Response to Referee Comment 1 on 'Impact of bias nonstationarity on the performance of uni- and multivariate bias-adjusting methods' by J. Van de Velde et al.

Bastien François (Referee)

We would like to thank Bastien François for his fair and thorough review. Below, we give a comment-by-comment response.

General Comments

Comment: Results are often not well explained when bad performances for (M)BCs are obtained. The validation metrics used in this study (residual biases relative to the observations (Rbo) and to the model bias (Rbmb)) tend to make, in my opinion, the analyses of results difficult for readers. In that respect, I would find helpful to provide raw values of the indices for, respectively, the calibration and projection periods (i.e. without computing directly their biases as in Table A1) and their changes between the calibration and projection periods for: the reference, the simulations and the 5 corrections from the (M)BC methods. In a general way, it would permit to better explain the results obtained from the (M)BC methods, i.e. whether good or bad results for a specific method come from the nonstationary problem (i.e. from the fact that simulated changes are "wrong", which alter the quality of results from MBC methods) or from the characteristics of the method (e.g., the statistical technique used, its stochastic nature, etc.). Identifying the reasons for good or bad results from (M)BC methods is of key importance for this study, to be able to conclude with more certainty that bias nonstationarity is the main source of problem.

Response: We used the RBo and RBmb metrics as a method to present detailed information on the full extent of changes in a visual way and to enable the reader an easier comparison without needing too many tables. However, these are new metrics and we agree that the raw values are a valuable source of additional information that can provide additional guidance to the reader. However, as referee #2 preferred seasonal or monthly data, providing all raw values and biases would have seriously lengthened the paper. As the RBo and RBmb values can help understand the bias adjustment behavior at some points, we preferred to keep those in the paper and to provide additional information where necessary.

Comment: The authors conclude that "the univariate bias-adjusting methods, computationally less complex and not taking (potentially changing) correlations into account, seem to be more robust." (L787). Concerning correlation, I am afraid that this result of robust-ness is only specific to the present application, and hence the generalization of this conclusion cannot be done as the authors do (e.g., L825 "we advise to use univariate bias-adjusting methods, until it becomes more clear how it can be ensured that multivariate methods certainly perform well in a climate change context."). Indeed, one of the advantages of considering MBC methods instead of univariate ones is that MBCs are able to adjust correlations between variables. Univariate BC methods are not designed to do so. For example, 1D-BC methods such as QDM globally conserve the rank structures of the climate model to correct, and hence the simulated dependence structure between variables is preserved. According to Table A1, simulated correlations often present little bias compared to observations (for example, for corr(P,E), corr(E,T), crosscorr(P,E,0), crosscorr(E,T,0), crosscorr(E,T,1)). Then, by preserving the simulated rank structures, 1d-BC methods like QDM mechanically present correlations with little bias as well. This result is really specific to this study, and would not be obtained if the raw climate simulations would have presented strong biases in correlations. Hence, concerning correlations, results from 1d-BC as QDM depend on how well the models simulate relevant dependencies between the climate variables. This is something already suggested by Zscheischler et al. (2019). This point is not explained in the present study, while it is one of the key points of discussion, and the principal reason why QDM often performs well in this study for

correlations. This point should be mentioned and discussed to provide the appropriate nuances to initial conclusions.

Response: The bias in correlation and the simulation of dependencies is a point that should have been mentioned better. We addressed this in the revised manuscript in the section on correlation results and the Discussion and conclusions section.

Comment: Linked with the comment 1. above, I would find interesting to provide (at least in Appendix) the results for the multivariate bias correction and the hydrological model for the calibration period. In my opinion, verifying that MBC methods perform well for the calibration period (which has to be verified) would validate the global methodology and would better support the conclusions on the effect of bias nonstationarity on the results from MBCs for future periods.

Response: We have calculated the adjustment for all methods in the calibration period. The information will be included in the revised version. This shows, in combination with the seasonal results requested by referee #2 that the original statement that multivariate bias-adjusting methods perform poorer than the univariate bias-adjusting methods was an overstatement. Nonetheless, the comparison illustrates that 1) for some seasons, there is a clear influence of bias nonstationarity on the results and 2) that the multivariate bias-adjusting methods are less well equipped to deal with the seasonal differences in bias nonstationarity.

Comment: As explained by the authors (L79), climate models can present statistical biases compared to observations that are nonstationary, that is, that the differences of bias between the calibration and future period are not the same. What is often not clarified clearly through the article is that, in a changing climate context, another way to see bias nonstationarity is that it results from the fact that observed and simulated variables do not present the same changes between the calibration and projection periods. This is of key importance, as some of the (M)BC methods included in this study can take into account the simulated changes in their correction procedures (such as dOTC or MBCn). It must be better highlighted through the article, and has to be used to provide a better analysis of the results.

