Reviewer #1 - 'Comment on egusphere-2023-1481' - September 14th, 2023

Comments by the reviewer in blue italics - responses by the authors in black

Dear reviewer,

Thank you for your constructive criticism of the GMD paper submission and the software package. We have addressed your comments, please find details of our response below.

I find the effort of making this tool for evaluating and inter-comparing bias correction methods important and highly relevant. The paper is well structured, and although I haven't tested the software itself, it seems from the manuscript that it is producing many relevant statistics and useful plots. The manuscript lacks important details, and I have several suggestions for how to compare the bias correction methods in my major comments, which I wish the authors to consider and make a revision.

Major comments:

I think the idea of implementing multiple methods in one software is a good idea, as is the common framework for evaluation.

One major remark I have is that some components are not tied to a certain method. One example is the treatment of dry days, where the threshold and the way they are bias corrected will have impacts on certain metrics that have nothing to do with the method applied to the rest of the distribution. An example is the results for your "QM" method. I suggest that this component is detached in a way that the same treatment is applied to all methods when making the inter-comparison. The dry day corrections themselves can be assessed separately. Another example is how extremes and data outside the calibration data range are handled. This is often not properly defined for different methods, but can have large consequences for indicators based on the extreme ends. When possible, the same tail handling should be applied to all (empirical) methods. If possible to implement in your software, this would make it a very useful tool to assess and find the method best suited for a particular case of bias correction. I am not expecting you to implement this for a new revision, but please think about it and add to a discussion section on future developments.

Response: We thank the reviewer for this helpful comment. We agree with the comment, and the modularity of different methods was certainly a key challenge we discussed in detail when we were developing the architecture of the ibicus software.

In principle, we agree that many of these different choices are interchangeable across methods (within certain limits). However, practically speaking and after some deliberation when developing the package, we decided that it would not be feasible to make the package fully modular (to enable, for example, the application of the same dry day treatment or handling of extremes across methods) for two related reasons: one is that different choices can be entangled, and the second is that some choices such as the dry day treatment can be seen as 'core characteristics' of a specific method. For example, in the case of the ISIMIP method, the dry day treatment fits into their general approach of bias adjusting bounded variables and is being developed further in new versions of the method (see

Lange 2021 eq. 1 in comparison to Lange 2019 eqs. 8 and 9). In the case of Scaled Distribution Mapping (SDM), the dry day treatment is defined as part of the core characteristics of the method itself (the number of rainy days is mapped using formula 1 defined in Switanek et al. 2017) and in the case of CDFt, the Stochastic Singularity Removal technique corrects both occurrence and intensity jointly non-parametrically based on Vrac et al. (2016).

The approach we therefore took in the design of the ibicus software architecture design was to define some components as core characteristics of each method, making selected options interchangeable for specific methods but not for others. This means the package ends up being partially modular. For example, the dry day treatment can be modified in the Quantile Mapping method: options range from a mapping using a censoring approach similar to the one outlined in Cannon et al. 2015, a fit of a precipitation hurdle model together with randomization to be able to map dry days into wet ones and an adjustment of intensity only as well as possible user provided different adjustments.

We agree with the reviewer that it would be a good idea to develop the modularity of the different methods further in future versions of the ibicus software, and we also agree that the discussion of the modularity of the different methods was previously not detailed enough in the paper. We have added several sections in the text (in the background section and description of the ibicus package) that highlight the modularity of methods, using the treatment of dry days as an example. Furthermore, we expanded table 1 that gives an overview of the choices made in different methods to include the treatment of dry days and extremes and added an additional table in the appendix to detail the treatment of dry days in different methods. For each individual method (family), the package documentation gives a further detailed overview of adjustments that are possible within the ibicus package.

Regarding the case study presented in the paper, our aim was to compare different methods as they are commonly applied. For ISIMIP and SDM, the dry day treatment is, in our interpretation, part of what defines the method in how it is most commonly applied, and we therefore chose to not change these aspects of the method in the case study. For QM and LS we have added some text commenting on the dry day treatment used.

