Reviewer #2 - 'Comment on egusphere-2023-1481' - October 2nd, 2023

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

Dear Jorn Van de Velde,

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

First, I would like to say that I'm impressed by the paper. Further professionalization and evaluation of bias adjustment is clearly necessary, and the authors make an important step forward by providing this software package. In general, the paper is clearly written and provides good examples and results of the code. However, there are still some major and minor remarks that I would like to see discussed and implemented in the paper.

General comments

Implications of software like this. Further standardizing (or at least standardizing evaluation) becomes clearly possible through this method. This has some consequences. First, it allows for answering questions on the seemingly 'detailed' components of methods, such as the applied time windows (e.g. seasonally vs. 90 days vs. 60 days), number of years for calibration and evaluation, number of data points selected. When implementing a new method, these questions are often sidelined, but they could affect the final result. Not to say that they do, but at least it should be evaluated through standardized tools.

Response: Thank you for this point. In response to both reviewers, we added text in the background section and ibicus description that discusses the modularity of different methods (see also our next comment). We also amended the discussion section in response to both your and Richard Chandler's comments, and, amongst other things, mention that ibicus can be used to explore the consequences of these different choices that might impact the results.

Second, building on one of the comments of Anonymous Referee #1, some method components are not tied to a certain method. This might be considered to be a slightly more philosophical note, but it is possible to consider a switch from methods as 'packages' to methods as 'build from a set of elements'. As elements, I consider e.g. the choice of distribution(s), the choice of dry-day treatment, the order in which steps are taken, additional post-processing steps… Software like this may thus eventually help to disentangle methods and compare their elements (and changes to these elements). Even if they were not originally implemented as such (e.g. a distribution not foreseen by the original author, or an additional post-processing step applied in another software package). According to the documentation, it seems that the way the code is set up, allows (to some extent) for this kind of experimentation.

Response: We thank both reviewers for the constructive comments regarding the modularity of methods. As we also mention in our responses to reviewer #1, this was a question we deliberated on while developing the package. We agree that many components are not tied to a specific method and as you note, we tried to reflect that in the design of ibicus by giving the user options to modify components such as the dry day treatment for some of the methods. However, as we note in the response to reviewer #1, we also retained different core characteristics for the different methods (for example we consider the dry day treatment a core characteristic of some methods such as SDM) to ensure the recognisability of some methods such as ISIMIP, and because some of the choices are entangled with each other which stands in the way of making the package fully modular. In response to the reviewers' comments, we added additional text on the modularity of methods in the background section and description of the ibicus package, extended table 1 to include a high-level overview of dry day treatment, and added an additional table in the appendix that details the treatment of dry days for the individual methods. However, while we attempted to provide a detailed description of all the different modifiable components in the software documentation, we believe that such a detailed description would be beyond the scope of this GMD submission. Based on your comments, we plan to further improve the modularity of the package in future releases and mention this in our amended discussion section.

To conclude, software like this could in time change and influence how we evaluate bias adjustment methods. Could you comment on this and discuss this in your paper? That would certainly further enhance the discussion/conclusions of this paper. Or would even merit a separate discussions section, as Richard Chandler also touches upon this point in comment #3.

Response: We thank all reviewers for highlighting the potential implications of this package as well as possibilities for future methodological development. We significantly amended the discussion section to discuss the implications of the package.

To take the previous point even one step further, it would be relevant to actually review and compare existing software packages. This is seriously out of scope for this paper, but it might be relevant mentioning this need in the discussion/conclusion.

Response: We agree and added a sentence on this in the new discussion section in the paragraph on future development.

Although the authors have taken the time to get acquainted with some of the important discussions in bias adjustment/statistical downscaling literature and touch upon a lot of subjects, I think there is still a lot of ground left to cover. If a reader interested in applying bias adjustment software starts from your paper, it should be possible to track down most of the papers discussing issues and steps forward. So far this is not always possible. In the specific comments, I have given some references related to topics discussed at specific points, which I think are all relevant to refer to in the paper.

Response: We thank Jorn Van de Velde for highlighting some additional literature that is relevant to refer to and his concrete suggestions. We have included a large number of these (see details below).

Additionally to reading the paper, I also did a check of the documentation and tutorials. It seems like a lot of work went into this, for which I would like to congratulate you. Given the amount of information available, I hope a lot of potential users and contributors will find, apply and contribute to your package! However, take note of the changes and additional literature suggested for the paper, and also implement them in the documentation.