Response: This information was added to the introduction of Section 3.3 and considered in the Results and Discussions and conclusions sections where necessary.

Comment: I think the reasons why the three MBCs (MRQNBC, dOTC and MBCn) are selected in this study must be better specified. In particular, their differences of assumptions and attributes could be better indicated, for example with a table (as recommended in 6.).

Response: The introduction for the multivariate bias-adjusting methods was extended and an overview table was added to the end of this discussion. With this table, the reader should be able to get a clearer overview of the differences between the methods.

Comment: If I understood well, all the three MBC methods are able to take into account (some of) the potential simulated changes in their correction procedures. It also exists in the literature MBC methods that assume dependence structure to be stable in time (R2D2, Vrac, 2018, Vrac and Thao, 2020). It would have been interesting to include in the study such MBC methods as benchmarks for a better assessment of the potential losses (or benefits) of considering stability of dependence structure, even in a changing climate context.

Response: The R2D2 method is indeed a valuable method to include in the analysis. We have now included the method. In addition, Section 3 has been extended to give a good overview of this method.

Comment: The quality of presentation should be improved.
- The "Bias-adjusting methods" section is too long. Please consider to summarize each of the BC methods and their main properties in a short text (and with the help of a table, as already explained)

to keep from overwhelming the reader with technical details that are not necessarily useful to get the main points. Technical details and algorithms can eventually be placed in Appendix.
- Please consider to indicate the letters for each figure when describing the results (for instance, for QDM: Figure 3a, mQDM: Figure 3b, etc). This would better guide the reader through the different plots.

Response: In the revised paper, the technical details have been cut from the paper. Adding an Appendix would've increased the length of the paper too much. As mentioned previously, a table was added in the subsection on multivariate methods to better guide the reader. Where needed, the figure panel was specified to better guide the readers.

Specific Comments

Comment: L55: I would replace the word "uncertainty" as it can be misleading here. I would rather say "error (or bias) that can propagate in the impact models".

Response: Thank you for the suggestions, this was updated. (L 52-53 in the revised paper).

If the correlation between these variables is biased w.r.t. the observations, then it can be expected that the model output is biased as well, which can further propagate in the impact models.

Comment: L72: in François et al. (2020), it is explained that all multivariate methods that are under study fail in adjusting the temporal structure of simulated times series. Of course, it already exists methods that can adjust (part of) the temporal properties: MRQNBC is one of them.

Response: The sentence was improved to better reflect this nuance.(L 70 in the revised paper)

Besides, they also noticed that all multivariate methods studied fail in adjusting the temporal structure of a time series.

Comment: L75: what do you mean by "more recent validation period". I think that just mentioning "validation period" is enough.

Response: This was updated in the paper. (L 73 in the revised paper)

Comment: L225: Did you verify that assuming the stationarity of the frequency of dry days holds for your application? Also, did you apply a thresholding after bias correction? I know that, for example, dOTC can generate negative values for precipitation. How did you deal with this problem?

Response: As is shown in the results (R index values) the number of dry days can be considered stationary for the time periods studied. A reference to the first Section of the results was added (L 224 in the revised paper). For dOTC, we applied a thresholding before calculating the indices, as this was suggested in Robin et al. (2019). However, because of space limitations, the technical details are not included in the paper and hence this will also not be mentioned.

However, as thresholding is used prior to all methods, the influence of possible bias nonstationarity on ΔN is assumed to be negligible. Besides, as is shown in Section 4.1 the number of dry days is stationary for the time frames studied in this paper.

Comment: L282: The successive conditional approach performs successive corrections conditionally on the variables already corrected. This can be applied to more than two variables. It should be specified. It does not "adjust a second variable conditionally on the second variable", as written.

Response: This sentence was indeed incorrect. We have clarified that it can be applied to more than two variables. (L 276 in the revised paper)

These two components are then recombined to obtain data that are close to the observations for both marginal and multivariate aspects. The latter approach consists of adjusting a variable conditionally on the variables already adjusted. This procedure is applied successively to each variable.

Comment: L286: This is not well explained. The problem of robustness of the successive conditional approach is specific for a high number of dimensions to correct. Indeed, in a successive conditional approach, as the number of variables already corrected increases at each successive step, it progressively reduces the number of data available for the correction, making it less and less robust.

Response: As successive conditional methods are not studied in this paper, these sentences were removed in the revised paper.

Comment: L396: Do you know why the results are exacerbated?

Response: The first repetition created some unphysical values (especially for precipitation). As the implementation is based on the correlations, every repetition exacerbated these nonphysical values. However, when running the method just once, the results were satisfactory. This was updated in the paper (L 335-340 in the revised paper). The method is not entirely suited for bounded variables such as precipitation and evaporation. Nonetheless, as some of the results were good in both calibration and validation periods, methods based on the same principles are worth investigating and using for precipitation.

