1. Introduction

Climate indices are widely used across a number of disciplines. They have become an important impact parameter in climate change studies [Christenson et al., 2006; Zubler et al., 2014], especially when considering extremes [Alexander et al., 2006; Fischer et al., 2013; Sillmann et al., 2013a, 2013b]. Indices offer the possibility to get a more direct quantification of what implications for the environment or certain economic sectors can be expected [Della-Marta et al., 2009]. Many of these indices are based on daily values such as daily maximum temperature or different daily percentiles. For example, wet days are defined as the number of days with a precipitation amount larger than the 90th percentile of the long-term precipitation climatology of this calendar day [Klein Tank et al., 2009]. Hence, for each calendar day the percentile-based thresholds need to be estimated to count the number of events. In order to evaluate the changes of such indices over time due to climate forcings, the number of events in the reference period is compared to the numbers of some future period [Zubler et al., 2014]. A nice example of an index that gives the user information about its own comfort today and compares it to the comfort in a future climate is the following: The heat stress index combines temperature and precipitation information to draw conclusions about human comfort in a warmer climate, which is more informative than a simple temperature increase [Fischer et al., 2012]. Furthermore, depending on the index the seasonal forecast of it can be more skillful than simple seasonal means [Brands, 2013].

Often, 30 years (as reference period or hindcast length) or less are available to compute such indices. When calculating such indices based on daily data of about 30 years (e.g., the 90th percentile of precipitation of one particular calendar day), the sample size consists of about 30 values for the observations and the individual climate model (only in case of seasonal forecasting with an ensemble hindcast data set the sample size is larger). This sample size is often too small for a robust estimation of thresholds. Therefore, the Expert Team on Climate Change Detection and Indices (ETCCDI) [Klein Tank et al., 2009] suggests using a 5 day moving window centered on the day of interest. However, depending on the region of interest, the daily variability of the parameter of interest (e.g., temperature or precipitation) may be large, which would infer that the sample size is still too small to get a good estimate of the thresholds which define the index. Furthermore, daily values are likely to be serially correlated, reducing the effective sample size even further. Zhang et al. [2005] discuss this issue focusing on inhomogeneities due to sampling problems at the border of the reference time periods. However, when working with seasonal forecasts, the time window considered is
limited to the available data; hence, no borders exist, as the reference period is also the period of interest. Plus, depending on the forecast model, only a few months and not the whole year is being forecasted. This means that one is forced to determine the thresholds within the same time window as, for example, the verification or debiasing is being done. Thus, the data used are always within the reference period. In this study it is shown that a fitting approach can improve the threshold estimation.

When working with long-range forecasts, the bias removal is an important topic [Doblas-Reyes et al., 2005; Hanlon et al., 2012; Gangsto et al., 2013; Hawkins et al., 2013, 2014a], especially when a daily bias needs to be removed in order to calculate, for instance, indices based on absolute thresholds (e.g., heating degree days) [Piani et al., 2010]. In order to remove the bias, the “true” observed daily climate needs to be known for proper estimation of the bias. Again, based on a hindcast period of 30 years or less, the estimate of the mean observed climate is not robust and strongly influenced by day-to-day or weekly variability. Fitting the averaged mean climate and therefore smoothing the short-term variability results in more robust estimates of the climate which can then be used to calculate the bias. Often, various methods are applied to long-range forecasts in order to improve the forecast skill. However, when applying a quantile mapping or a similar technique [e.g., Themell et al., 2012; Lafon et al., 2013] or a very sophisticated method as described in Piani et al. [2010], all these methods assume that the observed climate is known. However, little information is available on how the observed climate can be estimated properly. Also, when debiasing before calculating an index based on a fixed threshold (e.g., tropical nights) in case of long-range forecasts, it is important that a robust bias estimation is established first to ensure that only the bias is removed and not the variability.

This study is structured as following: Section 2 describes the data and the method, section 3 compares the newly introduced method with the standard method within the perfect model approach; in section 4 the method is analyzed when using reanalysis data, and section 5 summarizes the conclusions drawn from this study.

