Answer to report #2 of referee #1

Decadal prediction of heat periods based on regional climate model data – a complex network approach

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

Have you actually ever checked how well your network method is at identifying heatwaves in the observational data? I do not find anything on this matter in the manuscript. You should make a classic prediction skill analysis and state the rates of false positives, false negatives and so on.

For us, it is not really clear what this comment raised by the referee is addressing at. In our here presented analysis we quantify the prediction skill of a model in predicting heat waves. Therefore we use model data (CCLM) and observational data (E-OBS) and compare the number of heatwaves in the model with the number of heatwaves in the observational dataset. Additionally we compare the number of heat waves in the observations with the new network quantity link strength.

This is explained in Sect. 4.4: “To quantify the prediction skill of the model, we calculate the absolute mean difference (see Eqs. 6 and 7) between the number of heat periods in E-OBS (o) and CCLM (m) and the CCLM link strength.”

Furthermore, the main figures in the paper show exactly what the referee is addressing at. For instance Fig. 6 shows on the left y-axis the number of heat periods in E-OBS (observations) and on the right y-axis the CCLM link strength. How well these curves fit together shows how well the new method can estimate heat waves in observational data.

Or is the comment more related to the exact method of calculating the skill? A classic skill analysis, counting rates of false positives, false negatives and so on is not suitable for our analysis. Since we are working with decadal predictions, we cannot predict a single event. Thus we used the skill measure defined in Eqs. 6 and 7, which fits exactly to our needs and represents well the prediction skill on decadal time scales.

Or is it, that it is not clear that E-OBS data are observations? To exclude such a misunderstanding, we will better describe the E-OBS dataset in the new version.

So, what do you consider a "true" heatwave, then?? Is your whole approach based on a new idea of how heatwaves should be defined? If yes, then please explain this new idea.

We replace “true heat period occurs” with “heat period according to the definition at the very beginning of this section occurs”

Concerning the second question: Yes it is. The whole idea is described in Sect. 4.3 and we will refer to this section at this point.

The way you write it here ("static threshold" and so on), I get the impression that you do not intend to apply any bias correction method while, in fact, later you implicitly apply one in your percentile-based "standard" heatwave detection method.

In the sentence after “static threshold” on page 6 we write that this problem also exists if the threshold is adapted to the model.
This is what we do in our analysis: Of course, it right that we adapt our threshold to the model data. However, the threshold is static for the whole time series. Therefore, the threshold problem exists within each ensemble member of the decadal predictions.
We will add a subclause which points out that we do this adaption.

"[...] but also to consider complex long-term climate evolution in contrast to short-term weather": I do not understand what you mean by that.

We wanted to say that climate predictions generally include higher uncertainties and different dynamics than short-term weather forecasts. Thus new methods are needed to quantify the prediction skill of decadal model runs. We will rewrite this sentence.

Page 13, equations (6) and (7): This quantification of prediction skill is most probably useless. As far as I understand you want to predict the number of heatwaves a given summer will be struck by. You should look at how well your network metric actually reproduces the interannual variability of heatwave counts.

No, we don’t want to predict the heat waves occurring in a single summer, this is not possible in our decadal prediction experiment. This is e.g. mentioned on page 11: “... since we are interested in decadal variability, and since we do not expect the model to represent the year to year fluctuations ...” Thus we cannot expect to reproduce interannual variability.
As can be seen from Eqs. 6 and 7 they measure how well model and observations fit within a decade, they are perfectly suited for our experiment. Thus they show how well the model reproduces the decadal variability of the observations.

The figures in the supplement tell a pretty different story than Fig. 4. You should not choose your best example for the paper and hide your bad examples in the supplement. If you want to objectively assess the prediction skill of a method, you will have to talk about failures as openly as about successes.

The figures in the supplement do not “tell a pretty different story”. It is clear that the method works better in some regions than in others. We see no problem to show the region, where the method performs best to show the potential of the new method. There are other regions, where the method performs equally well e.g. Prudence Region 4.
Concerning the referee’s comment on “hiding” the other figures in the supplement: In our opinion, the supplement is easily accessible and an opportunity to show the different time series of all regions.

We discuss the strengths and weaknesses of our new method objectively and show all results e.g. in Fig. 7 in the rank matrix.
However, we will smooth out any potential misconceptions and include three new figures in the new version (now Fig. 7). They will show the “M” values of all decades and regions of Figs. 4 to 6. So, we show

1. how well the network method works with only observations (corresponding to the mentioned Fig. 4) for all regions and all decades.

2. how well the standard method (like Fig. 5) works for all regions and all decades.

3. and how well the network method (like Fig. 6) works for all regions and all decades.

Sorry to say this, but I think that this significance test is a good example for why quite many statisticians these days are worried about flawed research being hidden behind statistical significance. (See, e.g., http://amstat.tandfonline.com/doi/abs/10.1080/00031305.20 or read Nate Silver’s The signal and the noise.) I wrote above that I do not consider the metrics defined in equations (6) and (7) appropriate for any quantification of the skill of you heatwave identification method. Consequently, I consider any significance test based on any application of these methods useless in terms of telling anything about whether your method is any good.

Since the referee apparently cannot specify what the problem with our test might be, we cannot react on this comment. Regarding the ASA statement, we fully agree with this document and see no conflict to our analysis. A similar interesting paper is “Testing ensembles of climate change scenarios for "statistical significance"” by climate statistics experts Hans von Storch and Francis Zwiers, which we cite in our paper and which is worth to read. The main point is that it is a good idea to analyze results visually and against the backdrop of the underlying physics, models, observations, hypotheses and so on to finally achieve a satisfying answer. Often significance tests are difficult to design, often hypotheses are not well-posed or it is even not possible to capture all uncertainties. In these cases a well-suited statistical test (as in our analysis) can support the findings already made. In this sense, we see our test as an additional support to the found results, which can be analyzed very well visually. We also explained this in our manuscript on page 15.
As explained above, Eqs. 6 and 7 perfectly fit to our experiment. The problem is the misconception that the referee is thinking about year to year prediction, but instead we want to predict decades. Thus Eqs. 6 and 7 only consider means over whole decades and that is exactly what we want.

Answer to report #1 of referee #2

Decadal prediction of heat periods based on regional climate model data – a complex network approach

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

It is still not stated clearly that the authors do not propose a method to predict heat periods, but rather propose a different way to define a skill for the prediction of heat periods using a regional climate model: in the modified title, it says "prediction of heat periods", and in the abstract the authors write "We show that the skill of the network measure to predict the low frequency dynamics of heat periods is superior to the typical approach". In the main text, it is claimed "that the network method is clearly superior in three regions", and in the conclusion: "We found that the network approach is superior (significance is 5%) to the standard approach in predicting heat periods in Europe", and that the network approach is is "the more robust estimator of heat periods" These sentences are, at least, misleading.

We apologize for not properly having understood this issue in the first review. We agree with the referee that we cannot predict heat periods themselves with our new estimator and will adapt the points of the referee and where else it is not formulated clearly.

Especially, we will change the title to: “A new estimator for heat periods in decadal climate predictions – a complex network approach”

As explained in my previous review, the authors use P70 as a proxy for heat waves, and quantify the skill of heat wave prediction by CCLM by comparing the development of P70 in the CCLM data with observed heat waves, instead of comparing the observed number of heat waves with the modelled number of heat waves (the latter would be the standard approach)

Actually, we do both in our analysis, see Figures 5 and 6 where we compare the number of heat periods of E-OBS with that of CCLM and the link strength in CCLM, respectively.

However, P70 is only an indicator of coherent behavior of the time series. More heat waves trivially lead to more coherent behavior of the time series, but so do, as I already mentioned, more cold episodes, or any other periods of joint behavior. This is the reason why I agree that P70 is correlated to the number of heat waves, but still question the suitability of P70 as a predictor for heat waves (which the authors found contradictory). The standard way of assessing the forecast skill is to compare the number of observed and simulated heat waves. Instead, the authors propose to compare the number of observed heat waves with a characteristic of the simulated data (P70) which is not in one-to-one correspondence with the number of heat waves, but may instead react to many other climatic features as well (given that it only measures coherent behavior of the time series, no matter if daily maximum temperature are employed).

As described in Sect. 4.2, we restrict ourselves to daily summer maximum temperature time series and diminish the influence of cold periods by excluding the lower 10 %-quantile from the analysis. Thus most probably the dominating coherent phenomena are heat periods.

The essential question raised by the referee's comment is: Which other meteorological circumstances apart from heat periods influence the surface temperature on synoptic scale (i.e. essentially the scale in the order of the whole considered network area) and on time scales of several days?