Text modifications:

- Background and ibicus description see latex diff document
- New row in table 1: “Treatment of dry days and extremes – Methods have different ways of handling certain aspects of the distribution such as extreme values or dry days in the case of precipitation. For extremes some methods use an extrapolation based on parametric distribution, which can be modified by the user for example should a mapping based on extreme value theory be required. For dry days the ISIMIP, SDM and CDFt methods provide an explicit handling that might be appropriate in some situations but not in others. QDM treats the mapping of dry days as a censoring problem and adjusts them together with the body of the distribution whilst for methods like QM and ECDFM the user has the choice of different treatment methods.”
- New table on the treatment of dry days in the appendix.

I would like if the authors can add some information on how a user can implement their own method in Ibicus, what are the steps? Is there a guide in the software documentation etc. And why is It called ibicus?

Response: We thank the reviewer for this idea and agree that a guide for users on how to implement a new method in ibicus would be very useful to have in the software documentation. While such a guide is outside the scope of this GMD paper, we will include a new tutorial in the software documentation with the next ibicus release.

The package is called ibicus because we felt that it was a suitable name for a software package. It is not an acronym but rather just sounds similar to abacus as well as ibis which is the name of a bird and therefore in line with package naming conventions at ECMWF.

The method called "Quantile Mapping" is not properly defined and named. This is a category of methods that includes most of the methods that are used in this paper. A more precise name is necessary, and also a detailed description of how the quantile mapping is implemented, e.g. which quantiles, how extremes are dealt with, and especially for data outside of the calibration range.

Response: We agree that a large number of choices are possible when applying QM leading to fairly different "flavours" of the method and a "family" of QM methods (as detailed in the comment above regarding the modularity of different methods). In fact, most methods in the package have some sort of quantile mapping in this form as their core which is then expanded in various ways. Related to the idea of making the package partially modular (see response above), we believe that it is useful to include a generic quantile mapping method that can be modified in many different ways.

Based on publications such as Maraun (2016), we define Quantile Mapping (QM) as a fairly concrete method based on the mapping of two cumulative distribution functions as $F_{cm,fut}^{-1}(F_{cm,hist}(obs))$, as detailed in the appendix table 1, as well as the documentation of the software package. We believe that despite the many different variations of Quantile Mapping the package offers, this defines it as a method. We agree with the reviewer that further specification is necessary when applying the method in the case study and have added an additional sentence in the case study providing details of the type of quantile mapping applied there.

Generally, the authors need to more clearly defined how each method deals with dry days as this can have a large impact on some of the statistics. Please revise with some statements about this in the main text and in Table A1.

Response: We have adjusted the main text in a couple of locations (background section, ibicus description section and table 1, see below). As the table A1 is supposed to outline the implemented methods for all 10 supported climatic variables we opted against including additional information on the treatment of dry days directly in there as this would impede readability. Instead, we added a second table A2 containing additional information about dry day treatments. We also adapted the text in section 4 (the case study) stating what dry day treatment was used for the individual methods. As for ISIMIP and SDM the dry day treatment is closely entangled with the method design; a conscious choice was only made for QM, which we state in section 4.1:

New text in italic: "These four methods are applied to daily total precipitation (pr) and daily minimum near-surface air temperature (tasmin), chosen to cover two different types of variables (bounded vs unbounded, different distributions etc) that are both highly relevant for many impact studies. *The bias adjustment methods are used with their ibicus default settings for both variables (for more details see table A1 and the software documentation). This means that the ISIMIP and SDM methods provide an explicit adjustment of dry day frequencies, whilst for QM they are treated as censored and the method based on Cannon et al. (2015) is applied and LS provides no explicit adjustment, scaling all values.*"

Finally, we have added some words on future developments in the discussion section highlighting that extending the modular aspect of certain methods and the customizability is something that we plan to implement in future versions of ibicus and that will increase comparability of bias adjustment methods.

It is not clear what future is for which the climate trends are calculated. Is all done between periods in the historical range 1959 to 2005? I cannot find any other information about time periods, nor any information about SSP-scenarios used. Please clarify this point.

Response: We have adjusted the text as follows:

"Table B1 in the appendix provides more details on these models. We used the historical runs as well as the SSP5-8.5 experiments."

Original text: The data ranges from January 1st, 1959 to December 31st, 2005, with the initial 30-year period (1959-1989) serving as the historical/reference period and used as a training dataset and the subsequent 15-year period (1990-2005) used for validation purposes.