However, the nesting method cannot fully remove biases at all time scales, thus Mehrotra et al. (2016) suggested to repeat the complete procedure multiple times. Yet, in our case multiple repetitions exacerbated the results. Nonphysical outliers created by the first repetition influenced the subsequent repetitions, creating even more nonphysical values. This was most clearly seen for precipitation. As a bounded variable, precipitation is most sensitive for nonphysical values. Nonetheless, running the method just once yielded agreeable results.

Comment: L429-440: I am not sure that the equations to explain the Schaake Shuffle are necessary, as it can simply be explained with words.

Response: Not every written explanation of the Schaake Shuffle is equally clear, and we wanted to ensure that readers could follow by providing an exact mathematical formulation. However, these technical details have been removed from the paper to limit the length.

Comment: L442: Do you have a reference to account for ties by introducing small random values? There exists also other ways to compute ranks to handle problems of equal values.

Response: This was based on the approach used in Vandenberghe et al. (2010), in which ties were removed before copulas were fit. Vandenberghe et al. (2010) based this on the discussions in Salvadori and De Michele (2006) and Salvadori et al. (2007). However, these details were removed from the paper, as mentioned earlier.

Comment: L445: This result is really "case-specific". I would precise it, as it cannot be generalized for each application (depending on the bias-nonstationarity, for example).

Response: These sentences were removed from the paper to limit the length.

Comment: L454: Did you check if, even after early stopping, overfitting was not a problem? In particular, how did you choose the tolerance of 0.0001?

Response: We checked overfitting by studying the final distance scores: these show that there is still some margin for improvement. The tolerance was chosen by trial and error: this level of tolerance yielded fairly good results while not increasing the duration too much. This was not added to the revised paper, as the technical details were removed.

Comment: L463: I would precise that dOTC extends the CDFt method to the multivariate case.

Response: This was added to the paper.(L 370-371 in the revised paper)

Comment: L465: dOTC is indeed designed to preserve the trend of the model for marginal properties but also for dependence structure (or copula). It should be specified, as it is a major difference with univariate methods.

Response: This specification is indeed important and was added to the paper. (L 373 in the revised paper)

The combination of both optimal transport plans allows for bias adjustment while preserving the trend of the model for both marginal properties and the dependence structure.

Comment: L520: I don't think that the 10 calculations made for dOTC are necessary. What are the bin sizes chosen to implement dOTC? If the bin sizes are small, the influence of the stochastic components in dOTC is rather small, and hence introducing 10 calculations does not seem necessary to me.

Response: For each variable, 25 bins were used. More bins quickly increased the computational load. The bin size depends on the variable space. The results differed among the calculations, and thus it is better to keep them and average the results.

Comment: L593: I do not understand the sentence "Yet, QDM has the best RBMB values and might thus be preferable", that seems in opposition with the previous sentence.

Response: The results were rewritten based on seasonal results (a request of Referee #2) and this sentence was removed.

Comment: L644: "an influence" on what? Correlations? Please be more precise.

Response: The results were rewritten based on seasonal results (a request of Referee #2) and this sentence was removed.

Comment: L677: Actually, it surprises me as it has been found in François et al. (2020) that, for a small number of dimensions to correct, MBCn and dOTC preserve roughly the rank sequences of the model to correct. It normally does not have such "important shuffling" as specified. Do you know why?

Response: This should have been written differently. As can be seen in Figure 7, the RBmb and RBo values for the lag-1 autocorrelation and wet-dry transition probability (panel (c), MBCn) and the number of dry days (panel (e), dOTC) are close to 1, indicating that they are very close to the original climate simulation values. Thus, the methods indeed roughly preserve the rank sequence. As the results were rewritten based on the seasonal results, this sentence was removed. However, ensured that the writing is clearer on this specific aspect of the method.

Comment: L698: A difference between QDM and MBCn is also the adjustment of dependence structure. It should be specified.

Response: Thank you for addressing this point. Although the specific sentence was removed, this point was considered when revising the Results section.

Comment: L705-706: I do not understand. Please rewrite.

Response: The results were rewritten based on seasonal results (a request of Referee #2) and this sentence was removed.

Comment: Did you mean "possibly"?

Response: Yes, thank you for noting this mistake.

