Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2020-31-RC3, 2020

Earth System Dynamics
Discussions

Interactive comment on "Evaluating the dependence structure of compound precipitation and wind speed extremes" *by* Jakob Zscheischler et al.

Anonymous Referee #3

Received and published: 25 August 2020

Disclaimer:
Even though I read the whole paper and appreciated both the methodological and applied aspects of this research, my review mostly revolves around the statistical contributions of this paper, which I'm more confident to comment on.

Summary:
In this paper, the authors propose a new statistical metric to compare the bivariate joint tails of two different datasets. This metric, which relies on the Kullback-Leibler (KL) divergence based on the count of points in certain number of "extreme sets", provides a single number that can be used to assess whether or not the joint tails are different,

and if so, by how much they differ. It is proposed as being complementary to more classical measures, such as the χ-measure introduced in the paper that is widely used in extreme-value theory. The proposed KL metric depends on the number of sets, W, which has to be defined by the analyst, and is "non-parametric" in the sense that it does not rely on stringent model assumptions. In the paper, the proposed metric is used to estimate the likelihood of compound precipitation and wind speed extremes derived from different climate model outputs.

General assessment and general comments:
In my opinion, the paper is well written and concise with interesting practical results. Methodologically, the proposed metric is well-grounded but is not particularly novel as the Kullback-Leibler divergence (here based on the multinomial distribution) has been used extensively in other areas of statistics. The novelty probably relies on its specific use to study bivariate extremes and to compare bivariate joint tails of extreme precipitation and wind speed, although it is based on a previously published paper by one of the authors (Naveau et al., 2014, JRSS B). Overall, I like the paper and find the results quite interesting, yet several questions remain unanswered. My general and specific comments below mostly focus on the statistical contributions of the paper.

1. The χ measure is computed based on "local block maxima". I think it is easier to understand what the χ and $\bar{\chi}$ measures represent when used with the original daily data, rather than with block maxima. With original data, if $\chi = 0$ this implies asymptotic independence of daily data, but how should we interpret it with block maxima? It would be good to add a few lines or a short paragraph to better explain the statistical meaning and the practical interpretation of the proposed metrics (χ, and KL-based) when they are used with block maxima. And why did you choose to compute χ based on block maxima and not block means or block minima? What is the rational behind this choice? Is it somehow more informative to compare joint tails?

2. A major question that remains unclear to me is what do we gain with the proposed KL measure? As pointed out by the authors on page 5, we could compute a measure

$\chi^{(1)}$ based on the first dataset, and another measure $\chi^{(2)}$ based on the second dataset and compare their values. The authors argue that they want just a single number to assess whether the tails are different and by how much. I get that. But why not simply doing a formal statistical test of whether $\widehat{\chi}^{(1)}$ is statistically different from $\widehat{\chi}^{(2)}$?? The test statistic (or the corresponding p-value) would indeed be a single value that could be used to assess whether the tails are different, and by how much. Moreover, the proposed KL metric is χ^2-squared distributed ASYMPTOTICALLY, while testing for $\chi^{(1)} = \chi^{(2)}$ could—I think—be done EXACTLY for finite n (or be based on the corresponding asymptotic normal distribution). A partial answer to my question above ("what do we gain with the KL measure?") may be that the KL measure is probably more informative for testing whether the joint tails are different because it relies on full distribution of counts within extreme sets, rather than only on information about "the diagonal $F_1(X_1) = F_2(X_2)$"... but without a proper simulation study, this is difficult to claim (especially that the KL measure depends on the choice of W). It would be good if the authors could elaborate on that, and complement the paper with a short simulation study to assess the gains of the KL measure compared to a simple test $\chi^{(1)} = \chi^{(2)}$.