Heat periods in the sense defined in Sect. 3 always occur in situations where a controlling (i.e. vertically covering the whole troposphere) high-pressure system on the one hand leads to large-scale subsidence and evaporation of clouds. On the other hand, warm air masses are transported from the south to the European continent due to the high-pressure system. We conclude that heat periods are combined with controlling high-pressure systems.

Another dominant influence on the temperature of synoptic scale are fronts which usually are combined with low-pressure systems. However, warm and cold fronts by definition separate air masses of different temperature (and moisture). This means that the time series in the considered area will include non-coherent behavior since temperature does not change for all time series at the same time. Therefore, the link strength should be actually decreased by fronts.

A third reason for a change in temperature on synoptic scale is due to large-scale rain where low temperatures of the droplets and evaporation both lead to a decrease in the surface temperature. However, as in the previous example it is most unlikely that rain occurs on the whole Prudence region at the same time.

Therefore, we can conclude that the link strength applied to daily maximum surface temperature time series is a suitable measure for heat periods in the way we implemented it.

2. I don’t think that the significance test proposed by the authors is appropriate. [...] With this model, I get a probability of [about] 0.2 that I have 5 or more regions with more blue than red entries, indicating that the author’s result would only be significant at a confidence level of 0.2 [...] Alternatively, one could simply ask for the probability of having at least 18 blue entries (the observed number) in total (regardless of how they are distributed in the single rows) in the same setting as above (note that the expectation value for this is 16!), for which I get [about] 0.2 again.

First, we want to point to the discussion on significance, which we have included in the revised version, including the von Storch, Zwiers reference. Again, we think that in this study it is more important to visually inspect the results, e.g. Figs. 4,5,6,7, and while including the knowledge about decadal predictions, the model, the observations, the physical mechanisms and so on, deriving satisfying answers or solutions. We see the statistical test as a supporting instrument to the analysis, quite conscious about the difficulties in e.g. defining a null hypothesis.

The crucial point here is that many tests can be developed for our analysis, e.g. these suggested by the referee. Another test could be using only blue and red entries without any “undecidables” and so on. Further tests are thinkable e.g. in a Baysian framework. All these tests have their eligibility under certain assumptions and hypotheses. There is no perfect test, but there are tests, which fit better and tests which fit worse to our study. Thus, we developed a test, which is as close as possible to our results, and this test is to use 16 blue, 16 red and 8 white entries, as it has been observed. The suggestion by the referee to use 40% of entries red, 40% blue and 20% white is on the mean the same, namely 16 entries red, 16 blue and 8 white, but can crucially vary for single realizations. Thus the referee’s test can yield an outcome of e.g. 25 red, 12 blue and 3 white. Such an outcome and all others with white not equal 8 is not possible in our test. Hence, the referee’s test has more degrees of freedom than our test, thus more possibilities to achieve 5 or more regions being colored blue and so a larger probability of 0.2. Concluding, we think our test, which is as close as possible to the observed outcome, is more suitable than any other test, which includes additional assumptions (i.e. more or less than 8 whites), which cannot be proven.

[...] For example, there might be a period of intermediate (i.e., neither very high or very low) temperature anomalies that are coherent over a region of interest in the CCLM data, which is detected as a high P70 and according to the authors "translated" to the occurrence of a heat period when compared to the number of actually observed heat periods in that decade, ultimately resulting in a high skill value although something completely different was actually going on.

The referee here constructs a special situation to show that our method fails for such a situation. So, let’s have a look at this situation although Figures 4 and 7 indicate that our method works and that our method is superior to the standard method. Unfortunately the situation constructed by the referee is not very well explained and we hope to understand it correctly. The gedankenexperiment focuses on a decade. Within this decade, there is a period of intermediate temperature anomalies that are coherent. Let's summarize some facts:

- to dominate the link strength during this decade, the mentioned period of intermediate temperatures must be relatively long, otherwise it would have no effect
- if this is the case, the referee is right, the link strength would be high
- since it is a period with intermediate temperatures, no real heat waves occur in the observations during this period
- as said, the period must be relatively long, so it can be expected that only few heat periods have been detected in the observations in this decade

Concluding, we have a decade with high link strength in the model and few heat periods in the observations. Thus, the difference between link strength and number of heat periods would be large and thus the skill low (Eq. 7). That means, the conclusions from the referee of observing high skill in this situation is wrong. We would observe low skill indicating that the method does not work.

In contrast, in our paper, we have seen that the skill is good and the method works, thus it is quite unlikely that a situation as constructed by the referee occurs in reality. From a meteorological point of view it is the question why the temperature time series should be coherent in this situation. If there is no significant synoptic influence on the temperature it is dominated by local effects such as local rain, wind, cloudiness and sunshine duration. Therefore, the time series should not be highly correlated in this situation. This would yield a low link strength.

That means, in reality, a decade with a long period of intermediate temperatures would yield few heat periods in the observations and a low link strength, yielding a small difference (Eq. 7) and thus large skill, which shows that the method works.

Manuscript prepared for Nonlin. Processes Geophys. Discuss.
with version 2014/09/16 7.15 Copernicus papers of the LaTeX class copernicus.cls.
Date: 20 July 2016

A new estimator of heat periods for decadal climate predictions – a complex network approach

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

Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology,
Karlsruhe, Germany

Correspondence to: M. Weimer (michael.weimer@kit.edu)

Abstract

Regional decadal predictions have emerged in the past few years as a research field with high application potential, especially for extremes like heat and drought periods. However, up to now the prediction skill of decadal hindcasts, as evaluated with standard methods is moderate, and for extreme values even rarely investigated. In this study, we use hindcast data from a regional climate model (CCLM) for eight regions in Europe and quantify the skill of the model alternatively by constructing time evolving climate networks and use the network correlation threshold (link strength) as a predictor for heat periods. We show that the skill of the network measure to ~~predict~~ estimate the low frequency dynamics of heat periods is superior for decadal predctions with respect to the typical approach of using a fixed temperature threshold for estimating the number of heat periods in Europe.

1 Introduction

Decadal prediction is a relatively new field in climate research. Skillful prediction of climate from years up to a decade would be beneficial for our society, economy and for a better adaption to a changing climate. Within the large international Coupled Model Intercomparison Project Phase 5 (CMIP5 Taylor et al., 2012) global decadal predictions of climate key variables like temperature and precipitation have been performed with state-of-the-art Earth system models. In order to validate the prediction skill of the models so called hindcast experiments are conducted. That means, the models are initialized with observations e.g. in 1961 and then run freely for 10 years and stop at the end of 1970. In 1971, the models are again initialized and start to run for another 10 years and so on. More advanced approaches of initializing every year have followed as well. These hindcasts can be evaluated against observational data to quantify the prediction skill of the models depending on the lead time, which is the time range between the initialization and the forecast datum of interest. In recent years, several studies on decadal predictions have shown the potential of these initialized (global) model runs (e.g. Keenlyside et al., 2008; Müller et al., 2012; Matei et al., 2012; van Oldenborgh et al., 2012; Corti et al., 2012; Doblas-Reyes et al., 2013; García-Serrano et al., 2013; Smith et al., 2013; Meehl et al., 2014; Chikamoto et al., 2015). However most studies concentrate on regions like the Tropical Pacific or North Atlantic and on slowly evolving variables like sea-surface temperature. These regions receive their predictability from large scale processes like the Atlantic Meridional Overturning Circulation (AMOC) or Pacific Decadal Oscillation (PDO) and thus allow to extract predictable signals out of the noise. To be useful for society, and climate change adaption, regional climate predictions are required which should provide skillful forecasts on smaller regions, shorter periods, and include climate extreme events on populated land areas like the European continent. The European climate is more connected to short term processes like the North Atlantic Oscillation (NAO), which is to a certain extent predictable on seasonal scales, whereas the decadal predictable signal is weak (Scaife et al., 2014), which has been shown also for temperature and precipitation in large projects like ENSEMBLES (MacLeod et al., 2012). Further, the complex orography with the Alps in the center contributes to a manifold of general weather situations and hence to a complex climate (e.g. World Climate Research Program Coordinated Regional Downscaling Experiment for Europe (CORDEX-EU), Jacob et al., 2013; Giorgi et al., 2009). Nevertheless, the European continent is influenced by the AMOC and thus this process may yield to a certain predictability, although the signal to noise ratio is most probably small. Up to now, the prediction skill for Europe is weaker than for such regions as the South Pacific or North Atlantic. Mieruch et al. (2014) have used a regional decadal hindcast ensemble for Europe and detected moderate prediction skill for summer and winter temperature and summer precipitation anomalies within the lead time of five years. Eade et al. (2012) analyzed the predictability of temperature and precipitation extremes in a global model and found a moderate but significant skill (correlation) for seasonal extremes. They also found skill beyond the first year, but this skill arose from external forcing. Thus, Eade et al. (2012) compared initialized climate predictions with uninitialized projections to evaluate the skill gained by initializing and excluding the external forcing. They found that the "... impact of initialization is disappointing".