Response: Thank you very much for the appreciation of the work that went into the documentation. We will include any modifications made in the paper in the documentation of the package and will also link this paper in the software documentation as background material for users.

Note that 1) I agree with most of the comments of the other reviews (so far posted) and would like to see them addressed properly. Only where really necessary, I repeated a comment. 2) I consider this to be minor revisions, as the software, evaluation set-up and main conclusions are coherent and scientifically sound, but reading the suggested papers might of course take some time.

Detailed comments

L19: it might be good to provide a few examples for the interested reader. See e.g. Vautard et al. (2021) or Galmarini et al. (2019) for relatively recent papers discussing respectively model biases and the impact on agriculture.

Response: We have expanded this point a bit and included a few more examples:

<u>Original text</u>: "Even though climate models have greatly improved in recent decades, simulations of present-day climate still exhibit biases. This means that there are systematic discrepancies between model output and observations that become especially relevant when using the output of climate models for local impact studies, for example by running agricultural or hydrological models."

<u>Modified text</u>: "Even though climate models have greatly improved in recent decades, simulations of present-day climate of both global and regional climate models still exhibit biases (Vautard et al. 2021). This means that there are systematic discrepancies between statistics of the model output and observational distribution (Maraun, 2016). These discrepancies in the two distributions become especially relevant when using the output of climate models for local impact studies that often require focus on specific threshold metrics such as dry days, for example when running hydrological (Hagemann et al. 2011) or crop models (Galmarini et al 2019)."

L22: I would like to stress the comment by AR#1. There are many examples of parametric transfer functions out there.

Response: We have changed 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, but realised based on comments of both reviewers that using the term empirical here is potentially misleading. We had originally wanted to avoid using the term statistical as it sometimes implicitly implies machine learning methods as an alternative.

L24: many multivariate methods as well build on quantile mapping (e.g. by first applying univariate quantile mapping and then a multivariate adjustment procedure, the so-called marginal/dependence multivariate bias adjustment)

Response: This is true, and we have slightly modified the sentence. However, as this is still the introduction, we mostly want to focus on highlighting the breadth of methods, without going into too much detail yet.

<u>Original text</u>: "A variety of statistical bias adjustment methods have been developed and published in recent years, ranging from simple adjustments to the mean, to trend-preserving adjustments by quantile and multivariate methods (Michelangeli et al., 2009; Li et al., 2010; Cannon et al., 25 2015; Vrac and Friederichs, 2015; Maraun, 2016; Switanek et al., 2017; Lange, 2019, and more)."

Modified text: “A variety of statistical bias adjustment methods have been developed and published in recent years, ranging from simple adjustments to the mean, to trend-preserving adjustments by quantile and further multivariate adjustments (Michelangeli et al., 2009; Li et al., 2010; Cannon et al., 25 2015; Vrac and Friederichs, 2015; Maraun, 2016; Switanek et al., 2017; Lange, 2019, and more). “

L25: with regards to multivariate methods, I had to wait until L304 and further to find clarity on why multivariate methods where not implemented here. Although I understand the choice, it should be clear from the start, given the importance of multivariate methods (e.g. in relation to compound events).

Response: We have included the following sentence in the introduction after the sentence modified in the last reviewer comment. However, we kept a detailed discussion in the original position in the paper.

“While this paper focuses primarily on methods that are applied at each grid cell individually, the use of multivariate methods is further discussed in section 5.”

L40 and further: I could nowhere find a clarity on the implementation of the bias adjustment methods. Did you copy-paste them from existing code, implement them yourselves, or mix them? Did you compare results with the original code (whenever available) or contact the original authors to check the original code? Given that small differences in code implementation can have a potentially large impact, this has to be clear from the start (especially in a journal like GMD)

Response: The bias adjustment methods were implemented by the authors using the literature describing the individual methods and possibly available reference implementations in R or Python. Choices were made by the authors of ibicus primarily regarding the question of which aspects of the method to modularize. In case of divergence between the literature description and available reference implementations, the respective authors of the method in question were contacted and after the first alpha release of the software package, all bias adjustment developers were invited to comment on the implementation and review the package. Finally, extensive testing was done to ensure the correctness of outputs and consistency of implementations. We have amended the text in two locations to highlight this:

Introduction:

Original text: “The contribution of ibicus is two-fold: It provides a unique unified interface to apply eight different peer-reviewed and widely used bias adjustment methodologies, including Scaled Distribution Matching (Switanek et al., 2017), CDFt (Michelangeli et al., 2009) and ISIMIP3BASD (Lange, 2019).”

