Answer to comment of referee #2

Predicting climate extremes - a complex network approach

M. Weimer, S. Mieruch, G. Schädler and C. Kottmeier

Dear referee,

Thank you for your review of the paper. In the following, you can find our answers to your comments which are written in red text color.

1 General comments

The title and abstract are in my opinion misleading. The approach presented here does not aim to improve the prediction of climate extremes. Instead, the authors propose a different way to quantify the skill of an initialized regional climate model to reproduce observed numbers of heat periods. The standard way to define such a skill would be to compare the numbers of observed and modeled heat periods. As an alternative way, the authors propose to compare the (suitably normalized) number of observed heat periods with the correlation threshold used to construct climate networks from the model data.

We agree with the referee and adopt the title and improve the abstract.

Their result is that in some regions and for some decades, the model skill quantified using the correlation threshold is better than the standard one, and sometimes it is not. It should be emphasized here that the result does not imply any improvements of the model's predictions, but only shows that a different and less direct skill definition (based in the correlation threshold instead of directly comparing the numbers of heat periods) leads to different results, with sometimes higher skill. Furthermore, the result is not really quantified (see Fig.7 which compares the two skill definitions in terms of better/worse/tie), and no tests of statistical significance of this result is performed (in the sense of: "what is the probability of getting a rank matrix like the one in Fig.7 from some suitable null hypotheses"). See also point 4. below.

Although we agree with the referee that the statistical significance is an important point, we refer to a 2013 paper entitled "Testing ensembles of climate change scenarios for "statistical significance"" by climate statistics instances Hans von Storch and Francis Zwiers (von Storch and Zwiers, 2013) who claim that "... a statistical null hypothesis may not be a well-posed problem ..." and "Even if statistical testing were completely appropriate, the dependency of the power of statistical tests on the sample size n remains a limitation on interpretation.". We therefore followed and will follow von Storch and Zwiers (2013) who "... propose to employ instead a simple descriptive approach for characterising the information in an ensemble ...".

The problem is that we have so many factors involved in the analysis, i.e. the models themselves, the downscaling, the ensemble, the initialisation, the different regions, the filtering, etc. that any nullhypothesis would be not well-posed and any test would be questionable.

Nevertheless we will perform a significance test, but we will not completely rely on the significance, especially because we are dealing with small sample sizes and the power of the test is questionable.

According to the referees comment: "what is the probability of getting a rank matrix like the one in Fig.7 from some suitable null hypotheses" we construct the following significance test.

First, we have to define what is the possibly "significant" characteristic of the matrix in Fig.7. It is, as we concluded in the paper, that the network method is superior in 3 regions. Regarding Fig.7 we have 3 lilac entries in region 4, 4 lilac entries in region 5 and 3 lilac entries in region 6. Thus the question is: "**What is the probability to observe at least two regions with at least each having 3 lilac entries and at least one region with at least 4 lilac entries by chance?**". So, we developed a short algorithm and constructed matrices like in Fig.7. Given the nullhypothesis of random entries we coloured randomly 20 fields lilac and 20 fields red. Afterwards we coloured 10 fields white as in Fig.7. Finally we repeated the algorithm 1000 times and counted the cases, where we observed, as mentioned above, at least two regions (rows) with each at least 3 lilac entries and at least one region with at least 4 lilac entries. The result is that we found such patterns in 10% of the cases. Thus the probability of observing such a rank matrix by chance is 10%. This is not strongly significant, however far away from being random. We will include this analysis in the revised version.

