variability at different time scales while the phase tells us about the time shift in the data.*

- in point 4., what do you generate a random time series for both T and P? Or just one time series?

Reply: *We specified that we generate one random time series based on the temperature time series.*

- in point 5., how to you do the "rank-transform" exactly? Do you mean that you apply the probability integral transform?

Reply: *Yes, the rank-transformation is achieved by applying the probability integral transform, which we now specified in the text.*

Bottomline: I think it is needed to clarify the methodology. It seems necessary to me to provide further mathematical equations to clarify each point and to illustrate the wavelet transform with a simple example in order to faciliate interpretation.

Reply: *We added mathematical equations to explain the SEP and E-GPD distributions, the continuous wavelet transform and the Morlet wavelet. A schematic illustration of the procedure is provided in Figure 1.*

2. The methodology seems to have certain limitations that may be concerning:

- The authors mention that the same random phases are used at all sites and for both variables. Is this not too restrictive, and will this not create too strong spatial or cross-dependencies?

Reply: *Using the same phases across stations and variables allows us to model spatial and variable dependencies. Without following this procedure, we would produce local simulations, not retaining the spatial dependencies we would like to reproduce. An example of what happens to spatial dependencies if non-identical phases are used across stations is provided in Figure A2 in Brunner and Gilleland (2020; https://hess.copernicus.org/articles/24/3967/2020/). While this step is essential to model spatial and variable dependencies, it is true that neither spatial nor variable dependencies are perfectly represented. In the case of spatial temperature dependencies, we see a slight overestimation (Figure 4) while T-P dependencies are slightly underestimated (Figure 3g-h). Jointly* *modeling temporal, spatial, and variable dependencies is very challenging and we therefore consider model performance to be satisfactory for our application. We acknowledge these model limitations in the methods section and added an additional statement to the discussion section: 'However, spatial dependencies are slightly overestimated while variable dependencies are slightly underestimated. The model still has acceptable performance across three types of dependencies - temporal, spatial, and variable - and enables studying rare spatial multivariate events, which would not be possible using observations only.'*

- In point 4., a time series of one site is chosen at random. Are all sites "exchangeable"? What is the implication of this approach?

Reply: *Yes, the stations are exchangeable as the goal is to generate a random time series with some seasonality. Using a totally random series, e.g. white noise, won't allow us to reproduce temperature seasonality, which is why we are using a random series mimicking the temperature signal.*

- Again in point 4., a random time series is generated by bootstrap by resampling years with replacement. This implies that years are exchangeable and therefore that any time trend is ignored. Is this not a major issue for temperatures (and perhaps also precipitation)? If so, this should be further acknowledged and discussed.

Reply: *It is correct that PRSim.weather is a stationary model, i.e. potential time trends are not considered. This is not an issue in this study as we do not aim to look at temporal trends in compound event characteristics. Adapting the model to non-stationary conditions would primarily require the introduction of non-stationary distributions for P and T in Step 2 of the modeling procedure. If one would in addition want to consider potential non-stationarities in spatial and/or variable dependencies, one would have to use alternative resampling schemes in Step 4 retaining the temporal order of the original series. We add a short paragraph to the discussion stating that 'Extending model application to non-stationary conditions would require the implementation of non-stationary distributions for both T and P. For example, one could introduce covariates for certain parameters of the marginal distributions of T and P in Step 2 or introduce covariates with information about trends or variability in P and/or T to guide resampling in Step 4.'*

- Using a bootstrap-based approach implies implicitely that simulated events will NEVER be more extreme than what has been observed in the data. This is a major limitation since the goal here is to enrich the dataset with more simulations of compound extreme events.

Reply: *This statement is true for classical bootstrap approaches. However, PRSim.weather is a semi-parametric model, which combines a non-parametric bootstrap model to represent spatial and variable dependencies with two parametric models for temperature and precipitation. This may not have been clear in the previous version of the manuscript and we added that: 'Using theoretical instead of empirical distributions will allow us to generate extreme values more extreme than the observations.'*

- Estimating a copula using the empirical copula (based on ranks) implicitely implies that the data are stationary over time, thus without time trend (or seasonality) again. Is this a reasonable assumption here?

Reply: *The STI and SPI time series do not show a time trend in most grid cells and we think that using an empirical copula is appropriate especially because it is a non-parametric model.*

3. L129, p5, "site-specific Gamma distribution": should this not be the E-GPD distribution as specified in the methods section (point 2.)?
Reply: *Yes, we corrected this by replacing 'Gamma distribution' with 'E-GPD' distribution.*

4. p6, top: further details on copulas are required to introduce the notation properly...

- What is a copula => Joint distribution with uniform Unif(0,1) margins

- What is C(u,v)? => the copula of T and -P

- What are the ranks R_i and S_i? => ranks of T or P values across the time series

- In Figure 2, what does "Empirical copula" mean? => the values of C_n(R_i/(n+1),S_i/(n+1)), i.e., the empirical copula evaluated at the observed uniform values.