Modified text: “The contribution of ibicus is two-fold: For one, it introduces a unique unified interface to apply eight different peer-reviewed and widely used bias adjustment methodologies. The implemented methods include Scaled Distribution Matching (Switanek et al., 2017), CDFt (Michelangeli et al., 2009), Quantile Delta Mapping (Cannon et al. 2015) and ISIMIP3BASD (Lange, 2019).”

Start of section 3:

Original text: "ibicus implements eight state-of-the-art, peer-reviewed bias adjustment methodologies in a common interface that enables the user to modify aspects of an individual methodology to suit their target variable, region and impact of interest."

Modified text: "ibicus introduces a unified, modular, software architecture within which eight state-of-the-art peer-reviewed and widely used bias adjustment methodologies are implemented. This enables researchers to apply different methods through a common interface, and modify components of the methods, such as the treatment of dry days, based on region and impact of interest. The code implementation of each methodology is based on the cited academic publication, as well as available accompanying code that was re-written and modularised to fit the developed interface. Consistency with the original implementation was ensured through rigorous testing and correspondence with the authors of the different methodologies."

L45: Did you consult Maraun et al. (2015) on the aspect of evaluation and the validation tree? They build heavily on the dimensions you mention here, and follow this up in all papers of the VALUE experiment (see e.g. Maraun et al. (2019)). Although this experiment focuses more heavily on statistical downscaling instead of bias adjustment, the latter is also accounted for and the general principles and lessons should at least be mentioned in a paper on bias adjustment evaluation.

Response: We have included a reference to the VALUE experiment when introducing the evaluation framework in section 3.3:

Original text: "The ibicus evaluation framework offers a collection of tools to identify these issues and compare the performance of different bias adjustment methods for variables of interest."

Modified text: "The ibicus evaluation framework offers a collection of tools to identify these issues and compare the performance of different bias adjustment methods for variables of interest, building on previous efforts such as the VALUE evaluation framework for statistical downscaling (Maraun et al. 2019)."

L50: Here, you apply the standard 'section' titles, whereas further in the paper, you refer to sections as 'chapters'. I prefer the former, as it is more standard.

Response: Thank you for your comment, we changed chapter to section everywhere in the text.

L69: delta change is not limited to linear scaling. It is more correct to consider delta change as a principle of philosophy, where, in contrast to bias adjustment, not the climate model output is adjusted, but historical time series are adjusted. See e.g. Olsson et al. (2009) or Willems and Vrac (2011) for papers building on this principle.

Response: This is a helpful point. We included notes on the modularity of different methods and the fact that they are method families rather than methods at several locations in the text (see response to other comments). With regards to the delta change approach, we have adjusted the text to the following:

Original text: "The most common approaches to the bias adjustment of climate models include a simple adjustment of the mean (Linear Scaling or Delta Change), a mapping of the two entire cumulative distribution functions (Quantile Mapping), or more advanced methods that also aim to preserve the trend projected in the climate model (such as CDFt or ISIMIP3BASD). The practice of using bias adjustment methods to also downscale the climate model has been criticized in various publications (von Storch, 1999; Maraun, 2013; Switanek et al., 2022), therefore this paper focuses on bias adjustment of climate models purely for the purpose of reducing biases at constant resolution."

Modified text: "The most common approaches to the bias adjustment of climate models include a simple adjustment of the mean (Linear Scaling), a mapping of the two entire cumulative distribution functions (Quantile Mapping), or more advanced methods that also aim to preserve the trend projected in the climate model (such as CDFt or ISIMIP3BASD). An alternative approach, often termed Delta Change method, adjusts the historical observations to incorporate the climate model trend (see, for example, Maraun et al. 2016, Olsson et al. 2009 or Willems and Vrac 2011). The practice of using bias adjustment methods to also downscale the climate model has been criticised in various publications (von Storch, 1999; Maraun, 2013; Switanek et al., 2022), therefore this paper focuses on bias adjustment of climate models purely for the purpose of reducing biases at constant resolution."

L90: given the relative importance of trend preservation in your paper and evaluation, I think this concept should be discussed more in-depth. Consider for example Ivanov et al. (2018), which do not entirely seem to agree with Maraun (2016) (which you refer to), Hagemann et al. (2011) or Casanueva et al. (2018).