2. Data and Methods

Based on an observational data set or a single model run, the only dimension available to determine a percentile-based threshold or a mean is the time dimension. As introduced above, this sample is often too small. Hence, the neighboring points in time (5 days in total) are also included in order to improve the data sample. In order to assess the difference of the estimated threshold between the 5 day moving window (hereafter 5d-fit) as suggested by ETCCDI and our approach (see below), a perfect model approach (PMA) [Müller et al., 2005; Collins et al., 2006] is used to determine the “true” climate of a model. The PMA offers another dimension to sample the data; besides the time dimension, the ensemble member dimension is added. This provides a large increase in the sample size, which allows for a good estimate of the true climate. European Centre for Medium-Range Weather Forecasts (ECMWF)'s System4 seasonal forecasting model offers the possibility to explore the true model climate as the combination of 51 members and a total of 32 hindcast years (1981–2012) results in a sample size big enough to get a good estimate of the daily mean and distribution which is needed for climate index calculations. The number of ensemble members needed to estimate the true climate varies with parameter and region. The noisier (in this case, the larger the daily variability) the climate of the parameter and the region, the more members are needed. Hence, the number of ensemble members needed scales with the signal-to-noise ratio of the parameter and the region [Deser et al., 2012]. However, with a total number of 1632 years System4 provides a sample large enough for all cases. Note that this approach is applicable because of the relatively low skill over most areas over land, hence our region of interest. If the model had perfect skill, the sample size would reduce to the observations as all members would forecast the same [Della-Marta et al., 2009].

ECMWF System4 is a fully coupled atmosphere-ocean forecast system that provides operational seasonal predictions as well as reforecasts in order to evaluate the predictions. Details on the ECMWF System4 are described in Molteni et al. [2011] and http://www.ecmwf.int/products/forecasts/seasonal/documentation/system4/. The reforecasts used in this study in case of the perfect model approach are initialized on 1 November, include 51 members, and consist of 7 months simulations. The example illustrated in section 4 uses the 1 May initialization, also consisting of 51 members and 7 months simulations.

Following the idea of the perfect model approach (PMA), one of the 51 members was used as “observations” (hereafter observed climate), meaning that the estimate of climate is based on 32 years of data only from that one member (called member x). These were then compared to the full hindcast set which is based on 51
members each with 32 years of data (hereafter true climate). Although there may be some random aspects when working with one member, sensitivity tests with other members showed that the general pattern remains very similar. Furthermore, when studying one member only, the findings are, in theory, comparable to observations, which offer only one realization and therefore smoothing effects due to averaging might be misleading.