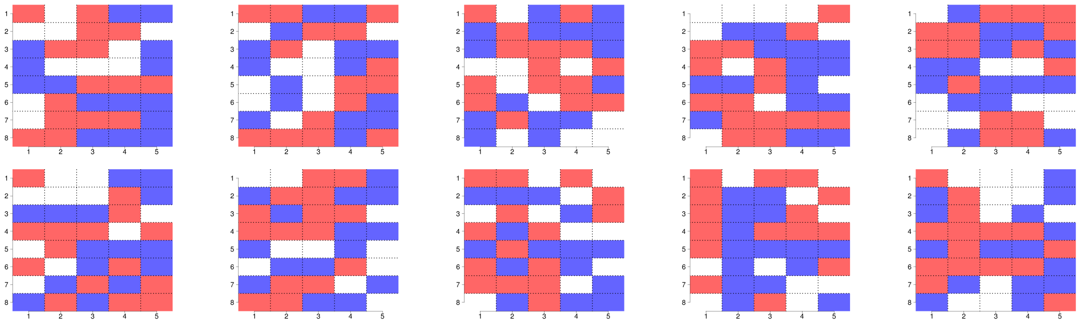

Figure 1: Example of randomly generated matrices. The probability of observing at least two regions with at least each 3 lilac entries and at least one region with at least 4 lilac entries is 10%. In this example, the matrix in the middle of the bottom row fulfills the criterion.

2 Specific comments

1. For all networks, a constant edge density of 0.3 is enforced, implying that the network correlation threshold is simply the 70th percentile of all correlation values for a given region and time span (P70). Hence, i think that the network terminology is quite overdrawn in this work, since the only "measure" employed here is this P70. The wording "link strength" is furthermore unfortunate in this context, since strength also refers to the sum link weights of a node in a weighted network.

First we like to thank the referee for pointing to this simple method of calculating the link strength, i.e. the P70. We will mention this fact and implement this effective way of calculation in our method.

However we disagree with the referee that the network terminology is overdrawn. As explained in our paper on page 1484, from line 19 we have a clear physical/statistical motivation why the link strength can be used as a heat period estimator. The whole idea would have not been emerged without the network terminology, thus it is of utmost importance to understand the analysis under the perspective of complex climate networks. The complex climate network theory is the frame for this analysis. Using P70 would be completely arbitrary without the network terminology, and as pointed out by the referee, it is just a method to effectively calculate the link strength as used by us.

One key point of the analysis, which highlights that the analysis would have been impossible without the network terminology is explained on page 1489, from line 7. We performed several sensitivity studies and investigated the network patterns. By visual inspection we found that correlation in the order of 0.7 to 0.9 yields good results. So what are good results? We think that it is optimal for our purpose to extract as much variability from the data as possible. Thus the networks observed with correlations of 0.7 to 0.9, i.e. using an edge density of 0.3 yield the most variable patterns, from very few connections to a lot of connections and nuances inbetween.

Finally, P70 is only a way to calculate something effectively, the network terminology is the necessary frame to justify the analysis.

Regarding the wording, as far as we know, in weighted networks the measure "node strength" is used, whereas we call the correlation threshold "link strength", according to Berezin et. al (2012). Thus, we think it is ok, especially because it is clear from the context.

2. The authors first show that in artificial (Gaussian) time series, P70 increases linearly with the number of heat periods included in the time series (Fig.3). Furthermore, based on the observations (E-OBS), they show that there is some dependence between the number of heat periods and P70 (Fig.4). The authors infer from this that P70 is a good predictor of the number of heat periods. I don't agree to this statement: In the case of artificial time series, you trivially increase their correlations by putting in additional heat periods, because the time series become "more similar" by construction. P70 can be interpreted as an estimator of the overall correlations between given time series at hand, and hence their coherence. If a region experiences a spatially extended heat period (more than 20% of the region with temperatures above a threshold, as you defined it), it is to be expected that the corresponding time series behave more coherently, as expressed in higher P70.

This comment from the referee seems to be contradictory. First the referee doesn't believe that there is a connection between heat periods and P70, although it is clearly shown in Fig.4. In the last sentence the referee states that it is expected that there is a connection between heat periods and P70.

We don't know exactly what to answer to this comment. But this is exactly our motivation behind the study, as the referee pointed out "it is to be expected that the corresponding time series behave more coherently, as expressed in higher P70.". Thus we see no problem here.