Another relatively new field in climate research has been established, namely the complex climate network approach. The general idea of climate networks is to consider climate time series e.g. at the grid points of a climate model as nodes of the network and the statistical connection between the time series as links of the network. A link between two arbitrary time series (geolocations) exists, if the correlation measure between the time series exceeds a certain threshold.

The climate network community has been very active in recent years. Tsonis et al. (2007) proposed “A new dynamical mechanism for major climate shifts” and explained e.g. decadal shifts in global mean temperature (Tsonis and Swanson, 2012). Radebach et al. (2013) discriminated different El Niño types using the network approach, Ludescher et al. (2013) developed a network method to improve El Niño forecasting and Donges et al. (2011) revealed a connection between (paleo-) climate variability and human evolution using recurrence-networks, which are similar to the complex climate networks. Generally, it has been shown that climate networks contain useful information for climate applications, e.g. the relation between climate and topography found by Peron et al. (2014), dynamics of the sun activity using visibility graphs (Zou et al., 2014) and the prediction of extreme floods Boers et al. (2014).

In this paper, we exploit the idea to use an alternative heat period estimator, based on complex climate networks, and show that its skill is superior to the typical approach of using a fixed temperature threshold for prediction of heat periods on time scales up to a decade.

In Sect. 2 we introduce the daily maximum temperature data used in this study and motivate our approach in Sect. 3. Section 4 describes our approach, which includes the preparation of the data, the definition of heat periods and the construction of time evolving climate networks. The results for applying the new approach to hindcasts are shown in Sect. 5. Finally, we give the conclusions and an outlook in Sect. 6.

2 Data

We apply the climate network approach to a decadal prediction ensemble generated within the German research project MiKlip (Mittelfristige Klimaprognosen, Decadal Climate Prediction, e.g. Kadow et al. (2015)) by the regional COSMO model in CLimate Mode (COSMO-CLM or CCLM) Doms and Schättler (2002). CCLM has been used in numerous studies recently e.g. in Kothe et al. (2014); Dosio et al. (2015), a comprehensive overview can be found here: http://www.clm-community.eu. CCLM has been used to downscale global decadal predictions from the Earth System Model of the Max Planck Institute for Meteorology (MPI-ESM, Stevens et al., 2013). From a suite of different decadal prediction experiments we have selected the so-called regional baseline 0 ensemble. This ensemble consists of 10 members, each covering the period 1961–2010 for the European region (according to CORDEX-EU Jacob et al., 2013; Giorgi et al., 2009) on a 0.22° grid. This ensemble has already been used by Mieruch et al. (2014).

The regional baseline 0 ensemble (based on the global MPI-ESM model) has been initialized every 10 years (1961, 1971, 1981, 1991, 2001). Within a decade the CCLM model runs freely, except for the prescription of the atmospheric boundary conditions by the global MPI-ESM model.

More details on the development of the ensemble and the initialization can be found in Matei et al. (2012), Müller et al. (2012), Mieruch et al. (2014).

In the study presented here we use daily maximum near-surface temperatures from the CCLM model and from the E-OBS v8.0 gridded climatology (Haylock et al., 2008) for the European continent. The E-OBS data basically are measurements interpolated to a regular latitude-longitude grid.

For our comparison, we use the so-called Prudence regions http://prudence.dmi.dk/, namely British Isles, Iberian Peninsula, France, Central Europe, Scandinavia, Alps, Mediterranean and Eastern Europe shown in Fig. 1.

3 Motivation

Generally, heat periods are maximum temperature values persisting for several days and occurring on spatially expanded regions. This means that many temperature time series (grid points) behave in a “cooperative mode” (see e.g. Ludescher et al. (2013)). This cooperative state can be described by the link strength, i.e. essentially the correlation between time series, of a climate network. Thus, the link strength of a climate network could turn out to be the better heat period estimator for model data, because it is independent of the typically critical thresholds used in classical extreme value detection.

The standard estimator for heat periods according to the World Meteorological Organization (WMO) is that the daily maximum temperature is 5 K above the 1961–1990 mean maximum temperature at five consecutive days at least (Frich et al., 2002). Thus, the standard method to compare the prediction skill of heat periods between observations and model would be to count the heat periods e.g. for each year in an observational reference data set and similarly in the model data, both according to the WMO definition (cf. Fig. 2).

A crucial problem of the standard estimator for model predictions is the inherent static threshold used to detect heat periods. Although this threshold can be adapted to the model climatology (as we do it in Sect. 4.2) the problem is that it is still likely that the model slightly undershoots or otherwise slightly misses the threshold if a ~~true heat period~~ heat period according to the definition at the very beginning of this Section occurs, assuming the model exhibits at least some predictive skill.

To account for this situation in decadal predictions we propose a new method, based on complex climate networks, to detect heat periods, which is independent of a fixed temperature threshold ~~.~~ (see Sect. 4.3). Again we want to emphasize that no new method for the detection of heat periods is needed, if past observational data ~~is~~ or short-term forecasts are used. The WMO based definition works well. However, ~~for detecting heat periods in 10-yearly initialized forecast/hindcast data, new methods are needed. Not only to overcome the threshold problem, but also to consider complex long-term climate evolution~~

~~in contrast to short-term weather.~~ the increased uncertainty in decadal predictions requires new methods to handle climate extremes like heat periods.

The following schematic examples in Figs. 2a-c and Fig. 3 illustrate why the complex network approach is able to detect heat periods without using a temperature threshold. The black curves represent (artificially generated) daily maximum temperature model data. Further we assume that one heat period has actually occurred in Figs. 2a-c persisting for 15 days from day 11 to day 25. Accordingly the black curves show different possible model results if the model exhibits predictive skill to detect a signal out of the noise.

Figure 2a depicts that using the standard approach the model correctly detects one heat period above the threshold. In Fig. 2b the model detects a signal, but this signal is too weak to cross the threshold, thus no heat period would have been detected and the model underestimates the number of heat periods. Overestimation of the number of heat periods happens in Fig. 2c, where the model detects two heat periods (5 days above the threshold at the edges and below the threshold in between). Now, the key point for our motivation is that a heat period constitutes an event in space and time, thus in a certain region, many time series would look like the ones in Fig. 2. The link strength of a network would be given by the correlation between these coherent time series. Since the signals in Figs. 2a–c look quite similar, the link strength of the network would thus be very similar in all three cases. Whereas the standard approach would correctly ~~predict~~ estimate the heat period in only one case (Fig. 2a), the networks' link strength would correctly ~~predict the heat period~~ estimate it in all three cases, given a proper relation between link strength and heat periods.

To test the relation in principle, we created 100 artificial time series (Gaussian noise) and included successively 0–9 heat periods. Figure 3a shows such a time series with three artificial heat periods indicated by the dashed lines. In a following step, we calculated the mean correlation (link strength) between these 100 coherent time series dependent on the number of included heat periods depicted in Fig. 3b. As can be seen, more heat periods are connected with a larger link strength. This simplified test shows that a proper relation between link strength and heat periods could exist. Note that Fig. 3b is not a calibration curve for real data, because we simply used Gaussian noise to create the time series.

It is clear that the argumentation above concerning the link strength as a heat period estimator is quite simplistic, but it elucidates our approach and the main idea.

4 Method

Our hypothesis is that complex network measures may be better estimators for climate extremes than standard measures like absolute threshold exceedances.

4.1 Data pre-processing

Before using the complex networks in general it is necessary to remove stationary biases and long-term variabilities from the climate time series (Donges et al., 2009).

We remove bias, trend and the average annual cycle by subtracting a standard linear regression including a Fourier series from the time series according to:

$$y_i(t) = \delta_i + \omega_i t + \sum_{j=1}^{2} \alpha_{i,j} \sin\left(\frac{2\pi j \cdot t}{365.25}\right) + \beta_{i,j} \cos\left(\frac{2\pi j \cdot t}{365.25}\right), \tag{1}$$

where $y_i(t)$ represents daily maximum temperature from 1961 to 2010, δ_i is the intercept, ω_i is the linear trend and $\alpha_{i,j}$ and $\beta_{i,j}$ represent the Fourier coefficients. Equation (1) is evaluated individually at each grid point $i = 1, \ldots, N$.

In order to minimize the influence of cold periods on the network approach (details below in Sect. 4.3), we remove the data lower than the 10 % quantile. This filtering has no influence on the standard estimator of heat periods. Then, the months from June to September are selected because we are interested in summer heat periods.