Climate indices are often defined based on thresholds, e.g., a wet day is defined as the daily precipitation amount larger than the 90th percentile of the long-term precipitation distribution of that day. Hence, to determine whether a day is a wet day or not, the threshold of the 90th percentile needs to be defined. As shown in this study, differences exist for thresholds depending on how these are determined. The reason for these differences is the small sample size that is often available when such thresholds are estimated. In order to estimate a certain percentile of a distribution, the sample needs to be large enough for a robust result. Two different approaches are presented in this study and the differences illustrated. Based on the PMA, it is assumed that the threshold derived using the full hindcast data set is the true threshold. The same actually applies for a simple measure such as the mean or the standard deviation. In our approach, the mean, the standard deviation, and the thresholds are first derived based on daily data. For example, for each of the simulated days in the hindcast period of member \( x \), the first estimate (hereafter raw estimate) of the mean is derived by averaging all years of the reference period (32 years) from the hindcast as would be done to determine the mean observed climate; in case of the thresholds and the standard deviation, the raw estimate is derived by using the distribution of all 32 values of the hindcast period (or reference period). Then, in a next step, a statistical fit, the local polynomial regression fitting (LOESS) which uses neighboring points of the time series \( x \) ([Cleveland and Delvin, 1988]) is used to estimate the true observed mean, standard deviation, and thresholds by smoothing the raw estimate. In order to obtain an optimal fit (hereafter LOESS-fit), it is important that the fitting method takes into account the annual cycle of the climate and smoothes the short-term variability. Hence, basically, a low pass filter needs to be applied to conserve the characteristics of the climate and at the same time reduce the importance of the short-term influences. Sensitivity tests showed that there is no single best fit for all purposes. Depending on the parameter and the percentile of interest, the fit needs to be optimized. The goal is to keep as much as possible from the characteristics of the annual cycle but to smooth as much as possible from the short-term variability. This means, the closer to the tail the percentage is, or the smaller a region is, the more noise can be expected, implying that a stronger smoothing needs to be done compared to a mean of a larger area, for instance. The 5d-fit also represents a low pass filter; however, the amount of smoothing needed in many cases cannot be provided by this method. In the examples presented in this study, the LOESS fitting is done locally. Hence, neighboring points in time of point \( y \) are used, weighted by the distance from \( y \). The distance considered, or the size of the neighborhood is controlled by parameter \( \alpha \); thus, \( \alpha \) controls the amount of smoothing needed for the optimal fit. Generally, for \( \alpha < 1 \) the neighborhood includes proportion \( \alpha \) of the points, for \( \alpha = 1 \) all points in the time series are used. In this study only values of \( \alpha \) smaller than 1 were used, varying from 0.25 to 0.5. The same \( \alpha \) was chosen globally such that the difference to the optimum (here the “true” climate) was minimized. However, different \( \alpha \) values were chosen for different thresholds and parameters. The greatest advantage of the LOESS-fit is that it does not require the specification of a function to fit a model. Hence, no assumptions need to be made about the distribution of the data which is often problematic with daily climate data. A caveat, however, is that LOESS needs a lot of information on the local level. This means, along the daily time series, the data needs to be densely sampled; hence, a long time record is needed for the best results possible. The LOESS-fit proves to be a good choice for the fitting of the seasonal cycle as it has been used for this purpose in studies before ([Cleveland et al., 1990; Kunz et al., 2007]).

In this study, the mean, standard deviation, 5, 10, and 20% thresholds are computed for member \( x \) but also for the whole hindcast ensemble. As mentioned above, all ensemble members and all years offer a large data sample; thus, no fitting is needed when working with all the data available. These results obtained from the whole hindcast data are considered to reflect the true climate of the model. Keep in mind that the true climate does not refer to the true observed climate as this analysis is done with the context of PMA. To show the improvement of the LOESS-fit, the standard deviation and the same thresholds for member \( x \) were also derived based on the standard 5d-fit approach. The mean, however, was derived by simply averaging as this is what is usually done. Both approaches are then compared to the true climate estimated based on the full hindcast.
Figure 1. (four middle panels) The time-mean absolute difference of the surface temperature (K) to the hindcast mean (true climate) for the mean based on the LOESS-fit of member x or the raw data for surface temperature (top maps), and the absolute difference for the threshold of the 5th percentile based on the LOESS-fit or a 5d-fit (bottom maps). (top and bottom) The time series for one particular grid point, the top lines represent the mean, the lower lines the 5th percentile. The black line shows the raw time series of member x, the red line the hindcast mean (true climate), the blue line the LOESS-fit, and the green line the 5d-fit (only in case of the 5th percentile). The grey shading shows the standard deviation based on the hindcast. (right) The daily variability averaged over the entire forecasted period.
The four time series in Figure 1 show for two different grid cells (one within a tropical region and one in the higher northern latitudes) to illustrate the cases of low and high daily variability; within the two regions the grid cells are chosen randomly: the mean (top series) and the 5th percentile (bottom series) over the forecasted period without the first month (November) for surface temperature. Shown are the raw estimate, the LOESS-fit, the true climate and its standard deviation, and in the case of the 5th percentile, the 5d-fit centered on the day of interest. In case of the mean, it is obvious that a simple average over about 30 years is not a robust estimate of the true mean climate as the data are still very noisy. The variability is too large on a daily basis; a sample size of about 30 is too small. This implies that when correcting the bias on a daily basis, a simple mean of the corresponding calendar day over the reference time period does not provide enough data for a robust estimate of the climatology. The LOESS-fit of the mean, however, proves to follow the true mean represented by the hindcast mean quite closely. Thus, when using a fit to estimate the mean daily climate, the results are very likely to be closer to the true climate than a simple average.

