Response to A. Mondal:

S. Sippel et al., 2015

The paper presents a novel bias correction technique for use in climate change impact assessment studies. The technique is claimed to preserve the physics as well as multivariate dependence structures. Benefits of the proposed technique in comparison to the existing methods are categorically brought out and an end-to-end application is also illustrated using an impact-assessment study. The paper is overall very well written and will be of interest to a wide range of researchers. Therefore, I would favor its publication.
We thank the reviewer for the encouraging feedback.

However, I have a few comments/suggestions as I detail below and would like to see the authors' responses to them. Since the proposed bias correction methods leads to a decrease in effective ensemble size, large ensembles such as the weather@home experiment is necessary for its application. In my opinion, this is a strong limitation of the method as such large experiments are rare, particularly for developing regions. Is the proposed technique also effective on GCM simulations directly?
We agree with the reviewer, this issue is an important consideration for the applicability of our method.
The methodology's applicability depends on a few factors (see also discussion in the paper):

1. (ensemble) sample size and the anticipated application
2. structure of the bias in the model (if the model produces unrealistic simulations in all ensemble members, then our bias correction cannot improve the simulations).

Regarding (1): As demonstrated in the paper, large ensembles are quite useful (but are increasingly available also for many parts of the world, see e.g. http://www.climateprediction.net/weatherathome).
However, also smaller of 5-10 member ensembles could be used for this type of bias correction, if one assumes ergodicity over a certain number of years. E.g. assume an CORDEX-RCM simulation with 10 members over each 50 years. In this case, one could still resample from 500 ensemble members, hence a useful sample size should remain. In this context, the type of application also plays a role: A so-derived ensemble could be well interpreted probabilistically, but of course "strict year-to-year continuity" would be lost.

Regarding (2): In addition to (1), of course the "magnitude" of the biases is important. Consider e.g. Fig. 2d in the revised manuscript: No loss in effective sample size would require that for each decile in the observations, 10% of the original ensemble members are available for resampling. In our case, for the "worst" deciles the effective sample size is reduced roughly by 50% (Fig. 2d). This way, one could estimate whether the effective sample size after resampling for a smaller ensemble is large enough for the anticipated application (which of course is a function of the "severity" of the bias and the number of ensemble members).

We have stressed both points more clearly in the discussion section.

I also have concerns with the quantile mapping based technique for more general applications of the proposed bias correction method. The retention of an ensemble member depends on q_mod as given by the transfer function. Therefore, if a model simulated value does not correspond to a quantile of the observed record, that value is rejected, thereby indirectly defining a prescribed range of possible values of the variable based on certain number of years of observations. For bias correction of future values, clearly, there is no way to ensure that the actual values belong to that range.

We agree with the reviewer in that the choice of the resampling constraint is a critical step for the applicability of the methodology. However, we would like to add two comments:

1) Present-day bias correction: For transient bias correction, the underlying assumption is that the actual range of variability in the observations is representative, as correctly pointed out by the reviewer (but not fully restricted to that as in quantile-quantile mapping or similar approaches, due to the kernel bandwidth that allows some flexibility). However, for this reason we argue in the paper that an "aggregated" constraint (i.e., aggregated both in time (JJA) and in space (Central Europe)) is more useful than constraints define on short periods or single grid cells, because the resampling would then be very sensitive to the observed variability. We have tried to discuss this point more clearly in the manuscript.
2) Bias-correction of future simulations: For future simulations, clearly a direct application of the bias correction based on resampling a range of absolute values is not meaningful (as correctly stated by the reviewer). However, if one assumes that the structure of the biases relative to the observations holds in the future (i.e. not in absolute values, but in the percentiles of model simulations relative to percentiles in the observations, see e.g. Fig. 2b in the revised manuscript), the bias correction is still applicable, as the transfer function is defined as a mapping between the percentiles and calibrated on the present (e.g. Fig. 2b, see also our response to Referee #1). Of course this is a rigorous assumption, and we are not arguing that this should be assumed, but this kind of "stationarity assumption of the bias structure" underlies implicitly all bias correction approaches for future simulations.

Further, selection of Gaussian kernels seem somewhat arbitrary. It is a subjective choice, and so is the choice of Cubic Hermite splines.

This is of course correct.

The choice of the Gaussian kernel is motivated by the choice of the resampling metric and the Central limit theorem: Since the constraint is quite highly aggregated (JJA means over a relatively large region, Central Europe), we believe that the choice for a Gaussian kernel seems somewhat "natural".

Cubic Hermite splines for interpolation are clearly an arbitrary choice, but the form of the transfer function is almost entirely determined by the two kernels over observations and the model ensemble (because both kernels allow resampling of an arbitrarily large number of random variables).

Additionally, in my opinion, more clarity is solicited in the description of the proposed bias correction methodology.