These summer anomalies are used for both the standard approach, defined in Sect. 4.2, and the new approach illustrated in Sect. 4.3. We introduce a skill measure to compare the number of heat periods with values of the link strength (Sect. 4.4). Finally we present a simple ~~calibration formula to predict heat periods with the link strength, which can be applied~~ approach to apply the new estimator to real forecasts in Sect. 4.5.

4.2 The standard approach for determining the number of heat periods

In this study, we define a heat period for E-OBS observational data as a time range when the anomaly maximum temperature (according to Eq. 1) exceeds a fixed threshold of $3\,\mathrm{K}$ at five consecutive days at least, and additionally includes not less than $20\,\%$ of the grid points in the area of interest. This choice has been made to observe events frequently enough for reliable statistics while simultaneously ensuring important impacts.

To account for the inherent model bias it is essential to adjust the temperature threshold to the model climate. Thus, we estimate the percentile $\mathcal{P}^{3\,\mathrm{K}}$ corresponding to the $3\,\mathrm{K}$ E-OBS threshold for the complete time from 1961 to 2010 and the area of interest. Accordingly, we use this percentile as the threshold for heat periods for the model data which is nevertheless fixed for the whole area and time range and the argumentation of Sect. 3 still holds for the model data. Table 1 shows this threshold in K for the eight Prudence regions, estimated from the CCLM ensemble means. In the following, we will refer to this definition as *standard approach*.

4.3 The new approach

As an alternative heat period estimator, we propose to use the time varying link strength W_τ (τ represents the years) of a network, based on modeled daily maximum temperature time series. The link strength W_τ is the correlation threshold between time series, which is needed to construct a network of a given edge density. Accordingly we want to show that W_τ has the potential to be a better estimator for observational heat periods than the standard estimator. This approach is similar to that used by Ludescher et al. (2013), who forecasted El Niño events using the link strength of a network and showed the superiority to standard sea surface temperature predictions by state-of-the-art climate models. By contrast to Ludescher et al. (2013), however, we use the predicted $2\,\mathrm{m}$ maximum temperature of CCLM to create the networks and to forecast the number of heat periods.

To apply the method we proceed as follows. Suppose we have initialized our climate model in the year 2001 with the ocean, soil, ice and atmospheric state at that time.

Accordingly the climate model runs freely for 10 years, i.e. a retrospective decadal climate prediction. Now we are interested in the capability of the model to represent heat periods in summer. Based on the standard approach of counting heat periods (see Sect. 4.2) we could determine the prediction skill of the model in forecasting (hindcasting) the number of heat periods. Our approach, in contrast, is to create a time-evolving complex network with fixed edge density (Berezin et al., 2012; Radebach et al., 2013; Ludescher et al., 2013; Hlinka et al., 2014) from the modeled daily maximum temperature time series and use, as mentioned, the dynamics of the link strength W_τ as a heat period estimator.

Following our aim to use a network measure as a heat period estimator we construct a complex network from the daily maximum temperature model data. Here we use an undirected and unweighted simple graph. Thus, the network consists of vertices V, which are the spatial grid points of our temperature data, and edges (connections) E, which are added between vertices and represent the statistical interdependence between the anomaly daily maximum temperature time series. This complex climate network can be represented by the symmetric adjacency matrix **A** with:

$$A_{ij} = \begin{cases} 0 & \text{if } ij \text{ not connected} \\ 1 & \text{if } ij \text{ connected} \end{cases}, \tag{2}$$

where i and j represent the vertices, i.e. time series at grid points $i,j = 1,\ldots,N$. Two grid points are connected if the correlation between their time series exceeds a predefined threshold. The statistical interdependence between pairs $\{ij\}$ (self-loops $\{ii\}$ are not allowed) of time series is measured using the Pearson (standard) correlation coefficient (Donges et al., 2009). From sensitivity studies we found that correlations between time series in the order of 0.7–0.9 yield patterns with not too few and not too many connections. This is important in order to resolve temporal dynamics of the network. Correlations in this order of magnitude are significant on the 5 % level for the here used summer time series with length of about 120 days. However, since we want to analyze different regions in Europe and to generate comparable results we decided to alternatively create our networks with a constant edge density (ratio of number of actual connections to maximum number of connections) of

$$\rho = E \Big/ \begin{pmatrix} N \\ 2 \end{pmatrix} = \langle k_i \rangle / (N-1) = 0.3\,, \tag{3}$$

where E is the number of edges and $\langle k_i \rangle$ is the mean *node degree* with

$$k_i = \sum_{j=1}^{N} A_{ij}\,, \tag{4}$$

which gives the number of connections of a vertex i.

As mentioned above we removed the data lower than the $10\,\%$ quantile, to avoid that the link strength W_τ is influenced by possible cold periods in the data. We tested smaller quantiles ($5\,\%$) and larger quantiles ($20\,\%$) and found that the results are robust, i.e. they changed only slightly. The above used parameters (like the density of 0.3) and the $10\,\%$ filtering turned out to be optimal for our data. For other data, these parameters most probably have to be adjusted. Additionally, by removing the data lower than the $10\,\%$ quantile, gaps in the time series are generated. To ensure significance, we take into account only correlation coefficients where the two underlying time series exhibit 60 common data points (days). An effective way to estimate the link strength of a network with an edge density of 0.3 is to calculate the $70\,\%$ quantile of all correlation coefficients involved in the network.

In a similar way as Berezin et al. (2012) we analyze the temporal variation of the link strength W_τ, i.e the correlation threshold between time series (grid points) for a single year τ (summer) from 1961 to 2010. Thus, instead of using the *node degree* as an estimator of heat periods we use the link strength W_τ.

Using the definitions above, we finally construct a network for the summer months of each year based on anomaly maximum temperature model data. The quantity whose year-to-year variation we are interested in is the link strength W_τ; however, since we are interested in *decadal* variability, and since we do not expect the model to represent the year to year fluctuations, we applied a 10 year moving average filter to both link strength and number of heat periods, subsequently. Since the CCLM model has been initialized every decade (1961, 1971,..., 2001) we apply the filter only within a decade in order to avoid transferring information between decades. At the boundaries of the decades, the time range for the running average is shortened: For instance at the beginning of the decade, we use only the six years' mean (e.g. from 2001 to 2005), in the second year seven, and so on.

4.4 Comparison of the different quantities

To quantify the prediction skill of the model, we calculate the absolute mean difference (see Eqs. 6 and 7) between the number of heat periods in E-OBS (o) and CCLM (m) and the CCLM link strength (W_τ). To be comparable we normalized the time series to the range $\{0,1\}$ by a subtraction of the minimum of the time series and accordingly a division by the maximum for the whole time span, e.g. for the number of heat periods in CCLM:

$$\mu^r_{d,\tau} = \left(m^r_{d,\tau} - \min_{\tau=1}^{50} \left(m^r_{d,\tau} \right) \right) \Big/ \max_{\tau=1}^{50} \left(m^r_{d,\tau} - \min_{\tau=1}^{50} \left(m^r_{d,\tau} \right) \right) , \tag{5}$$

where r denotes the European region, d stands for the decade and τ represents the years. The similarly rescaled E-OBS number of heat periods will be denoted as Ω and the rescaled CCLM link strength as ψ.

Thus the absolute mean difference (based on normalized data) between observation and model heat periods for a region r and a decade d is given by

$$M^r_d(\mu) = \left| \frac{1}{10} \sum_{\tau=1}^{10} \left(\Omega^r_{d,\tau} - \mu^r_{d,\tau} \right) \right| = |\overline{\Omega}^r_d - \overline{\mu}^r_d| \,, \tag{6}$$

and the mean difference between observation heat periods and model link strength is

$$M^r_d(\psi) = \left| \frac{1}{10} \sum_{\tau=1}^{10} \left(\Omega^r_{d,\tau} - \psi^r_{d,\tau} \right) \right| = |\overline{\Omega}^r_d - \overline{\psi}^r_d| \,, \tag{7}$$

where the bars in the above equations denote temporal averages. Therefore, if the absolute mean difference is about 0, observations and model agree well, whereas a difference of about 1 denotes the maximum discrepancy.

4.5 ~~Prediction~~ Usage of ~~heat periods~~the new estimator in predictions

For a real application of our method to ~~predict~~ estimate the number of heat periods in ~~a forecasting sense~~forecasts, a calibration step using observational data o is needed to convert the link strength of the model to the number of heat periods m_y (the index y stands for year in the future). Therefore long hindcast data are needed. Based on our analysis we suggest as a first attempt to apply a linear conversion from link strength W_y to the number of heat periods m_y, which is also supported by our tests shown in Fig. 3:

$$m^r_{y,W} = \frac{W^r_y - \min_{\tau=1}^{50}(W^r_\tau)}{\max_{\tau=1}^{50}(W^r_\tau) - \min_{\tau=1}^{50}(W^r_\tau)} \cdot \left(\max_{\tau=1}^{50}(o^r_\tau) - \min_{\tau=1}^{50}(o^r_\tau) \right) + \min_{\tau=1}^{50}(o^r_\tau) \tag{8}$$

This linear approach corresponds to our skill analysis, where a linear connection between the link strength and the number of heat periods is assumed as well. Again we note that this study presents only the skill analysis of hindcast data and Eq. 8 is actually not used now.