When determining thresholds for percentile-based indices toward the tail of the distribution, the differences to the true thresholds increase and the thresholds become “noisier.” The time series of the 5% threshold in Figure 1 illustrates this clearly. Even the 5d-fit does not provide enough data for a robust daily threshold estimation. It still follows rather closely the daily time series. However, when applying the LOESS-fit, the time series resembles much more the true one based on the hindcast data as shown in the enlargement of the time series in Figure 1. Expectedly, the LOESS-fit and the true climate are not identical and can differ from each other. These differences are due to the model’s internal variability and are therefore irreducible [Hawkins et al., 2014b].

The four panels between the time series in Figure 1 show the global distribution of the absolute differences of surface temperature to the true climate averaged over the forecasted period without the first month (about 175 days). Although the difference cannot be expected to be zero, the differences for the LOESS-fit are generally smaller than for the 5d-fit. The numbers vary by region, and generally, the larger the daily variability (shown in the rightmost panel), the larger the differences. Hence, the middle to higher northern latitudes show the largest deviations from the true climate as the variability is larger in those regions [Mahlstein et al., 2011]. It is not surprising that the differences scale with increasing variability, as a particular member of the model is allowed to vary more from the mean climate in these regions. However, when using the LOESS-fit the differences are regionally reduced by up to 0.5 K.

Figure 2 shows a zonal mean summary of the absolute difference of the three different thresholds for the 5th, 10th, and 20th percentiles, the mean, and the standard deviation over land only, based on daily values but averaged over the forecasted period (excluding the first month due to model drift and dependencies...
(or the average in case of the mean). Again, the regions with a large daily variability generally show larger absolute differences. Furthermore, the closer to the tails of the distribution the threshold is (e.g., 5th percentile), the larger are the differences. However, when LOESS-fitting the data instead of using a 5 day moving window to estimate a particular threshold, the differences can be reduced substantially. In seasonal forecasting the standard deviation can be used to recalibrate the forecast to improve the reliability based on the differences found between the observed variability and the modeled one [Weigel et al., 2009]. Figure 2 shows that for a robust quality improvement of the forecast, not only the observed mean for the bias correction but also the observed standard deviation needs to be estimated robustly.

Hence, these findings have potentially strong implications for various applications. In case of seasonal forecasting, the bias removal before calculating indices can be an important step to improve the quality of the forecast, especially when indices based on fixed thresholds are used. A biased forecast only uses its actual skill, whereas a proper debiased forecast uses the full potential skill in the forecast [Murphy, 1988]. This fact becomes even more apparent when analyzing threshold-based indices as a biased forecast is not capable of distinguishing between events and nonevents when biased toward one side. Hence, this step potentially increases the skill of a forecast. However, when removing the bias, it is important that only the bias is removed and not part of the daily variability, as would be the case when working with a simple daily average. The same applies when validating seasonal forecasts of indices. Whether an event, for example, a wet day, happened or not depends on the threshold. A systematic false identification of events, or nonevents, can have skill-relevant consequences in case of seasonal forecast verification (see section 4).

Figure 3. Same as Figure 1 for the 90th percentile of precipitation (mm).
The same findings shown in Figures 1 and 2 apply to precipitation data as illustrated in Figure 3. As expected, the day-to-day variability is larger in case of precipitation compared to surface temperature. Hence, the fitting needs to be done very carefully, especially when the annual cycle has more than one maximum or minimum. Such a case would require a stronger dependency on the local environment, hence, a lower $\alpha$ term. Figure 3 (bottom right) illustrates how large the variability can be for some places. And it also illustrates that, in such cases, member $x$ may deviate from the true threshold, as due to variability, its realization was drier (in this case) than the model climate.