3. This point is related to the point 2 above, but I split it into two parts so the authors can more easily address the several questions that I have. Another major question related to the proposed KL measure is how to set the number of extreme sets, W, to use. In the paper the authors choose $W = 3$, but there is no optimality with this choice. In fact, while the proposed KL measure is not well-defined when at least one of the sets is empty, the more classical χ-measure is always well defined (so testing $\chi^{(1)} = \chi^{(2)}$ is always possible). This is a major drawback of the KL measure, I think, since under asymptotic independence we should EXPECT that the probability mass will concentrate on the axes (on the Pareto scale) with no point in the interior (so extreme sets should be empty in the limit!). Of course, in practice, there will always be points in the interior and ways to ensure that the extreme sets are non-empty, but it still raises the question of how to choose the number of sets W and the sets themselves. A related question is what is the efficiency of the statistical procedure for different numbers of

sets, W? In my opinion, it would be good to complement the paper with a simulation study, in order to investigate this issue in more details and come up with some concrete advice for practitioners on the selection of sets. Is there an "automatic" way to do this "well"?

4. Another major point that is unclear to me is the treatment of marginal distributions. I assume that margins are estimated non-parametrically (with ranks) to compute the χ-measure, and that the extreme sets are defined based on data transformed to a common scale (e.g., Pareto), but there is no mention of marginal modeling in the paper. Does it matter here, since the KL-measure is non-parametric? I think this should be clarified. Marginal modeling usually has a major effect on the final results and their interpretation, so care is needed. In particular, how was the uncertainty related to marginal modeling taken into account (if it was)? The authors mention a bootstrap procedure for the χ-measure, but does it take marginal estimation uncertainty into account or does it only account for the estimation of the dependence structure?

5. Figures 5-6: Even if I understand why the authors chose different block sizes (i.e., spatial lags and temporal windows), I find it difficult to interpret the results in Figure 5 given that the color in each pixel represents the tail dependence of potentially completely different events based on different block sizes. This may also explain why the figure looks a bit "noisy". Wouldn't it make more sense to produce such a map for each block size separately, and then present only the "most relevant" one (or potentially 2 block sizes of interest)? In my opinion, this would be much easier to interpret.

6. Although the authors cite relevant papers related to extreme-value theory, some general review papers (or classical textbooks) could be added in my opinion to help non-experts navigate through this extensive literature.

Specific comments:

1. Page 2, Line 29: "studies studies"
2. Page 5, Line 119: If I'm not mistaken, the $\bar{\chi}$ measure has been introduced in a paper

by Coles, Heffernan and Tawn (1999) published in Extremes, not by Ledford and Tawn (1996). Please add this reference.
3. Page 5: Line 126 says "inspect their behavior as $q \to 1$" but Line 128 says "We generally estimate χ at $q = 0.95$". I agree and I get what the authors want to say, but these two sentences sound a bit contradictory. Please reformulate.
4. Page 5, Lines 149-150: you mention the sum and the minimum as the risk function $r(x)$. Why not considering the maximum, as well, which is perhaps more commonly used than the minimum?
5. Page 6, Line 155: write "$A_w^{(j)}$" instead of "A_W"?
6. Page 6, Line 164, "The statistic d_{12} follows a $\chi^2(W-1)$ distribution is the limit": Do you mean "in the limit as $n \to \infty$"? Also, is this valid under the null hypothesis that the tails are the same? Please clarify.
7. Page 6, Lines 181-182, "$q = 0.95$ and $u = 0.9$": why did you choose different numbers? Does it matter?
8. Page 7, Line 199: write "In particular, in the south of the Alps" (add "in the")
9. Page 7, Line 213-215: Table 1 shows the results are different as u increases. What do you conclude? And what if q increases?
10. Page 8, Line 224: write "Because the model setting determines the dependence structure" (add "the")
11. Page 8, Lines 228-229: the sentence "This is to ensure ... (e.g., Foehn)" sounds odd to me. Please consider rewriting.
12. Figure 3: The difference in tail behavior for the two datasets from $q = 0.8$ is already quite clear based on the χ-measure. This comes back to my general comments above: do we really need the new KL metric to detect this?

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2020-31, 2020.