### Review of Manuscript hess-2018-151

### 'Cross-validation of bias-corrected climate simulations is misleading'

### by D. Maraun and M. Widmann

Dear Editor, dear Authors,

I have reviewed the aforementioned work. My conclusions and comments are as follows:

## 1. Scope

The article is within the scope of HESS.

## 2. Summary

The authors start by giving a short overview on the scope and methods of bias correction (BC) as used in the field of climate model simulations. In this context they discuss the usage of cross validation to evaluate bias correction and distinguish the two cases 'perfect predictor setting' (where the boundary conditions of the problem are known) and 'climate change application' (where the free running model is not bounded by known boundary conditions). The authors argue that for the latter case, random realisations of long-term climate variability can render a classical split-sample cross-validation approach inapt to evaluate a BC method.

In section 3, the authors analyze a simple holdout cross-validation experiment for both an additive and multiplicative BC method: The BC value (factor) is determined in one partition of the available data and applied to the remaining data. The authors demonstrate that the residual bias in the validation data set, which is typically used as a measure of BC effectiveness, is sensitive to both the relative magnitudes of the simulated **and** observed change signals (changes between calibration and validation time). In short, climate variability between the calibration and validation period can lead to false conclusions about the effectiveness of a BC method.

In section 4, these findings are demonstrated at an exaggerated modelling example.

The authors conclude that cross validation should not be used when evaluating BC methods in free running climate simulations.

## 3. Overall ranking

The topic the authors tackle is highly relevant. The conclusions are convincingly supported by the analytical derivations shown in section 3. Therefore I think this is an important contribution to climate change research.

Nevertheless, I suggest the following **major** revisions:

- The analytical derivations in section 3 are straightforward, convincing and sufficient to support the author's arguments. I found the additional example in section 4 hard to understand. Also, as the example is constructed in an exaggerated manner to make the authors points clearer, I was not sure to which degree the conclusions based on this extreme case are transferable to 'normal' cases. Therefore I suggest deleting section 4 altogether. This will make the paper much clearer. If the authors wish to present an example, they can refer to the recent, excellent paper by the same authors (Maraun et al. 2017) where a similar example is presented. If section 4 is deleted, I further suggest to change the article from 'Research Article' to 'Technical Note', as it will be both short and dealing with a single, specific, technical question, which is what Technical Notes are meant for.
- I welcome the authors' discussion throughout the paper about when cross-validation approaches are valid and when not. What would be really helpful for the reader in the paper would be a short summary of the general problem by naming its main components and their interrelations (maybe

with a drawing) and strategies to decide whether cross-validation is appropriate for a given problem-setting or not. This could also serve as a sketch of the framework the authors mention in the last sentence of the conclusions. An incomplete list of the components:

- Length of available observational records [time]
- For a given observable (e.g. rainfall):
    - For a chosen spatial aggregation (e.g. 30x30 km grid): What is the aggregation time until (stationary) temporal variability has become statistically insignificant? [time]
    - For a chosen temporal aggregation (e.g. 1 year): What is the spatial aggregation until (stationary) spatial variability has become statistically insignificant [space]
- Existence and intensity of instationarities (trends): E.g. expressed by the time until a statistically significant change between the first and second half of a trend-afflicted time-series is detectable [time]
- Envisaged time of extrapolation beyond the observed period [time]
- By comparing the resulting timescales, it would be e.g. be possible to analyze:
    - With the available observational records, at which spatial/temporal aggregation can we separate instationarity from variability? This also puts a limit to the resolution of extrapolations and also allows to decide whether cross-validation makes sense or not.
- All of these components have already been mentioned by the authors in this or other papers, what I suggest here is to put them together in a compressed manner to frame the problem as a starting point for solution strategies.

- A last point: On page 6 line 5, the authors dismiss a discussion about the validity of BC as beyond the scope of the paper. However, when reading the paper, the natural question arising was 'If we do not have valid and agreed-upon methods to evaluate the effectiveness/appropriateness of a BC method in the observation period in the first place, how can we evaluate the validity of BC methods in the context of extrapolation, which is an even more involved problem?' So while I agree with the authors that an exhaustive discussion of this matter is impossible and beyond the scope of the paper, it should definitely be addressed here.

Yours sincerely,

Uwe Ehret

**References**

Maraun, D., G. Shepherd, T., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J., Hagemann, S., Richter, I., Soares, P. M. M., Hall, A., and Mearns, L.: Towards process-informed bias correction of climate change simulations, nclimate3418 pp., 2017.