The results presented here can also be important for climate change studies, as the situation is quite similar. Depending on the question asked, model data are either compared to an observed reference period, or two time periods within the model framework are compared to each other [Sillmann et al., 2013a, 2013b]. Usually, these periods consist of only one model/member; hence, the sample size is not sufficiently large in order to estimate robust thresholds. Consequently, these thresholds might not be robust. Zhang et al. [2005] proposes a methodology to improve the threshold estimation for climate trend studies (which was applied by Sillmann et al. [2013a, 2013b]), taking into account the inhomogeneities in percentile exceedance between estimates within and out of the reference period. Their methodology uses a Monte Carlo approach to increase the sample size. The approach presented here can also be applied in climate change studies and can be seen as an alternative as it improves the threshold estimation without a large computational effort. However, our approach does not take into account the boundary issues discussed by Zhang et al. [2005].

Figure 4 shows an example for one specific index, in this case, for wet days. The area in orange (over land only) shows where the LOESS-fit is as good or better than the 5d-fit when estimating the number of wet days for December-January-February (DJF) for one particular year. The DJF period is shown in order to illustrate the difference over one season. Specifically, the estimated numbers of wet days of the LOESS-fit and the 5d-fit are compared to the number of wet days estimated from the full hindcast data set for 1 year in DJF. However, the results of other years or the mean over the hindcast period look very similar. Averaged over all years, 72% of the grid cells are as good as or better than the 5d-fit. Furthermore, the area where the LOESS-fit shows no difference to the estimated number of wet days compared to the full hindcast data set is substantially larger than that compared to the 5d-fit. Hence, the LOESS-fit yields more often the true estimated number of wet days than the 5d-fit. The stippled area in Figure 4 shows the area that is significantly improved, meaning that in all 32 years used, these grid cells are improved or as good as 5d-fit, which poses a very rigid criteria of significance. The stippled area happens to be the area with the largest variability [Mahlstein et al., 2012], which implies that especially for noisy climate data, a fit will yield a better result.

4. Examples Based on Reanalysis

The remaining question is how the method introduced above performs outside of the perfect model approach. Therefore, the two approaches are compared to each other with the aid of the example of
A warm night is defined as the minimum temperature that is above the 90th percentile of the long-term climatology of minimum temperatures. This index was chosen as it is threshold based depending on the percentile, yet the percentile is still quite moderate. Hence, for each day of the year, the 90th percentile was determined based on the LOESS-fit and the 5d-fit. Each approach yields a number of warm nights a year; the time period considered are 30 years starting in 1981. Then, the absolute difference between the numbers

Figure 5. (a) Mean absolute difference in number of warm nights a year between the 5d-fit and the LOESS-fit averaged over the time period 1981–2010 of the ERA-Interim data set. (b) Difference in skill (anomaly correlation) averaged over a 30 year hindcast period (1981–2010) between the two approaches for JJA. Positive values indicate an improvement in skill by the use of the LOESS-fit compared to the 5d-fit and negative a decline in skill. (c) Anomaly correlation between System4 and LOESS-fitted ERA-Interim data for warm nights over the hindcast period 1981–2010 for JJA. (d) Difference in skill (anomaly correlation) averaged over a 30 year hindcast period (1981–2010) between the two approaches for JJA for the number of wet days. Positive values indicate an improvement in skill by the use of the LOESS-fit compared to the 5d-fit and negative a decline in skill.
of each year were computed and then averaged over the 30 year period. Figure 5a shows that the mean differences between the two approaches vary regionally but can be quite substantial. Globally averaged, the difference is about six nights, which is a difference of about 16% on average. Locally, the differences can be quite large which changes the number of warm nights significantly. This implies, that in case of a seasonal forecast verification, quite a large number of events are either false events (false alarms) or false non-events (missed events). As mentioned above, the false identification of an event, in this case, a warm night, or a false non-identification can have skill-relevant consequences. To quantify these consequences, the hindcast data set of number of warm nights in June-July-August (JJA) based on the System4 May initialization data was evaluated twice; once against the observed number of warm nights from ERA-Interim based on the 5d-fit, and once based on the LOESS-fit. The difference in the anomaly correlation can be as large as 20% as shown in Figure 5b. On average it is around 5%, but for areas with low skill, even this can be substantial. The difference in skill can be positive or negative. The example shown in Figure 5b shows an increase in skill of about half of the land area, the actual skill of the LOESS-fit is shown in Figure 5c. Depending on the characteristics of the variability, the LOESS-fit of the observed thresholds to identify a warm night can be closer to the model mean climate or further away; hence, the skill can improve or decline, however, not in a systematic way. For instance, the threshold in the observations for an event without fitting is 10.2, while the fitting yields a threshold of 10.3. Whether the skill of the model improves depends solely on the model forecast, whether it forecasted the event or not. If the observed value is 10.25 and the model forecasted an event, without the fitting the model would yield a better skill than with the fitting of the observations. On the other hand, if the model did not forecast an event, the skill is improved when the fitting approach is applied to the observational data. Hence, it is important to note that a proper characterization of events in the observations does not result in an overall improvement of the model skill. The method simply helps to quantify the true skill of the model as the reality can be estimated more adequately. Or in case of a bias removal the method improves the skill of the model as the bias can be estimated more precisely because the mean observed climate is known to be more precise. The important finding however is that the method of estimating percentile thresholds for indices has skill-relevant consequences.