5 Results

Figure 4 depicts the number of observed heat periods (solid line) and the corresponding link strength (dashed line) retrieved from the complex evolving network, both from E-OBS data for France (Prudence region 3), and shows that the link strength W_τ is a suitable estimator of heat periods. It shows that the network contains climate information in the sense that the dynamics of the link strength W_τ is similar to the dynamics of heat periods, both based on the same data. So, the link strength can here be considered as an estimator for heat periods which is comparable to the standard heat period estimator. Jumps between the decades occur as the running mean filter is only applied within the decades (see Sect. 4.3). The corresponding figures for the seven other Prudence regions can be found in the supplementary material.

As an example, Prudence region 8 (Eastern Europe) is a region where the network method performs better than the standard approach (Figs. 5 and 6). Figure 5 shows the E-OBS number of heat periods o (black) and the CCLM ensemble mean number of heat periods m (blue) for Eastern Europe together with the interquartile range (25th and 75th percentiles), and Fig. 6 shows again the E-OBS number of heat periods now compared to the CCLM link strength. Comparing the absolute mean differences, denoted as M in the two figures reveals that our network approach enhances the skill in four decades, namely 1970, 1980, 1990, 2000. Especially the 1970s, 1980s, 1990s show a clear improvement and our network approach better reflects the low frequency dynamics of the heat periods. The 2000s seem to be off in both model cases, the number of heat periods and the link strength, which indicates a failed model initialization.

In order to see how the prediction skill of the standard as well as the network heat period estimators vary with the considered region, we performed the same analysis as above for the eight Prudence regions in Europe and for the 1960s, 1970s, 1980s, 1990s and 2000s. The corresponding figures for the other regions can be found in the supplement. ~~To summarize our~~

To summarize the results we calculated ~~as the prediction skill~~the absolute mean differences (Eqs. 6 and 7) for all the Prudence regions, see Fig. 7. Blue colors in the panels stand for low values (high skill) whereas red colors depict high values (low skill) in the absolute mean difference~~within a decade between~~ .

The left panel of Fig. 7 shows how the network method performs using only E-OBS ~~heat periods and CCLM heat periods (~~data similar to Fig. 4. Therefore we estimated the number of heat periods in the E-OBS data and the link strength (from the complex network) of the E-OBS data and accordingly calculated the differences (after normalization, cf. Eq. ~~6)~~ 5) between these two estimators. As can be seen blue colors dominate the plot, i.e. low differences and hence high skill. Thus, this reference test shows that the link strength is coupled to the number of heat periods in maximum daily summer temperature data and so can be used as an alternative, possibly better, heat periods estimator. There are some exceptions like the 1990s and 2000s of Prudence region 6. Further investigation on the reasons of these cases has to be performed.

The middle panel of Fig. 7 shows how well the standard method performs in predicting heat periods using E-OBS ~~heat periods~~ observations and CCLM model data. The right panel indicates the performance of the new network method in estimating heat periods using E-OBS and CCLM data. In contrast to the relatively low values in the left panel, the values in the middle and right panels on the one hand are higher for many decades and Prudence regions. On the other hand, the visual impression is that the absolute differences of the right panel are slightly smaller than those of the middle panel, especially during the 1990s and ~~CCLM link strength (Eq. 7). Figure 8~~2000s.

Thus, we can conclude with Fig. 7 that the method works in principle but that the uncertainties in the model simulations lead to increased differences between observations and model simulations. In addition, the link strength seems to work better than the standard approach for the model simulations with respect to the observations.

To quantify this last statement, we calculated the difference between the middle and right panels of Fig. 7, see Fig. 8. This basically shows which method performs better regarding the eight regions (columns) and five decades (rows). Blue color in Fig. 8 indicates that the network approach performs better ($M_d^r(\psi) < M_d^r(\mu)$) and red color stands for a better performance of the standard approach ($M_d^r(\psi) > M_d^r(\mu)$). White boxes in Fig. 8 denote a tie between the methods in the case of too small differences ($|M_d^r(\psi) - M_d^r(\mu)| \leq 0.05$). The matrix of Fig. 8 shows that the network method is clearly superior in three regions (5,7,8) and slightly superior in two regions (4,6), the standard approach is superior in two regions (1 and 3) and in region 2 we observed a tie, i.e. no clear result.

The crucial question is if this result indicates that the network method performs significantly better than the standard approach or not. However, testing for statistical significance bears serious problems. There are so many factors involved in the analysis, i.e. the models themselves, the downscaling, the ensemble, the initialization, the different regions, the filtering, etc. that any ~~nullhypothesis~~ null hypothesis would be not well-posed and any test would be questionable. This issue is discussed in detail in 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 nullhypothesis 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.” and finally “... propose to employ instead a simple descriptive approach for characterising the information in an ensemble ...”. Although we totally agree with the argumentation by (von Storch and Zwiers, 2013) that a “classical” significance test would most probably fail in our analysis, we think that alternative significance tests, based on bootstrapping or surrogate data, could definitely help for a better interpretation of the results. Thus, we construct the following significance test based on surrogate data to answer the question: “What is the probability of getting a rank matrix like the one in Fig. 8 by chance?”.

First, we have to define what is the possibly “significant” characteristic of the matrix in Fig. 8. It is, as we concluded above, that the network method is superior in five regions. Thus the question is: “What is the probability to observe at least five regions, where we have in each at least one blue matrix element more than a red one by chance?”. Accordingly we constructed matrices like in Fig. 8 by randomly coloring 20 matrix elements blue and 20 red. Afterwards we colored 8 matrix elements white as in Fig. 8. Finally, we repeated this surrogate procedure 1000 times and counted the cases (regions) where the blue matrix elements dominate. Table 2 shows the probabilities that blue matrix elements dominate in n regions. Since we have 16 blue elements and 16 red, it is sure that blue dominates in $n = 1$ region and impossible to dominate in $n = 7$ and $n = 8$ regions. As can be seen from Tab. 2 the probability of dominating in $n = 5$ regions by chance is only about $5\,\%$, thus the results of our network approach have to be stated significant. Due to the symmetry of the test, the same argumentation is valid for red matrix elements. Dominating in $n = 2$ regions, as achieved by the standard approach (Fig. 8), can be realized easily by chance with a probability of approx. $99\,\%$. Page 1 in the supplementary material shows an example of 12 of these randomly generated matrices, where one matrix, depicted by a black frame, fulfills the "significance" criterion.

6 Conclusions and outlook

We presented a novel approach examining heat periods using a complex network analysis. We have investigated the predictability of the slow dynamics of the occurrence of heat periods in Europe based on daily maximum near-surface temperature data.

We found that the network approach is superior (significance is $\approx 5\,\%$) to the standard approach in ~~predicting~~ estimating heat periods in Europe, hence highlighting the potential of network methods to improve the skill estimation in decadal prediction experiments. Picking up our hypothesis and simplified argumentation from Sect. 3, the crucial point why we detect heat periods with the network link strength is that heat periods are cooperative events in space and time. Thus, the link strength can be used as an estimator of heat periods. The drawback of the standard approach is most probably the inflexible threshold for the detection of heat periods (cf. Fig. 2). If the climate model contains the signal of a heat period, but with a slightly too small amplitude, the threshold will not be crossed and no heat period will be detected. In contrast, the complex climate network does not depend on such fixed thresholds, and can use this information, which makes it the more robust estimator of heat periods.

The general prediction skill of climate in Europe using standard measures is still moderate. In this sense our work adds new aspects to our previous study (Mieruch et al., 2014) and also the work of Eade et al. (2012) who found a strong variation of skill with region and decade. In essence, we found regions and decades in Europe where our climate model output, or more specifically the used network estimator, follows the slowly evolving dynamics of observed heat periods. We also found regions and decades, where the network estimator is not able to represent the observational reference. Understanding of this variability in prediction skill is one of the future challenges of decadal predictions.