We thank the reviewer for highlighting the need for a "cleaner" methodological description.
We have redrawn the figures for the methodological illustration (Fig. 2b and 2c). We hope that these now better reflect the procedure how the transfer function is obtained as a mapping between percentiles in the observations and model ensemble?
Additional changes to the methodological description are highlighted below as a response to the reviewer's comments.

For example, do the authors simply concatenate observed data listed in Table 1? How do they fit the kernel density 'over the observed meteorological constraint in various observational datasets' (blue cdf in Figure 2(a)? How are the 800 ensemble members merged to obtain the red cdf of Figure 2(a)? The authors also mention that they derive a bias-corrected sample by 'randomly resampling n times from f_obs': what is the length of the sample?
Further, q_mod_X and q_obs_X represent a given quantile in the model ensemble and observation, respectively. Does this then imply that bias correction is carried out individually for each quantile?

One dataset is used at a time. Concatenation or another form of combination of different observational datasets would be an option, but could result in somewhat "strange" distributions (e.g. if one dataset is simply offset relative to another one would yield a bimodal distribution). Therefore, figures 4 and 5 contain several lines in the return time plots, one for the correction with each observational dataset separately. Since the different observational datasets were very similar to each other for the aggregated temperature constraint (Fig. 4a, b), only the ERA-Interim constraint was used for the LPJmL simulations. This has been clarified in the manuscript.
The Gaussian kernel density fit uses a bandwidth estimation procedure following Sheather and Jones (1991). Subsequently, resampling is done by 1) sampling n times with replacement from the respective observations (e.g. x_i), and 2) sampling from a Gaussian distribution with mean x_i and the bandwidth h as the standard deviation (definition of the Gaussian kernel). The Gaussian kernel fit is identically applied to he observations and the model ensemble (i.e. the temperature constraint is obtained from each ensemble member and each year and the Gaussian kernel is fitted over all of them). The length of the sample is n=800 for the illustrative application and probabilistic interpretation in the manuscript.
Bias correction is done by resampling percentiles from the observations and retaining the ensemble member that corresponds as given by the transfer function (in that sense, bias correction is applied individually to each resampled (random) percentile).

We have clarified these issues in the Methods section of the manuscript.

For fitting the GEV distribution, though the length of all the observed records listed in Table 1 is greater, the authors mention about a 'relatively small sample size (1901-2014)'. I did not understand why (why not all 26 years?). Also, statistical extreme value theory requires certain conditions to be held true for application of the GEV distribution to the block maxima. If 10-year samples are 'randomly concatenated', the tail behaviour may change, thereby questioning the application of extreme value theory to the concatenated datasets. Another, more fundamental issue concerns the random nature of the model output. The bias corrected variables are after all output of models

that are deterministic in nature; therefore, whether they can be considered as random variables remains a question.
We agree with the reviewer's concerns regarding concatenation of observational datasets. This might change the tail behaviour and could lead to many other problems. Therefore we have clarified in the revised manuscript that we do not concatenate observational datasets, but perform the analysis with each observational dataset separately. Hence, 10-year sample are concatenated only from the same observational dataset (following the procedure outlined in the manuscript), thus assuming ergodicity. Regarding 10-year samples from the model ensemble, we would like to add that each ensemble member (i.e. each year in the ensemble) has been initialized separately, i.e. each year can be considered as a random realization (therefore we believe that this resampling procedure is appropriate).
The reviewer also addresses a more fundamental issue, namely the random nature of the model output. Here, we also agree with the reviewer, but would like to add another comment: For deriving the initial-condition ensemble that is used in our study, initial conditions for each ensemble member are perturbed randomly at the beginning of each year. Therefore, different years in the ensemble (but not different months...) can indeed be regarded as random realizations.

Other points:
Abstract, last line: 'uptake of our methodology. . .for accurately quantify- ing past. . .extremes' – how is bias correction important for quantifying past extremes which have been already observed? Perhaps the authors mean 'quantifying changes in past extremes'?
This is correct. Thanks.
Page 2011, first sentence – this information is repeating for the third time here.
This sentence has been removed.
Page 2021, Para 15: 'Although more sophisticated. . .in this study' – perhaps a 'that' missing?
For readability, sentence has been rewritten.
All references listed contain two years of publication each – please correct this. Also, Coles, 2001 is a single-author book. The reference to Coles, 2001 is incorrect in the list.
Thanks for pointing this out. It's corrected in the revised manuscript version.
Figure 3 (and similar figures) and Section 4.1 – Figure 3 is not self-explanatory. If the x-axis doesn't consist of values/units, then what to the width of each shape represent?
The description was added to the figure caption of Fig. 3 and Fig. 6: "Both sides of each violin are constructed as rotated, equal-area kernel density estimates, and a standard boxplot is drawn inside each violin."