Figure 5d shows the same as Figure 5b but for the wet days index. Again, this index is threshold based but the percentile is quite moderate. The results are very similar to the ones of the warm night index. This illustrates that also in case of a precipitation-based index the skill is neither improved nor decreased in a systematic manner. Depending on the variability, the skill changes according to the better quantification of the events.

5. Conclusions

Climate indices are very useful tools to describe the current climate, the forecasted climate, or the expected climate changes in an applied way to users or even the public. Indices aggregate information in a nonlinear way, and this information tells more about possible impacts than simple temperature averages, for instance. Indices are often based on daily data and/or thresholds. These thresholds often define whether an event happened or not. Hence, an accurate estimate of the threshold itself is very important, and therefore, it needs to be estimated carefully. This issue is further complicated by the strong biases present in current climate models. These biases have to be corrected by comparing the model data to observational data covering the same period. This study employs the large sample of historic seasonal forecasts to determine the effect of smoothing the daily climatology by a fitting approach. It is shown that the estimates can be improved and the thresholds are closer to the “true ones” compared to a 5 day moving window being used in previous studies [Klein Tank et al., 2009] or a simple average as in the case of the estimation of the mean climate. A fitting procedure such as the LOESS fitting [Cleveland and Delvin, 1988] reduces the short-term variability substantially by keeping, as much as possible, a complex annual cycle. No particular fitting method is advocated in this study for the simple reason that there is no perfect fitting for all purposes. The fitting needs to be done carefully and can differ depending on the percentile, region, and parameter of interest. However, especially in regions with high variability, a proper fitting approach will yield better results than a moving window. Other approaches have been discussed in the literature, such as a Monte Carlo simulation [Zhang et al., 2005], which focuses on homogeneity issues at the border of the reference period. Compared to this approach, the fitting approach can be easily applied in long-range forecasting where usually the whole time period available is taken into account and therefore no boundaries exist. Alternatively, to increase the sample size, a larger time window could be used, but by doing so it is likely to attenuate the annual cycle.
of the threshold, which is particularly problematic for extremes. Folland et al. (1999) suggest to fit a statistical distribution to the daily data sample in order to get a robust estimate of the threshold. Compared to this approach, the method described here is simpler to use as the fitting needs to be done only once and keeps the annual cycle in the focus. This could be an advantage for operational use because of its simplicity or when dealing with large data sets (e.g., CMIP5, decadal and seasonal forecasts).