Modified text: "The historical data ranges from January 1st, 1959, to December 31st, 2005, with the data from January 1st, 1959, to December 31st, 1989, serving as the historical/reference period and used as a training dataset and the subsequent period: January 1st 1990 to December 31st 2005 used for validation purposes. Bias adjustment is applied to the validation period as well as the future period: January 1st, 2080, to December 31st, 2100."

It is also important to state something about the magnitude of the climate trend, as it gives some information about the signal to noise levels and whether differences between methods are significant or not.

Response: Thank you for this comment, the magnitude of the climate change trend is indeed relevant to specify when showing these results. We have added the range of the raw trend across locations for both dry days and mean minimum daily temperature in the respective figure captions.

Added sentences in the figure captions:

Figure 6: [...] The magnitude of the raw projected change in dry days depends on the climate model and, across different locations, lies between 10 fewer and 30 more dry days on average per year.

Figure 7: [...] The magnitude of the raw projected change in mean minimum daily temperature again depends on the climate model and, across different locations, lies between 2-5K.

Detailed comments:

L22: "an empirical transfer function" this could equally well be a parametric, so please remove the word "empirical".

Response: Thank you for your comment, we changed the term empirical to statistical in all instances. We had originally called the transfer function empirical as it is 'based on data' as opposed to 'based on theory', even if a parametric fit is used. We had wanted to avoid using the term statistical as it sometimes implicitly implies machine learning methods as an alternative. However, based on the comments of both reviewers, we concluded that using the term empirical here is potentially misleading and therefore changed it.

L26: This sentence is a bit difficult, which has to do with the vague part "ranging from" which lists generic reports which are not throughout using bias adjustment. Please reformulate and be more precise in your statement.

Response: We disentangled the sentence as follows:

Original text: Despite widespread use, ranging from the IPCC AR6 WGI&II (IPCC, 2021, 2022) reports to the climate scenarios used by central banks across the world (NGFS, 2021), Maraun et al. (2017) and others highlight fundamental issues with statistical bias adjustment and show that the approach can destroy the spatiotemporal and inter-variable consistency of the climate model and is prone to misuse.

New text: Despite widespread use both within the scientific community (see, for example, IPCC, 2021, 2022), as well as by climate service providers and practitioners (see, for example, climate scenarios used by central banks across the world, NGFS 2021), bias adjustment is known to suffer from fundamental issues. These issues have been highlighted, among others, by Maraun et al. (2017) who show that bias adjustment not only has limited potential to correct misrepresented physical processes in the climate model but can also introduce new artefacts and destroy the spatiotemporal and inter-variable consistency of the climate model.

L61: please remove "empirical".

Response: Corrected, see comment above.

L114: "MIdAS" according to the reference.

Response: Corrected in the text.

L150: Please clarify what is meant with "optional data information for running windows".

Response: We have adjusted the sentence as follows:

“This takes a 3-dimensional numpy array of observations, as well as historical and future climate model simulations as input. Bias adjustment methods using a running window might require date information as 1-dimensional arrays.”

L153: In which contexts are these methods “most widely used”? Can you quantify this?

Response: We have adjusted the sentence (see below). Furthermore, we have included a sentence in the future developments section that ibicus might be extended for more methods in future releases.

<u>New text:</u> “The methods were chosen to cover some of the most widely used bias adjustment methods. These methods were cited thousands of times, have been used in major coordinated projects like the ISIMIP project (Lange, 2019) and are based on fairly different assumptions, making them suitable for different applications.

L157: “Quantile Mapping”. I do not think this is a good way to describe this method in contrast to the others. They are all of the quantile mapping family, and there is no single clear definition of what quantile mapping is, but it needs to be clearly defined. If you are referring to detrended quantile mapping (as in Table A1), you could use the abbreviation DQM instead.

Response: Please see the response on the comment above. To differentiate plain Quantile Mapping from the mean trend preserving variant (detrended Quantile Mapping – dQM) we prefer to use the abbreviation QM here.

L170: The case is a bit more complicated when threshold-based indicators are used. Then it is not possible, or wanted, to preserve the original trend.

&

L268: Note no single method is attempting to preserve trends for all possible indicators, but target a single or more moments or quantiles of a distribution.