Concluding, our approach shows that the complex climate networks approach yields meaningful climate information and has the potential to improve skill measures within the framework of climate prediction. 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) yields 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 analyzing predictions of natural variability years ahead than using methods from short- or medium range forecasting. Further, from the network perspective it would be interesting to analyze other network measures like clustering, similarities or path lengths and how they are connected to climate evolution. The incorporation of other relevant variables like precipitation, wind or soil moisture into the network is an appealing aspect. From a physical or climatological point of view it is important to understand why the network measures are able to represent climate dynamics, which could also contribute to a better understanding of the sources of decadal predictability. Thus, the incorporation and investigation of processes like the AMOC, PDO or NAO together with complex networks and climate prediction might be an option for the future.

Acknowledgements. We acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://www.ecad.eu).

The research program MiKlip is funded by the German ministry of education and research (BMBF).

We also acknowledge the ETOPO1 data set (Amante and Eakins, 2009) from the NGDC (National Geophysical Data Center, Boulder, Colorado).

The article processing charges for this open-access publication were covered
by a Research Centre of the Helmholtz Association.

References

Amante, C. and Eakins, B.: ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum, Tech. Rep. NESDIS NGDC-24, National Geophysical Data Center, NOAA, Boulder, Colorado, USA, doi:10.7289/V5C8276M, 2009.

Berezin, Y., Gozolchiani, A., Guez, O., and Havlin, S.: Stability of climate networks with time, Sci. Rep., 2, 666, doi:10.1038/srep00666, 2012.

Boers, N., Bookhagen, B., Barbosa, H. M. J., Marwan, N., Kurths, J., and Marengo, J. A.: Prediction of extreme floods in the eastern Central Andes based on a complex networks approach, Nat. Commun., 5, 5199, doi:10.1038/ncomms6199, 2014.

Chikamoto, Y., Timmermann, A., Luo, J.-J., Mochizuki, T., Kimoto, M., Watanabe, M., Ishii, M., Xie, S.-P., and Jin, F.-F.: Skilful multi-year predictions of tropical trans-basin climate variability, Nat. Commun., 6, 6869, doi:10.1038/ncomms7869, 2015.

Corti, S., Weisheimer, A., Palmer, T. N., Doblas-Reyes, F. J., and Magnusson, L.: Reliability of decadal predictions, Geophys. Res. Lett., 39, 21, doi:10.1029/2012GL053354, 2012.

Doblas-Reyes, F. J., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., Mochizuki, T., Rodrigues, L. R. L., and van Oldenborgh, G. J.: Initialized near-term regional climate change prediction, Nat. Commun., 4, 1715, doi:10.1038/ncomms2704, 2013.

Doms, G. and Schättler, U.: A Description of the Non-Hydrostatic Regional Model LM, Part I: Dynamics and Numerics, Tech. rep., Deutscher Wetterdienst, P. O. Box 100465, 63004 Offenbach, Germany, LM_F90 2.18, 2002.

Donges, J. F., Zou, Y., Marwan, N., and Kurths, J.: Complex networks in climatic dynamics, The European Physical Journal Special Topics, 174, 157–179, doi:10.1140/epjst/e2009-01098-2, 2009.

Donges, J. F., Donner, R. V., Trauth, M. H., Marwan, N., Schellnhuber, H.-J., and Kurths, J.: Nonlinear detection of paleoclimate-variability transitions possibly related to human evolution, P. Natl. Acad. Sci. USA, 108, 20422–20427, 2011.

Dosio, A., Panitz, H.-J., Schubert-Frisius, M., and Lüthi, D.: Dynamical downscaling of CMIP5 global circulation models over CORDEX-Africa with COSMO-CLM: evaluation over the present climate and analysis of the added value, Clim. Dynam., 44, 2637–2661, doi:10.1007/s00382-014-2262-x, 2015.

Eade, R., Hamilton, E., Smith, D. M., Graham, R. J., and Scaife, A. A.: Forecasting the number of extreme daily events out to a decade ahead, J. Geophys. Res.-Atmos., Res., 117, D21110, doi:10.1029/2012JD018015, 2012.

Frich, P., Alexander, L., Della-Marta, P., Gleason, B., Haylock, M., Klein Tank, A., and Peterson, T.: Observed coherent changes in climatic extremes during the second half of the twentieth century, Clim. Res., 19, 193–212, 2002.

García-Serrano, J., Doblas-Reyes, F. J., Haarsma, R. J., and Polo, I.: Decadal prediction of the dominant West African monsoon rainfall modes, J. Geophys. Res.-Atmos., 118, 5260–5279, 2013.

Giorgi, F., Jones, C., and Asrar, G.: Addressing climate information needs at the regional level: the CORDEX framework, Bulletin of the World Meteorologic Organization, 58, 175–183, 2009.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, J. Geophys. Res., 113, D20119, doi:10.1029/2008JD010201, 2008.

Hlinka, J., Hartman, D., Jajcay, N., Vejmelka, M., Donner, R., Marwan, N., Kurths, J., and Paluš, M.: Regional and inter-regional effects in evolving climate networks, Nonlin. Processes Geophys., 21, 451–462, doi:10.5194/npg-21-451-2014, 2014.

Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O. B., Bouwer, L., Braun, A., Colette, A., Déqué, M., Georgievski, G., Georgopoulou, E., Gobiet, A., Menut, L., Nikulin, G., Haensler, A., Hempelmann, N., Jones, C., Keuler, K., Kovats, S., Kröner, N., Kotlarski, S., Kriegsmann, A., Martin, E., Meijgaard, E., Moseley, C., Pfeifer, S., Preuschmann, S., Radermacher, C., Radtke, K., Rechid, D., Rounsevell, M., Samuelsson, P., Somot, S., Soussana, J.-F., Teichmann, C., Valentini, R., Vautard, R., Weber, B., and Yiou, P.: EURO-CORDEX: new high-resolution climate change projections for European impact research, Reg. Environ. Change, 14, 1–16, 2013.

Kadow, C., S. Illing, O. Kunst, H. W. Rust, H. Pohlmann, W. A. Müller, and U. Cubasch (2015), Evaluation of forecasts by accuracy and spread in the miklip decadal climate prediction system, *Meteorologische Zeitschrift*, pp. –.

Keenlyside, N. S., Latif, M., Jungclaus, J., Kornblueh, L., and Roeckner, E.: Advancing decadal-scale climate prediction in the North Atlantic sector, Nature, 453, 84–88, 2008.

Kothe, S., Panitz, H.-J., and Ahrens, B.: Analysis of the radiation budget in regional climate simulations with COSMO-CLM for Africa, Meteorol. Z., 23, 123–141, doi:10.1127/0941-2948/2014/0527, 2014.

Ludescher, J., Gozolchiani, A., Bogachev, M. I., Bunde, A., Havlin, S., and Schellnhuber, H. J.: Improved El Niño forecasting by cooperativity detection, P. Natl. Acad. Sci. USA, 110, 11742–11745, doi:10.1073/pnas.1309353110, 2013.

MacLeod, D. A., Caminade, C., and Morse, A. P.: Useful decadal climate prediction at regional scales?, a look at the ENSEMBLES stream 2 decadal hindcasts, Environ. Res. Lett., 7, 044012, doi:10.1088/1748-9326/7/4/044012, 2012.

Matei, D., Pohlmann, H., Jungclaus, J., Müller, W., Haak, H., and Marotzke, J.: Two tales of initializing decadal climate prediction experiments with the echam5/mpi-om model, J. Climate, 25, 8502–8523, 2012.

Meehl, G. A., Teng, H., and Arblaster, J. M.: Climate model simulations of the observed early-2000s hiatus of global warming, Nature Climate Change, 4, 898–902, doi:10.1038/nclimate2357, 2014.

Mieruch, S., Feldmann, H., Schädler, G., Lenz, C.-J., Kothe, S., and Kottmeier, C.: The regional MiKlip decadal forecast ensemble for Europe: the added value of downscaling, Geosci. Model Dev., 7, 2983–2999, doi:10.5194/gmd-7-2983-2014, 2014.

Müller, W. A., Baehr, J., Haak, H., Jungclaus, J. H., Kröger, J., Matei, D., Notz, D., Pohlmann, H., von Storch, J. S., and Marotzke, J.: Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology, Geophys. Res. Lett., 39, L22707, doi:10.1029/2012GL053326, 2012.

Peron, T. K. D., Comin, C. H., Amancio, D. R., da F. Costa, L., Rodrigues, F. A., and Kurths, J.: Correlations between climate network and relief data, Nonlin. Processes Geophys., 21, 1127–1132, doi:10.5194/npg-21-1127-2014, 2014.

Radebach, A., Donner, R. V., Runge, J., Donges, J. F., and Kurths, J.: Disentangling different types of El Niño episodes by evolving climate network analysis, Phys. Rev. E, 88, 052807, doi:10.1103/PhysRevE.88.052807, 2013.

