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

In the second draft, the authors of "Impact of bias nonstationarity on the performance of uni- and multivariate bias-adjusting methods'' took into account part of my comments, as well as comments from the second reviewer. In particular, it results in evaluating the performance of 2 univariate and 4 multivariate BC methods under climate change conditions in order to determine the influence of bias nonstationarity on results. Instead of evaluating their results over the whole year (as in the first draft), the authors followed the advice from the second reviewer and performed a seasonal evaluation. They conclude that non-stationarity can have an important influence on the performance of the MBC methods and the propagation of biases in impact models. The authors found that the importance of the influence varies depending on seasons and variables. They finally advise to account for seasonality for a robust bias adjustment under bias-nonstationarity, and advice to use univariate methods instead of multivariate ones until it becomes more clear how MBCs perform under bias nonstationarity.

While I appreciate the work done by the authors to modify the initial draft and take into account the comments from the reviewers, I think that several major limitations still remain in the present study and can be improved.

General comments:

1) The design experiment does not permit to assess properly the influence of non-stationarity on the performance of the MBC methods and should be modified. If I understood well, the 2 univariate BC (QDM and mQDM) are applied over a 91-day moving window, while the 3 multivariate BC (MBCn, R2D2 and dOTC) are applied over the full time period. Applying the 3 multivariate BC (MBCn, R2D2 and dOTC) over the full time period and evaluating them at a seasonal scale is not appropriate, and hence presents a major issue for the interpretation of the results. Indeed, as pointed out by the authors in Section 4.1, biases can vary considerably depending on the season: for example, a climate model can present very little bias in winter and be drastically biased in summer. Thus, the correction of the statistical properties of the model (such as mean, variance, correlations) can be very different depending on the season. By applying MBC methods such as MBCn, R2D2 and dOTC that do not include seasonal components in their procedure over the full time period, it generates data that potentially present bias introduced by the design experiment: the MBC methods would correct the different seasons by applying a similar statistical transformation. Consequently, model data for seasons with strong biases won't be corrected enough, whereas model data for seasons with little biases could be deteriorated. In the study, applying these 3 methods (MBCn, R2D2 and dOTC) as such potentially introduces a bias by construction, placing them at a disadvantage in the intercomparison study compared to QDM and mQDM. Moreover, it makes it impossible to identify if bad performances from these MBCs are due to either bias nonstationarity or artefacts from the design experiment. This problem is known by the authors, as discussed in L638 (« It is thus unclear whether the poor seasonal performance obfuscates the effect of nonstationarity, or if the similar performance is a sign of robustness. ») and L640 (« Hence, the set-up does not allow to clearly discern between the various categories of multivariate bias-adjustment, such as the 'marginal/dependence' or 'all-in-one' categories. »). But assessing the effect of nonstationarity on the performance of MBCs is initially the main objective of the study. The design experiment must be established in order to isolate the effect of nonstationarity as much as possible. For example, it can be done by applying all the BC methods that do not include seasonal components in their correction procedure (QDM, mQDM,

dOTC, R2D2 and MBCn) over the same seasonal period (e.g. over winter or summer, but separately), and then performing seasonal evaluations to fairly compare them. It would permit the intercomparison of BC methods, all things being equal.

Response: We have updated the methods to work with seasonal input. This has, as suggested, an impact on the results. Consequently, the Results and Discussions and Conclusions sections of the text were rewritten to reflect these changed results.

2) The new implementation of a MBC method, named Rank Resampling for Distributions and Dependences (R2D2, Vrac et Thao, 2020), is unclear. This MBC method relies on an analogue-based technique for which some conditioning information is required to adjust dependence structure of the simulated time series. The conditioning information can be multivariate, by considering a set of variables at a given time t. It can also be extended to ranks sequences, i.e. conditioning by not only one but several lagged time steps. The choice of the conditioning information is crucial to interpret the results from R2D2, as it can have impacts on marginal, inter-variable and/or temporal properties. This information, and its influence on bias-corrected data from R2D2, is not precisely given in the paper. Consequently, results from this MBC method cannot be analyzed in an appropriate way by the readers. Moreover, at L362 is indicated: « Each variable (precipitation, evaporation and temperature) was in turn used as the reference dimension. ». This implies that 3 bias-corrected data were produced for R2D2, but, surprisingly, only one result for R2D2 is presented in the study. Thus, further clarification is required to better present the results from the R2D2 method.