Response: We thank you for pointing out some limitations of our discussion of trend preservation which we agree should be discussed more in-depth. We have expanded the paragraph in the background section significantly, modified Table 1 slightly and included a number of additional references:

Original text: Furthermore, bias adjustment can modify the climate change trend, in particular, that of threshold-sensitive climate indices (Dosio, 2016; Casanueva et al., 2020). This holds overall for non-trend-preserving methods, as well as for trend-preserving methods if underlying assumptions are not met. Trend modification might be justifiable in specific cases (Boberg and Christensen, 2012; Gobiet et al., 2015), but is not justified as a default practice, therefore requiring a decision on a case-by-case basis.

Modified text: Furthermore, bias adjustment can modify the climate change trend simulated by the model, in particular, that of threshold-sensitive climate indices such as dry days (Dosio, 2016; Casanueva et al., 2020). This holds in general for non-trend-preserving methods but can also be the case for any trend-preserving methods such as ISIMIP3BASD. Reasons for the modification of the trend by 'trend-preserving' methods can be traced to the underlying statistical method and assumptions, such as the specific treatment of values between a variable bound and another threshold, or parametric and non-parametric distribution fits used in different stages of the bias adjustment.

To justify any kind of trend modification by the bias adjustment method, it is necessary to make an assumption about how present-day bias relates to biases in the future period (Christensen et al 2008). This can be based on the assumption that climate model biases are stationary in time (Gobiet et al 2015): for example, based on this assumption, Ivanov et al (2018) developed a theoretical model to justify future trend modifications by the bias adjustment method based on present-day biases. However, Chen et al (2015) and Hui et al (2019), show that while temperature biases can be approximated as stationary, precipitation biases cannot. Similarly, van de Velde et al. (2022) show a clear impact of non-stationarity on bias adjustment, in particular for precipitation. Trend-preserving bias adjustment methods on the other hand assume, at least to some degree, that the raw climate model trend constitutes our best available knowledge for subsequent impact studies. In line with this, Maraun et al 2017 argue that the modification of the trend of a climate model based purely on statistical reasoning is not defendable, and should, rather be based on physical process understanding and reasoning about the large-scale drivers involved.

We have also made a minor change in Table 1:

<u>Original text:</u> Methods such as quantile mapping can modify the trend in the climate model. This might be sensible if the trends are taken to be unrealistic or due to state-dependent biases which need correction (Boberg and Christensen, 2012; Gobiet et al., 2015; Doblas-Reyes et al., 2021). However, in other cases, the trend might be considered credible and should be preserved. Methods can be designed to preserve trends in the mean (DC, LS, dQM), mean and variance (dQM) or all quantiles (CDFt, ECDFM, QDM, ISIMIP3BASD, SDM) - although even then they are not guaranteed to do so. Often trends are distinguished between additive trends (as for temperature) and multiplicative trends (as for precipitation where trends in intensity occur), however not all methods share this distinction. The question of trend preservation is related to the assumption made that the bias is ’stationary’. The assumption is explicitly made by Quantile Mapping. SDM explicitly relaxes the assumption, CDFt and QDM account for it by including a running window over the future period in addition to one over the year.

<u>Modified text:</u> Methods such as quantile mapping can modify the trend in the climate model. This might be sensible if the trends are taken to be unrealistic and related to present-day biases, as discussed in the background section (Boberg and Christensen, 2012; Gobiet et al., 2015; Doblas-Reyes et al., 2021). However, in other cases, the trend might be considered credible and should be preserved. Methods can be designed to preserve trends in the mean (DC, LS, dQM), mean and variance (dQM) or all quantiles (CDFt, ECDFM, QDM, ISIMIP3BASD, SDM) - although even then they are not guaranteed to do so. Often trends are distinguished between additive trends (as for temperature) and multiplicative trends (as for precipitation where trends in intensity occur), however not all methods share this distinction. The question of trend preservation is related to the assumption made that the bias is ’stationary’, as mentioned in the background section. The assumption is explicitly made by Quantile Mapping. SDM explicitly relaxes the assumption, CDFt and QDM account for it by including a running window over the future period in addition to one over the year.

Table 1: the literature concerning the bias stationarity assumption has been growing recently. In the context of evaluation, some of these papers should be referred to explicitly. Consider e.g. Dekens et

al. (2017), Christensen et al. (2008), Chen et al. (2015), Hui et al. (2019), Chen et al. (2020), Wang et al. (2018), Van de Velde et al. (2022) and references therein.

Response: We agree and have added citations to Van de Velde et al. (2022), Maurer et al. (2013), Chen et al. (2015) and Hui et al. (2015) (see response above).