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., Eade, R., Fereday, D., Folland, C. K., Gordon, M., Hermanson, L., Knight, J. R., Lea, D. J., MacLachlan, C., Maidens, A., Martin, M., Peterson, A. K., Smith, D., Vellinga, M., Wallace, E., Waters, J., and Williams, A.: Skillful long-range prediction of European and North American winters, Geophys. Res. Lett., 41, 2514–2519, 2014.

Smith, D., Eade, R., and Pohlmann, H.: A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction, Clim. Dynam., 41, 3325–3338, doi:10.1007/s00382-013-1683-2, 2013.

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K., Brokopf, R., Fast, I., Kinne, S., Kornblueh, L., Lohmann, U., Pincus, R., Reichler, T., and Roeckner, E.: Atmospheric component of the MPI-M Earth System Model: ECHAM6, Adv. Model. Earth Syst., 5, 146–172, doi:10.1002/jame.20015, 2013.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, B. Am. Meteorol. Soc., 93, 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012.

Tsonis, A. A. and Swanson, K. L.: Review article “On the origins of decadal climate variability: a network perspective”, Nonlin. Processes Geophys., 19, 559–568, doi:10.5194/npg-19-559-2012, 2012.

Tsonis, A. A., Swanson, K., and Kravtsov, S.: A new dynamical mechanism for major climate shifts, Geophys. Res. Lett., 34, L13705, doi:10.1029/2007GL030288, 2007.

van Oldenborgh, G. J., Doblas-Reyes, F. J., Wouters, B., and Hazeleger, W.: Skill in the trend and internal variability in a multi-model decadal prediction ensemble, Clim. Dynam., 38, 1263–1280, 2012.

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

Zou, Y., Donner, R. V., Marwan, N., Small, M., and Kurths, J.: Long-term changes in the north–south asymmetry of solar activity: a nonlinear dynamics characterization using visibility graphs, Nonlin. Processes Geophys., 21, 1113–1126, doi:10.5194/npg-21-1113-2014, 2014.

Table 1. Ensemble mean variation of the temperature threshold calculated for heat periods with the standard approach in CCLM data (see Sect. 4.2).

Prudence region	1	2	3	4	5	6	7	8
Temperature threshold (in K)	3.16	3.38	2.81	2.52	2.66	2.85	3.46	2.79

Table 2. Probability that blue matrix elements in Fig. 8 dominate in n regions by chance. Half of the elements are colored blue and red, respectively, and eight white elements are randomly added subsequently.

Number of regions n	1	2	3	4	5	6	7	8
Probability in %	100	99	82	35	5	0.1	0	0

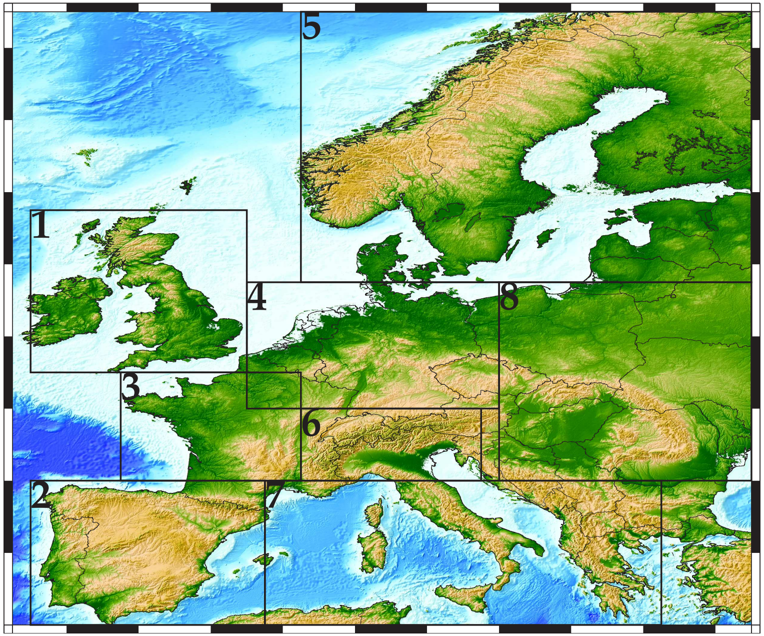


Figure 1. The 8 Prudence regions. (Topography: ETOPO1, Amante and Eakins, 2009.)

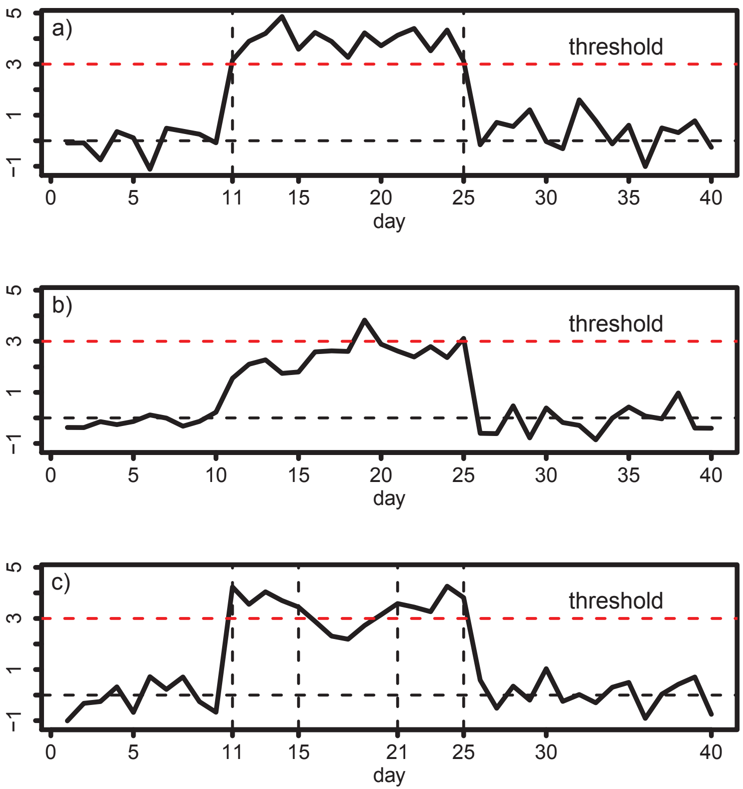


Figure 2. Schematical illustration of our approach (temperature anomaly on the y axis): **(a)** model detects correctly one heat period above the threshold, **(b)** model underestimates the number of heat periods, **(c)** model overestimates the number of heat periods (details see text).

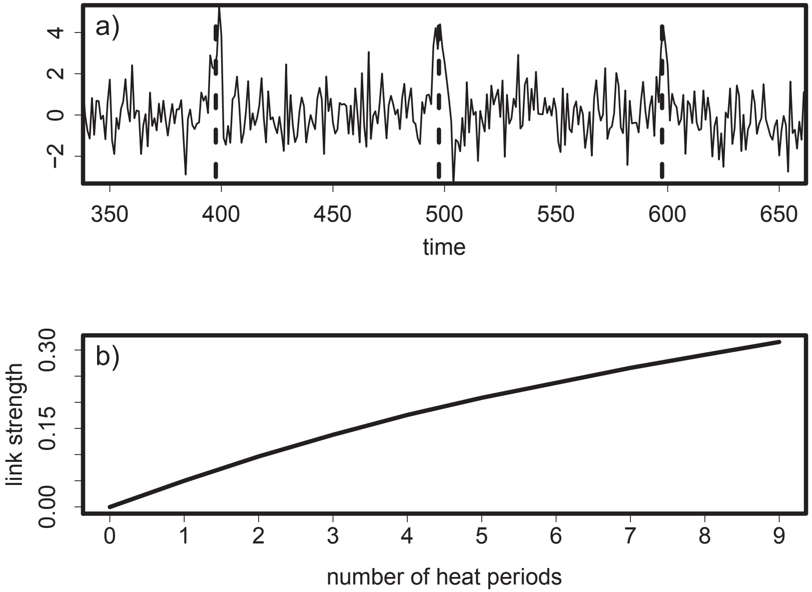


Figure 3. (a) Artificial time series including 3 heat periods (dashed lines). **(b)** Relation between the network link strength and the number of heat periods, based on 100 artificial time series.