References

Salvadori, G. and De Michele, C.: Statistical characterization of temporal structure of storms, Advances in Water Resources, 29, 827–842, https://doi.org/10.1016/j.advwatres.2005.07.013, 2006

Salvadori, G., De Michele, C., Kottegoda, N. T., and Rosso, R.: Extremes in nature: an approach using copulas, vol. 56, Springer Science &Business Media, 2007

Vandenberghe, S., Verhoest, N. E. C., and De Baets, B.: Fitting bivariate copulas to the dependence structure between storm characteristics: A detailed analysis based on 105 year 10 min rainfall, Water resources research, 46, https://doi.org/10.1029/2009WR007857, 2010

Response to Referee Comment 2 on 'Impact of bias nonstationarity on the performance of uni- and multivariate bias-adjusting methods' by J. Van de Velde et al.

Anonymous referee

We would like to thank the referee for the time spent reviewing our manuscript. Below, we give a comment-by-comment response.

General Comments

Comment: The authors did not explicitly present the study area of this paper. After looking for a while, I'm surprised to find that only one station and one grid cell were used (lines 115-143). In addition, there is only one GCM-RCM combination was used, and no information on its spatial resolution. As such, the results of this study are subjected to large uncertainty.

Response: We believe our research design is still informative, as it is too date scarcely addressed and can provide a base-line for more extended follow-up studies. Yet we simultaneously agree on the limitations of our design, and we have now properly stated this in both the Introduction and the Discussion.

Comment: The manuscript is lengthy and hard to follow. I think some information does not need to be presented in detailed in the manuscript. For example, the introduction of the bias correction methods (section 3.1 to 3.3, almost 14 pages of the main body). All these information can be found in literatures.

Response: We have removed the technical information and other minor unnecessary parts of the paper, such as the information on climate change in Uccle. Besides, we integrated both 'Discussions' and 'Conclusions' sections in one 'Discussion and conclusions' section

Comment: Although the authors have stated in line 527 "As the effect on discharge is the overarching goal of this paper", I think the information on the hydrological model and hydrological simulation this quite poor in the manuscript. Firstly, I did not find how the hydrological model performed in the study area (e.g. Nash value or some other criteria). Although the goal of this study is to compare the difference between the univariate method and multivariate method in the hydrology perspective, the hydrological model should be well calibrated. Secondly, as is stated that the PDM model was not calibrated in Uccle but Grote Nete watershed, please give the evidence to show that it is feasible to drive the PDM using the climate data in Uccle.

Response: Some details on the calibration performance were added. Besides, the Grote Nete watershed is roughly 50 km from the Uccle station and the effect of topography on weather is negligible. Hence, the assumption that the Uccle data can be used for the watershed is acceptable, as discussed in the referenced and other earlier papers. (L. 185-196 in the revised paper)

Comment: The authors used scatter plots throughout the whole paper. It is difficult to see how many dots are located within the 0-1 square or to do the comparison (e.g. Figure 6). I suggest that maybe the authors can use some more quantitative metrics and figures to show the results.

Response: The scatter plots allowed for the inclusion of multiple indices or interesting percentiles of the variables' distribution in one plot and are hence practical for conveying information. However, we understand that using only the RBmb and RBo values might impede some readers in quickly grasping the results. As such, we have include other measures, such as Perkins Skill Score and ETCCDI indices, wherever applicable. The inclusion of these measures allowed to give a more balanced overview of the results.

Comment: Figure 3 to figure 9, the authors did not mention which period (calibration or validation) and which month the results are based, or is it based on the whole year? In this case, readers cannot understand and assess the results. For example, for the correlations between two variables (e.g. P and T), the correlation coefficients are quite different for each month (e.g. summer and winter), therefore, at least, they should be evaluated separately for each month.

Response: These figures were based on the validation period, with data for the whole year. All figures are updated to clarify the time frame. Besides, as suggested, we have calculated monthly and seasonal data. This indeed allows for a better assessment of the inter-seasonal differences, which provides important insights in the propagation of bias nonstationarity. To provide a balanced overview of the results and keep a proper length of the paper, we preferred the seasonal evaluation. Hence, all results are rewritten to reflect the updated validation.

Comment: Table 1, I am wondering whether the Spearman correlation coefficients and lag-0 cross-correlation between two variables reflect the same thing.

Response: The Spearman correlation is based on the rank of the value, the crosscorrelation is based on its effective value; these two values thus do not reflect the same thing. This can also be seen in Figure 8 of the revised paper, where the RBmb and RBo values for the Spearman correlation and the always differ by some extent.

Comment: In general, I found that many expressions in the current manuscript are not very accurate. For example, in line 568 the authors write “A surprising result for P is the high RBMB value for P99.5 for MRQNBC”, but I found in figure 3 that the corresponding value for MRQNBC is quite small. Therefore, I’m quite confused by many expressions in the current manuscript.

Response: This was an incorrect statement and we thank the referee for pointing this out. After revising the manuscript, we made sure to properly proofread the manuscript.