Response: The information on R2D2 has been extended to be clearer and to better reflect our application.

In the present application of R2D2, QDM was used as the univariate bias-adjusting method to ensure consistency with the other multivariate bias-adjusting methods. This ensures the preservation of the changes in the marginal distribution. Each variable (precipitation, evaporation and temperature) was in turn used as the reference dimension. As the present study was limited to a single grid cell, the use of additional data was limited. However, to ensure that the selection of analogues is diverse enough, five lags were used to search for analogues, three of which were retained in the resampling. Finally, the results for the three variables were averaged to present the final R2D2 result.

3) The Section 4.1 'Bias change' is hard to read. Is a table missing? The authors describe index values for bias change between the calibration and projection period, but a table seems necessary to better present the results and facilitate reading.

Response: A table has been added.

4) I would like to thank the authors for providing the results for the calibration period. However, linked with my first comment, it highlights that MBCn, dOTC and R2D2 methods are not applied in an appropriate manner compared to QDM and mQDM: For example in Table 4, PSS values indicate poor performances for these 3 MBC methods during the calibration period, principally because they are applied over the full time period and evaluated by seasons. If MBC methods do not produce good results on the calibration period on indices that are supposed to be adjusted, then MBC methods are not well "calibrated" and no good results can be expected for the validation period. This point should be considered if my first comment is taken into account.

Response: Applying the multivariate methods on a seasonal basis has clearly improved the PSS results, and this is now reflected in the text. In case the PSS value is not similar among the methods, this is taken into account in the discussion of the results.

5) Also, linked with 4), the advantages of the "marginal/dependence" methods such as MBCn is that, for evaluation criteria on marginal properties such as PSS in Table 4, same performances must be obtained between QDM and MBCn (trend preservation), by construction. It would be nice to consider retrieving these results on marginal properties before analysing other indices, such as correlation or discharge.

   Response: As described for 4), the performance of the multivariate methods, and hence MBCn, improved with seasonal inputs. The PSS values for MBCn are now (for the marginal properties) the same as those for QDM, indicating that these methods can be compared for other indices.

Specific comments:

6) L318: In this study, dOTC is not the most recent method used, but R2D2 2.0

   Response: This has been adjusted. We now refer in this sentence to dOTC as 'the last method' (as mentioned in the paper).

7) L526 « Both the univariate and the multivariate bias-adjusting methods can adjust the simulated biases well » and L530 « the good adjustment by univariate methods is trivial: they will adopt the correlation of the simulations and only slightly adjust this by adjusting the marginals. » I was wondering if you can rephrase these sentences in order to avoid saying that univariate bias-adjusting methods adjust correlations. Improvements of correlations are only due to an indirect effect of the adjustments of marginal properties.

   Response: The first sentence has been removed, while the second was slightly rephrased.

8) L527: « The univariate methods will adopt the dependence structure of the raw simulations » I am not sure if it is true for mQDM, that will have exactly the same rank correlation structure than the observations by construction (at least for the calibration period).

   Response: This has been adjusted to refer to only QDM.

9) Table 2: As requested, a table is introduced in order to summarize the different characteristics of the MBC methods. This table can be very useful for the readers, but the actual one presents some formulations that are not clear enough or misleading. Some examples: for the row « Temporal properties » and column « dOTC », the information « Future, adjusted » is misleading. dOTC is not designed to adjust temporal properties and must be clearly indicated. Another formulation must be used instead to add more nuances and to specify that potential unexpected behaviors of temporal properties can be obtained with dOTC. For the column « R2D2 », the information « Shuffle based on observations » is not clear enough: temporal properties of the bias-corrected data depend on the conditioning information used (see my second point). For the column «MBCn», the information « Shuffle based on observations » is wrong: temporal properties from the model are modified in an uncontrolled manner by the decorrelation/recorrelation procedure and the univariate correction. However, empirical findings in François et al., 2020 indicate that MBCn (and hence the decorrelation/recorrelation procedure) tends to conserve partially the rank sequences from the model, in particular in the context of bias-correction of a small number of statistical dimensions. Moreover, it might be necessary to change the order of the rows in order of importance. I find it odd

to have the row « Temporal properties » at the beginning of the table and « Statistical technique » almost at the end of the table.