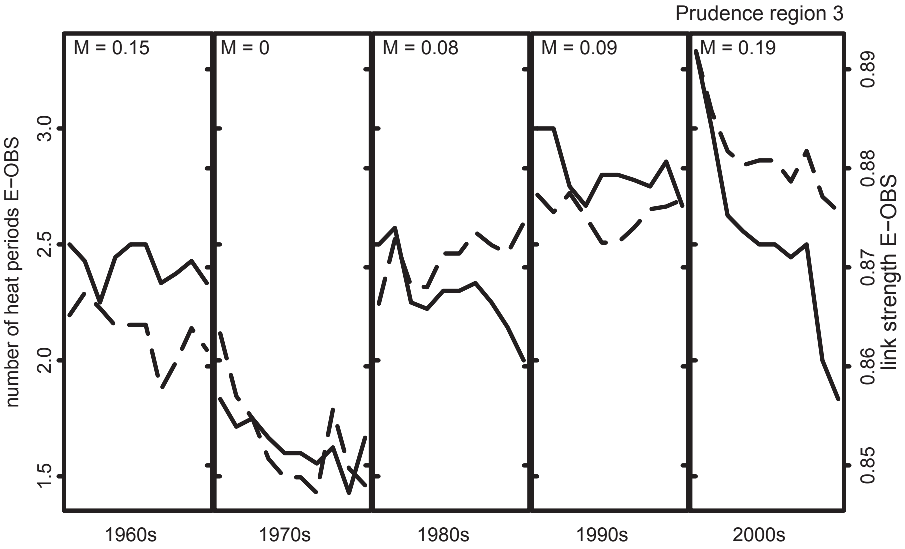


Figure 4. Number of heat periods (1961–2010) in France (Prudence 3) in summer from E-OBS o (solid line) and corresponding E-OBS link strength W (dashed line). The "M's" denote the absolute mean difference within a decade between E-OBS standard approach and the ~~CCLM ensemble mean~~ E-OBS link strength after normalization, ~~see~~ cf. Eq. 7.

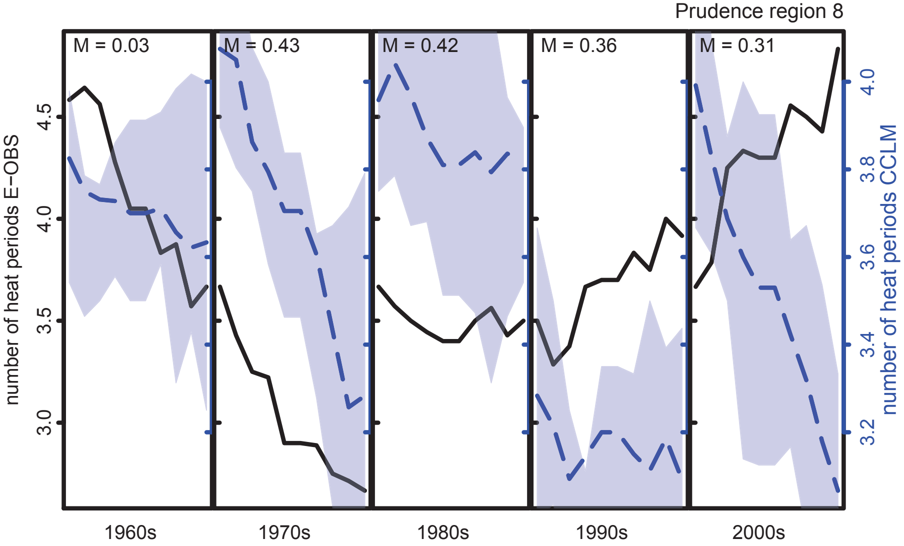


Figure 5. Number of heat periods (1961-2010) in Eastern Europe (Prudence 8) in summer from E-OBS o (black) and CCLM number of heat periods (blue: ensemble mean and interquartile range). The "M's" denote the absolute mean difference within a decade between E-OBS and the CCLM ensemble mean after normalization, see Eq. 6.

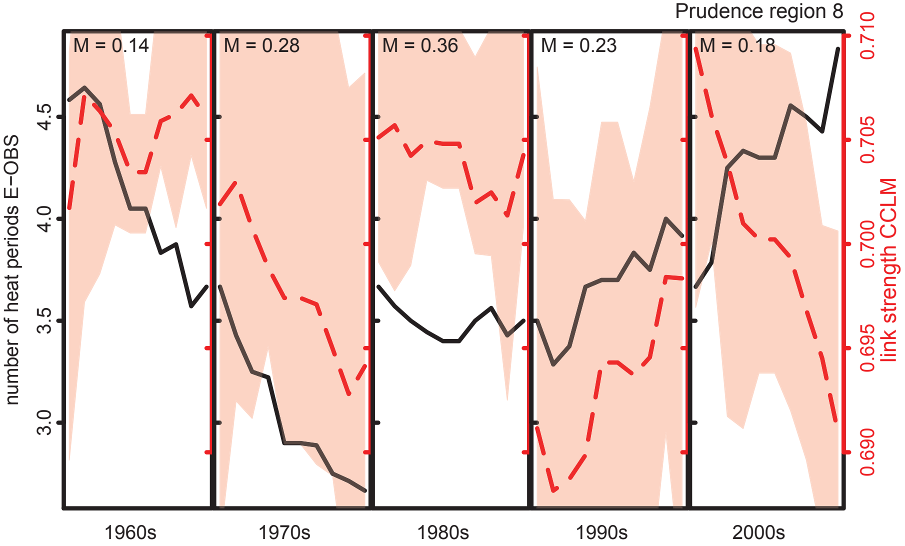


Figure 6. Number of heat periods (1961-2010) in Eastern Europe (Prudence 8) in summer from E-OBS o (black) and CCLM link strength or correlation threshold W (red: ensemble mean and interquartile range). The “M’s” denote the absolute mean difference within a decade between E-OBS and the CCLM ensemble mean after normalization, see Eq. 7.

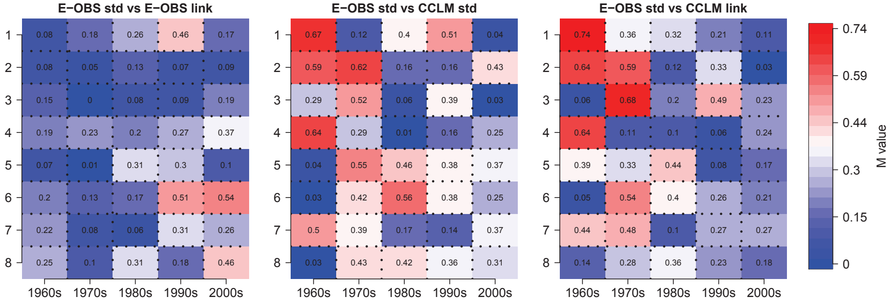
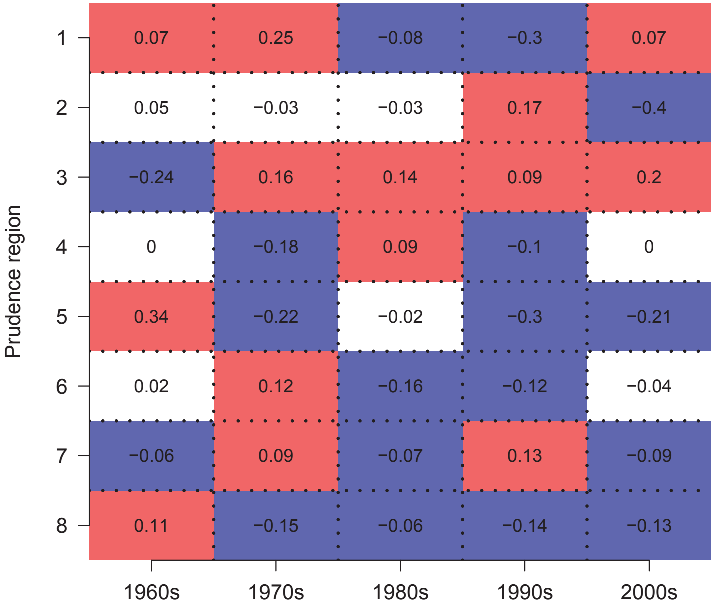


Figure 7. ~~Rank matrix of the performance of the two methods~~Absolute mean differences like in Figs. ~~Blue: network~~ 4 to 6 of all Prudence regions for E-OBS standard approach ~~performs better~~compared to E-OBS link strength (left), ~~red:~~ to CCLM standard approach ~~performs better, white: tie~~(middle) and to CCLM link strength (right).



	1960s	1970s	1980s	1990s	2000s
1	0.07	0.25	−0.08	−0.3	0.07
2	0.05	−0.03	−0.03	0.17	−0.4
3	−0.24	0.16	0.14	0.09	0.2
4	0	−0.18	0.09	−0.1	0
5	0.34	−0.22	−0.02	−0.3	−0.21
6	0.02	0.12	−0.16	−0.12	−0.04
7	−0.06	0.09	−0.07	0.13	−0.09
8	0.11	−0.15	−0.06	−0.14	−0.13

Figure 8. Differences between the two right panels of Fig. 7. Blue: network approach performs better, red: standard approach performs better, white: tie.