Additionally, the tests with artificially data are of course constructed to produce the wanted outcome. But this is the general idea behind such approaches and shows that some simplified ideas, can yield good assumptions for real problems.

However, any other spatially coherent behavior, such as a cold period, would lead to higher values of P70 as well. P70 measures nothing but the "spatiotemporal" coherence, and heat periods can be examples of this, but many other phenomena could be just as well.

This is an important point, which have also been raised by the other referee. First of all, we use daily maximum temperatures in summer. Thus coherent phenomena are most probably heat periods. Second, Fig.4 shows the good agreement between the number of heat periods and the link strength, both from observations. Thus, it shows that the method works well. Nevertheless we will improve the method by leaving out the coldest 10% of data from the analysis to absolutely remove any coherent cold phases.

3. I agree with you that comparing the observed number of heat periods with the modeled number of heat periods suffers from the "threshold problem" that the model amplitude might for a given observed heat period just closely not exceed the threshold. However, I can think of some simple ways to overcome this: In particular if an ensemble of projections is available, on could define that a heat period is detected in the model runs if some percentage of the ensemble members exceeds the temperature threshold.

We absolutely disagree with the referee. The ensemble cannot overcome the threshold problem. The problem would only be shifted from a deterministic prediction to a probabilistic one. If a heat period truly occurs and the e.g. 10 member ensemble shows this heat period in 3 out of 10 ensemble members, one would conclude that the probability of the occurrence is low, i.e. only 30%. Our approach is absolutely independent from the threshold and although only 3 members would have crossed the threshold, the other members could have been slightly below the threshold, still able to coherently correlate with their neighbours. Thus, the network method has the potential to detect the heat period, independently of how many ensemble members cross any static threshold.

I am not convinced that taking P70 instead, which is ambiguous in this context for the reasons explained above, is a good alternative. Instead of taking the modeled number of heat periods, you propose a much less precise "proxy" - namely P70 - for heat periods. Loosely speaking, this proxy could be seen as a "randomization" around the modeled number of heat periods, and the results shown in Fig.7 could hence simply reflect this randomization: Sometimes the model skill is better according to your skill definition than the standard skill, and some times vice versa.

We disagree with the referee. As explained above our approach is not ambiguous, it is physically/statistically rooted in the network ideas and it works well using observational data (Fig.4). The interpretation of Fig.7 is indeed not an easy task. Too many factors are involved to really uncover the causes why one method is better than the other or why they perform similar in some cases. Crucial points are the general low skill of the model, the initialisation, the different regions, the downscaling and so on.

There are other questions, e.g. why should the method perform equally well in whole Europe? Perhaps it works very well for some regions (4,5,6) and in others not. If this would be true it would be great success to use our method in these regions and the standard method in other regions.

Or our method can be used in addition to the standard approach to gain confidence. If two independent methods yield the same results, the confidence is definitively increased.

Finally, assessing our results being random is a too simple conclusion as shown in the significance test above. We have shown that a network approach can successfully be used to predict heat periods. It overcomes the problem of static thresholds used in the standard way of detecting heat periods. Additionally, it is the first time that such network techniques have been used in climate predictions. Since climate or decadal predictions aim to predict natural variability in the order of years suitable statistics are needed. Natural variability in the order of years evolves highly dynamical and often nonlinear. Thus, the complex climate networks could bear the potential to be very useful in climate predictions. Our approach, which is even based on the most simple network measure, the node degree (or as we used it the link strength) yield optimistic results. So, we think that our analysis could be the starting point for using the complex networks in climate predictions, using other measures and/or multivariate data could turn out to be the better way of analysing predictions of natural variability years ahead than using methods from short- or medium range forecasting.

References

von Storch, H. and Zwiers, F. W.: Testing ensembles of climate change scenarios for "statistical significance", Climatic Change, 117, 2013.