Response: The table has been adjusted based on your comments. The information was reordered and some of the characteristics were updated to better reflect the specifications of the methods. In addition, some of these points are now better clarified in the main text.

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

This manuscript presents an attempt by the authors to investigate the effect of non-stationary biases may have on the performance of multivariate bias-adjusting methods for regional climate models. To do so they have used four multivariate methods (MBCa, MRQNBC, R_2D_2, dOTC) and two univariate ones (QDM, mQDM), to adjust bias in the output from a single GCM-RCM run (12.5km RCA4 forced with boundary conditions from the MPI-ESM-LR global model) for three climate variables (temperature, precipitation, and evaporation). Precipitation and evaporation have been used to drive a rainfall-runoff model as a test of the impact that bias adjustment can have on variables used for impact studies. The authors conclude that non-stationary biases are important for bias adjustment procedures without reaching a firm conclusion on the issue of the relative performance of uni-and multivariate bias adjustment. I believe this is the result of poor methodological choices by the authors, which constrained their ability to reach a more meaningful conclusion in what is an interesting and relevant topic. More specific comments regarding the methods follow:

1) All the bias adjustment calibrations and validations were carried out using model output from a single model cell, of a single RCM-GCM combination, with a single observed dataset as reference. Therefore, the results presented in the paper may be unrepresentative of the behaviour of the bias adjusting methods, which could be explored much more robustly by exploring multiple locations and models. The data choices give the authors a single comparison point for the bias correction methods, a larger sample would help reduce the uncertainty present in the results and possibly lead to stronger conclusions.

Response: Although the data is indeed limited, we believe that it is still possible to derive useful results. Considering the comments of Referee #1, the results clearly illustrate how biases can be influenced by bias nonstationarity, and how this can propagate to an impact model. Yet, we acknowledge that this is only a case study and would like to present our study as such. Nevertheless, this provides a framework that other studies could expand upon. In addition, we have expanded the discussion to compare the climate models used here with other models from the EURO-CORDEX ensemble. This provides only a limited overview, but still allows other researchers to place this study within a broader context.

2) The results are unclear; they lack clear trends and demonstrate the limitation of the single-location approach as each seasonal index is plotted once and therefore no conclusion can be drawn as to how representative the results really are. While the indices used are useful in representing different portions of the distribution of each variable, no statistical evaluation of the significance of the differences was attempted, either through the use of summary statistics or graphically.

Response: It was our goal to indicate the significance using the RB_MB and RB_O metrics. Whereas RB_MB indicates the effectiveness of the bias adjustment, RB_O is an indicator of the significance in comparison with the observations. However, the explanation of these metrics was not clear enough and has now been extended. In addition, we have paid more attention to this throughout the text. Also, the comments of

referee #1 helped to clarify the results. Finally, the goal of this manuscript is not to clarify trends, but to give an overview of how bias-adjusting methods might respond to bias nonstationarity.

3) The emphasis given to hydrological models in the abstract is lost throughout the paper. Very little detail is provided about the rainfall-runoff model in the methods, in particular there is a single RMSE value as sole evidence of model calibration with the reader referred to a previous paper for details despite this being a key aspect of the paper. In addition, the data point used to evaluate the bias methods lies outside the model catchment and no evidence is provided to support the similarities between the sites.

Response: As we wanted to study the sensitivity of an impact model to biases and bias nonstationarity, small discussions of the (possible) impact of bias nonstationarity on the marginal properties, correlation and occurrence are now included in each Results subsection. This allows for a better understanding of the propagation of the bias nonstationarity and the role of the hydrological model. However, as indicated in the abstract, the hydrological model serves only as an illustration and the main goal is understanding the sensitivities of an impact model to bias nonstationarity. As such, adding more information on the calibration does not seem necessary, as it is not a key aspect for us. The issue of the distance between the climate model grid cell and the catchment was already discussed in the text, albeit limited.