Generally, the higher the day-to-day variability, the more difficult it is to estimate the true thresholds. However, while the 5d-fit preserves much of the short-term variability, the LOESS-fit is able to smooth these fluctuations. Applied on an index, it shows that the LOESS-fit distinguishes better between an event and a non-event. Hence, it is recommended to use a fitting approach to estimate thresholds for index calculations. Also, for bias removal procedure, it is important to gain good knowledge about the true climate based on which the bias is derived from. Again, the fitting approach is usually very close to the true mean climate. Based on this information, the calculation of the bias removal can be improved and therefore also the forecast quality. The main advantage of the method presented here is that the LOESS fitting improves the estimate of the seasonal cycle of daily means or daily thresholds and thus allows for a good comparison between two different reference periods (e.g., observations and models, or model past to model future).

Acknowledgments
We would like to thank Claudia Mignani for the support in code developing. The research leading to these results has received funding from the European Union’s Seventh Framework Programme [FP7/2007–2013] under grant agreement 308291. The data for this paper are available at ECMWF [System6: http://old.ecmwf.int/data/d/interim_daily/].

References
Alexander, L. V., et al. (2006), Global observed changes in daily climate extremes of temperature and precipitation, J. Geophys. Res., 111, D05109, doi:10.1029/2005JD006290.

Brandis, S. (2013), Skillful seasonal predictions of boreal winter accumulated heating degree-days and relevance for the weather derivative market, J. Appl. Meteorol. Climatol., 52(6), 1297–1302, doi:10.1175/JAMC-D-12-0303.1.

Christensen, M., H. Manz, and D. Gyalistras (2006), Climate warming impact on degree-days and building energy demand in Switzerland, Energy Convers. Manage., 47(6), 671–686, doi:10.1016/j.enconman.2005.06.009.

Cleveland, R. B., W. S. Cleveland, J. E. McRae, and I. Terpenning (1990), STL: A seasonal-trend decomposition procedure based on Loess, J. Offic. Stat., 6(1), 3–73.

Cleveland, W. S., and S. J. Delvin (1988), Locally weighted regression: An approach to regression analysis by local fitting, J. Am. Stat. Assoc., 83(403), 596–610.

Collins, M., B. B. Booth, G. R. Harris, J. M. Murphy, D. M. H. Sexton, and M. J. Webb (2006), Towards quantifying uncertainty in transient climate change, Clim. Dyn., 27(2–3), 127–147, doi:10.1007/s00382-006-0121-0.

Dee, D. P., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137(656), 553–597, doi:10.1002/qj.828.

Della-Marta, P. M., H. Mathis, C. Frei, M. A. Liniger, J. Kleinn, and C. Appenzeller (2009), The return period of wind storms over Europe, Int. J. Climatol., 29(3), 437–459, doi:10.1002/joc.1794.

Deser, C., A. Phillips, V. Boulanger, and H. Teng (2012), Uncertainty in climate change projections: The role of internal variability, Clim. Dyn., 38(3–4), 527–546, doi:10.1007/s00382-010-0977-x.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer (2005), The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination, Tellus, Ser. A, 57(3), 234–252, doi:10.1111/j.1600-0870.2005.00104.x.

Fischer, E. M., K. W. Oleson, and D. M. Lawrence (2012), Contrasting urban and rural heat stress responses to climate change, Geophys. Res. Lett., 39, L03705, doi:10.1029/2011GL050576.

Fischer, E. M., U. Beyerle, and R. Knutti (2013), Robust spatially aggregated projections of climate extremes, Clim. Change, 111(3), 313–338, doi:10.1007/s10584-012-0491-6.

Folland, C. K., C. Miller, D. Bader, M. Crowe, P. Jones, N. Plummer, M. Richman, D. E. Parker, J. Rogers, and P. Scholefield (2002), A comparison of methods, A Statistical Calibration and Combination, 31(3), 437–459, doi:10.1017/S0002264702000107.

Gangsto, R., A. P. Weigel, M. A. Liniger, and C. Appenzeller (2013), Methodological aspects of the validation of decadal predictions, Clim. Res., 59(3), 181–200.

Hanlon, H. M., G. C. Hegerl, S. F. B. Tett, and D. M. Smith (2012), Can a decadal forecasting system predict temperature extreme indices?, J. Clim., 26(11), 3728–3744, doi:10.1175/Jcli-D-11-00512.1.