Response: We completely agree with these comments. Regarding the first comment, we added a more detailed discussion of trend modification through bias adjustment in the background section. Regarding the second comment, we argue not that we expect the trend in all indicators to be preserved, but rather that it is necessary to evaluate how the trend is modified. By pointing out that even trend-preserving methods can change climate change trends we want to convey the two points highlighted by the reviewer, namely: 1) No method will preserve the trend in every indicator. In our experience, this is quite commonly misunderstood by users of trend-preserving bias adjustment methods. For example, the preservation of the trend in dry days can be stated in the publication of some methods, and assumed when the method is used, but not evaluated (eg. Lange, 2019). 2) Even if a method aims to preserve certain trends the success of this is dependent on whether the underlying assumptions of the method are met.

L195: It is not clear what time periods are used, and what the “future” is for which the trends are assessed. See major comment.

We have added the following:

“Table B1 in the appendix provides more details on these models. We used the historical runs as well as the SSP5-8.5 experiments.”

“The historical data ranges from January 1st, 1959 to December 31st, 2005, with the data from January 1st 1959 to December 31st 1989 serving as the historical/reference period and used as a training dataset and the subsequent period: January 1st 1990 to December 31st 2005 used for validation purposes. Bias adjustment is applied to the validation period as well as the future period: January 1st 2080 to December 31st 2100.”

L209 and 210: There are 31 and 16 years in these periods. Please check your statements.

Response: That is correct, we have amended the text accordingly (see response to the last comment).

L215: Please define “temperature” in this sentence. Is tasmin still indended, or some other temperature measure?

Response: We have adjusted the sentence to refer specifically to tasmin:

“– tasmin greater than the seasonal 95th percentile of the daily minimum temperature in each grid cell during the historical period (1959- 1989). This can be an indicator of the impacts of heatwaves (Raei et al., 2018).”

L222: Bias should be near zero for the calibration period, and it would be good to know if that is the case as it is a confirmation that the implementation is correct.

Response: We tested the implemented methods extensively and, for example, provide a notebook showing the correspondence of our implementation of the ISIMIP method with the reference implementation. However, it is important to note that the bias on the calibration period is not necessarily zero if the method is implemented correctly: for example, if the parametric fit does not fit the data well or other method-specific assumptions are not met, a correct implementation of the method might result in “residual” or in extreme cases even increased bias over the calibration period.

L232 and Figure 2 – dry days. It is necessary to explain how dry days are handled in the different methods to understand what is happening with “QM”.

Response: We have included some text outlining the dry day treatment of the different methods in the case study (see above).

L237: I do not understand the use of the word “assimilate” here. Please reformulate of explain.

Response: We changed the word, adjusting the sentence as follows:

“When investigating the spatial distribution of the bias (figure 3), we find that certain methods can homogenize the spatial pattern of the bias across climate models.”

Figure 3: This plot would be more efficient with model names only at the left and method names only on the top, and larger panels. If it is a direct output of the software, you can state that as it will justify the less optimal layout.

Response: The plot itself is not a direct output of the software as by operating on a numerical level it does not have information on the geographical position. We have tried adjusting the figure with model names on the left and method names on top, however that actually decreases the possible size of each individual panel. For legibility reasons we therefore opted to keep the current layout.

L260: Please defined the time periods used and if any emission scenario was used. Some measure of signal to noise or significant would be good to include as well.

Response: we added text on the time periods and emission scenarios (see responses above), and made amendments regarding the trend signal to noise question, detailed in the response to the comment above.

L275: Again, please defined the dry day definitions and treatment for each method as it has large impacts on the results, and shall in my opinion not be confused with the general method for the rest of the distribution.

Response: Our response and amendments to the text are covered by our response to the related comments above, as well modifications already made to the text on the basis of these other comments.

Figure6-caption: "change in the number of dry days" right?

Response: We changed the text, it now reads:

"Distribution of location-wise change in the additive climate trend in dry days introduced through the bias adjustment method, computed by computing the additive trend between the validation period and the future period in both the raw and the bias adjusted model and taking the percentage difference between the two trends."

Table A1: last sentence in CDFt "[SSR] can be applied." But is it applied here?

Response: This is applied as default setting; however, the user can choose to deactivate it in which case the full distributions (including dry days) are mapped using the CDFt method. We have changed the sentence to:

"To correct precipitation occurrences in addition to amounts Stochastic Singularity Removal (Vrac et al., 2016) is applied."