L288: There is a very relevant discussion on the issue of uncertainty in Maraun and Widmann (2018). I think it would be a proper addition to your paper.

Response: Thank you for this comment. We agree that there were points missing from this paragraph discussing uncertainty and have amended the paragraph as follows:

Original text: Figure 8 shows that the climate model ensemble spread of the trend in mean seasonal precipitation is modified when applying bias adjustment. This means that not only the trend but also the range of uncertainty and possible worst-case scenarios analysed in impact studies depend on the bias adjustment method used to pre-process the climate model. As shown in the previous sections, the 'best' bias adjustment method for a given use case depends on the variable, region and impact variable studied. The result shown in figure 8 demonstrates that bias adjustment can add an additional source of uncertainty if the method is applied blindly and not evaluated properly. Interestingly the uncertainty range is not necessarily narrowed as has been postulated by some authors (Ehret et al., 2012), but even extended and shifted in some cases.

Modified text: Figure 8 shows that the climate model ensemble spread of the trend of mean seasonal precipitation is modified in different ways by different bias adjustment methods which is in line with previous findings in the literature (Maraun and Widman 2018, Lafferty et al. 2023). Interestingly the variation (often interpreted as the uncertainty range) is not necessarily narrowed as has been postulated by some authors (Ehret et al., 2012), but even extended and shifted in some cases. From this finding, it follows that the range of uncertainty and possible worst-case scenarios analysed in subsequent impact studies might depend on the bias adjustment method used to pre-process the climate model.

The interpretation of this shift in uncertainty is related to the previously discussed questions on trend preservation, namely whether the change in the climate model trend through a statistical bias adjustment method is justified or not. This issue was mentioned by Maraun and Widmann (2018), who discuss that a minimum requirement to justify a change in the uncertainty spread through bias adjustment should be a critical evaluation of the validity of the results and the assumptions of the underlying statistical model. Given the finding in the previous section, namely that the best bias adjustment method depends on the variable, region and impact variable studied, it follows that indiscriminately applying a bias adjustment method across regions and variables without evaluation can shift the spread of the results of subsequent impact studies in a non-justified manner.

Added citation:

- Lafferty, D.C., Sriver, R.L. Downscaling and bias-correction contribute considerable uncertainty to local climate projections in CMIP6. *npj Clim Atmos Sci* **6**, 158 (2023). https://doi.org/10.1038/s41612-023-00486-0

L311: François et al. (2020) (which you refer to earlier in the paragraph) should also be referenced w.r.t. the difficulties with multivariate methods, as could Van de Velde et al. (2022)

Response: Thank you for mentioning these additional references, we added them to the modified discussion section in the amended document.

Table A1: 1) How were the experimental settings found and defined? Could you give a more expanded explanation? 2) are the references considered to be 'the' references, or just 'standard' references. Especially for linear scaling and delta change (but also quantile mapping), much older references are also available, but are also potentially less clear on the implementation. Please clarify this. 3) Wang and Chen (2014) also further expand on ECDFM and provided the first implementation of the relative version. 4) Cannon et al. (2015) discuss that ECDFM and QDM are practically equivalent. Is this also clear from your evaluation? If not, how come? 5) QDM is at the moment one the commonly applied quantile mapping methods (especially in multivariate methods, see e.g. Mehrotra and Sharma (2016), Nguyen et al. (2016), Cannon (2018)). This could be discussed in function of your evaluation.

Response: 1) The experimental settings were defined based on variable characteristics (bounded, etc), the default settings implemented for these variables in other methods, and manual testing by the package authors. 2) The references indicated are the ones used for the implementation of the method in the ibicus package. We changed the text in the caption of table A1 to specify this. 3) We have added a reference to Wang and Chen (2014) to the references under ECDFM. 4) The "core" mapping of the two methods, QDM and ECDFM are equivalent which is why we describe them jointly in Table A1. We also find that the two methods produce similar results. However, the results are not equal as the two methods use different types of distributions for the parametric CDF fits and QDM in contrast to ECDFM includes a running window over the future period. 5) This was our reasoning to include QDM, and we mention in the main text that we implemented some of the most widely used bias adjustment methods. We did not add an extra comment in the Table.

<u>New caption text</u>: The references given are the references used for the implementation of the method in the ibicus package.

References: please clean up your reference section. There are too many 'book:' and 'publisher:' in there, unless this is the current style adopted by GDM

Response: Thank you for pointing this out. We have improved our reference section.