Hawkins, E., T. M. Osborne, C. K. Ho, and A. J. Challenor (2013), Calibration and bias correction of climate projections for crop modelling: An idealised case study over Europe, Agric. For. Meteorol., 170, 19–31, doi:10.1016/j.agrformet.2012.04.007.

Hawkins, E. B., D. Robson, R. Sutton, and D. Smith (2014a), The interpretation and use of biases in decadal climate predictions, J. Clim., 27(8), 2931–2947, doi:10.1175/JCLI-D-13-00473.1.

Hawkins, E., et al. (2014b), Uncertainties in the timing of unprecedented climates, Nature, 511(7507), 3–E5.

Klein Tank, A. M. G., F. Zwiers, and X. Zhang (2009), Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation, Geneva, Switzerland.

Kunz, H., S. C. Scherrer, M. A. Liniger, and C. Appenzeller (2007), The evolution of ERA-40 surface temperatures and total ozone compared to observed Swiss time series, Meteorol. Z., 16(2), 171–181, doi:10.1127/0941-2948/2007/0183.

Lafon, T. S., N. Dodson, G. Buys, and C. Prudhomme (2013), Bias correction of daily precipitation simulated by a regional climate model: A comparison of methods, Int. J. Climatol., 33(6), 1367–1381, doi:10.1002/joc.3518.

Mahstein, I., R. Knutti, S. Solomon, and R. W. Portmann (2011), Early onset of significant local warming in low latitude countries, Environ. Res. Lett., 5(3), 034009, doi:10.1088/1748-9326/5/3/034009.

Mahstein, I., R. W. Portmann, J. S. Daniel, S. Solomon, and R. Knutti (2012), Perceptible changes in regional precipitation in a future climate, Geophys. Res. Lett., 39, L05701, doi:10.1029/2011GL065738.

Molteni, F., T. Stockdale, M. Balmaseda, M. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, and T. Palmer (2011), The new ECMWF seasonal forecast system (System 4).
Müller, W. A., C. Appenzeller, and C. Schär (2005), Probabilistic seasonal prediction of the winter North Atlantic Oscillation and its impact on near surface temperature, Clim. Dyn., 24(2–3), 213–226, doi:10.1007/s00382-004-0492-z.

Murphy, A. H. (1988), Skill scores based on the mean square error and their relationships to the correlation coefficient, Mon. Weather Rev., 116(12), 2417–2424, doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

Piani, C., J. O. Haerter, and E. Coppola (2010), Statistical bias correction for daily precipitation in regional climate models over Europe, Theor. Appl. Climatol., 99(1–2), 187–192, doi:10.1007/s00704-009-0134-9.

Sillmann, J., V. V. Kharin, X. Zhang, F. W. Zwiers, and D. Bronaugh (2013a), Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate, J. Geophys. Res. Atmos., 118, 1716–1733, doi:10.1002/jgrd.50203.

Sillmann, J., V. V. Kharin, F. W. Zwiers, X. Zhang, and D. Bronaugh (2013b), Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections, J. Geophys. Res. Atmos., 118, 2473–2493, doi:10.1002/jgrd.50188.

Themeßl, M., A. Gobiet, and G. Heinrich (2012), Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal, Clim. Change, 112(2), 449–468, doi:10.1007/s10584-011-0224-4.

Weigel, A. P., M. A. Liniger, and C. Appenzeller (2009), Seasonal ensemble forecasts: Are recalibrated single models better than multimodels?, Mon. Weather Rev., 137(4), 1460–1479, doi:10.1175/2008MWR2773.1.

Zhang, X., G. Hegerl, F. W. Zwiers, and J. Kenyon (2005), Avoiding inhomogeneity in percentile-based indices of temperature extremes, J. Clim., 18(11), 1641–1651, doi:10.1175/JCLI3366.1.

Zubler, E., S. Scherrer, M. Croci-Maspoli, M. Liniger, and C. Appenzeller (2014), Key climate indices in Switzerland; expected changes in a future climate, Clim. Change, 123(2), 255–271, doi:10.1007/s10584-013-041-8.