Reply: *We clarified the notation as suggested.*

5. In Figure 3, the results are almost too good to be true in my opinion. Does this not hide some issues of overfitting? Again, how do you simulate the zeros in precipitation for example?
Reply: *We use a four-parameter distribution to fit temperature and a three-parameter distribution to fit precipitation. These distributions are flexible enough to reproduce the main distributional characteristics of P and T. Non-zero precipitation values are added separately as often the case with precipitation distributions, e.g. when combining a Markov Process with a parametric distribution. One could use less flexible distributions with less parameters, which would decrease simulation performance.*

6. When the goal is to simulate many more compound events, it is crucial to check if the marginal and joint tails are captured correctly. For marginal tails, I would suggest to consider comparing long-term return levels of simulated vs observed data (on a scale that zoom into the tail rather than the bulk). For joint tails, a possibility is to look at the tail correlation coefficient (\lambda(u) = P(U1>u | U2>u)) for increasing thresholds u=0.8,0.9,0.95,0.98,0.99,0.995,0.999, say. Such diagnostics would complement the results in Figure 3.
Reply: *Thank you for these suggestions. We estimate the 100-year return levels for T and P and all grid cells for both observed and simulated series. The comparison of observed and simulated return levels shows that observed and simulated return levels estimated using the SEP for temperature and the E-GPD for precipitation are very similar (Figure 1 in this response to the reviewers). This additional analysis confirms that the SEP and E-GPD are indeed good choices to model T and P, respectively. We also compute upper tail dependence for different thresholds for high T and low P values. The tail dependence between extremely low P and high T is 0. The simulations reproduce this behavior.*

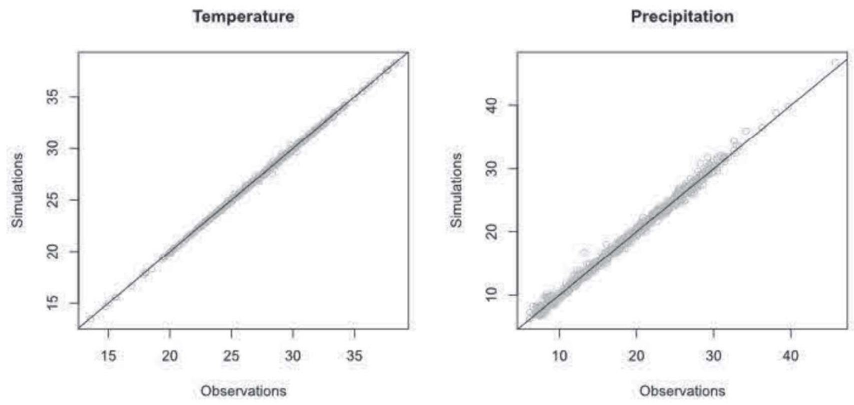

Figure 1: Observed vs. simulated 100-year return levels for (a) temperature (°) and (b) precipitation (mm/d).

7. In Figure 4, the simulated fields appear smoother than observations. Why is that the case?
Reply: *In the case of temperature, this is indeed true. This slight overestimation in spatial dependence possibly comes from the phase randomization procedure which relies on random phases generated from bootstrapped temperature time series. As mentioned earlier, we added this point to the discussion.*

8. In Figure 5, it seems like the spatial extent of very extreme events is largely overestimated. Is this because a single random phase is chosen across sites? Or is this a false impression due to the fact that there are less extreme events available in observations than simulations?
Reply: *Figure 5 maps median observed and simulated STIs and SPIs at a grid scale. It therefore shows that simulated STIs and SPIs are more extreme than observed ones. The simulated medians are more extreme because the model is able to generate more extreme events than in the observations thanks to the theoretical distributions used to simulate T and P distributions. We specified in the figure caption that the median events refer to 'a certain grid cell' and that 'While the simulated spatial STI and SPI patterns look similar as the observed ones, they are more expressed because of the larger sample available, which contains yet unobserved extremes because of the use of parametric distributions for simulating T and P'.*

9. In Figure 6, simulations severley underestimate the joint probability of concurrent events for severe and extreme events... and also for moderate events in the Southeastern part of the US... Is this due to using the empirical copula approach? What is the cause of this and how to remediate this (fairly severe) issue?
Reply: *The underestimation of the co-occurrence probability of compound events is related to the underestimation of T-P dependence as illustrated in Figure 3, acknowledged in the Methods section and discussed in the Discussions section. The reduction in variable dependence is introduced in the backtransformation step, which can hardly be avoided (Embrechts et al. 2002; Correlation and dependence in risk management: Properties and pitfalls). A potential improvement of the representation of variable dependence may be achieved by using phase annealing, which modifies the phases in an iterative way in order to optimize certain statistics but increases the computational effort (Hoerning et al. 2018; Phase annealing for the conditional simulation of spatial random fields).*

10. Figure 8 plots the "median spatial extent of concurrent events affecting grid cell". How was that calculated? I don't think it is clearly explained in the text...
Reply: *Thank you for pointing out the need for clarification. We specified in the Methods section that: 'To assess the spatial extent of compound events at different time scales, we define the spatial extent of the compound event as the percentage of grid cells affected by the compound event at any given time scale. Then, for each grid cell, we determine the median spatial extent of those events it is affected by.'*

11. Figure 10 reports the values of Kendall's tau between T and the bivariate empirical copula, as well as between P and the bivariate empirical copula. However, given that the empirical copula is itself calculated from T and P, I'm not convinced that such "correlation" values make sense... Wouldn't it make more sense to report the actual ranks R_i/(n+1) and S_i/(n+1), which already give the importance of T and P in the calculation of the empirical copula?
Reply: *With this part of the analysis, we intend to explain which of the two variables is related most strongly to the empirical copula, i.e. represents the main driver of the compound event. Reporting just the ranks would not allow us to provide a measure of association and ranks for both variables would range from 1 to n.*