Response to Reviewer Comments

We would like to thank the reviewers for their helpful and supportive comments. Please find our point-by-point response below.

Reviewer 1 (Seth McGinnis)

> *I can make only one comment of any substance, which is that I think it is a very slight exaggeration to say (page 5, lines 10-11, and again on page 7 line 10) that the result of cross-validation is *purely* random and says *nothing* about the sensibility. That would be the case if the difference between simulated and observed changes were caused entirely by internal variability and not merely dominated by it. Strictly speaking, the result of the cross-validation is *almost entirely* random, and says *vanishingly little* about the sensibility of the cross-validation.*

We agree with the reviewer. Therefore we have adjusted the text as suggested (we used mostly instead of almost entirely though).

> *the shapes of the red symbols in Figure 1 are difficult to make out at their current size. It might be beneficial to make the symbols somewhat larger or to give them a border in a contrasting color (e.g., black) to make it clearer which symbol is which.*

We have added a border and slightly increased the size.

> *Page 1, line 7: move "also" after "is".*

We have kept the also, as the emphasis is slightly different (both versions are grammatically correct).

> *Page 2, line 9: move "also" after "may".*

Same as in the case before.

> *Page 2, line 21: change "has been" to "was", remove "already".*

Changed.

> *Page 2, line 29: move "optimally" after "use the data".*

Changed.

> *Page 3, line 1: "lead time" is two words.*

Changed.

> *Page 4, line 19: "residual" is misspelled.*

Changed.

> *Page 5, line 13: typo after "time-scales".*

Changed.

> *Page 7, line 6: replace hyphens with em-dashes.*

Changed.

Reviewer 2 (Uwe Ehret)

> *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.*

Regarding the possible deletion of section 4, we strongly support the reasoning of reviewer 1 to keep the example as is. The reason is twofold:
1. we believe that the analytical derivation might be helpful for some readers, whereas others may prefer a strong illustrative example. This holds in particular as - to our experience - the role of internal variability in climate research is still often underestimated. The analytical derivation might then be dismissed as being purely academic reasoning without practical relevance.
2. choosing the Maraun et al. (2017) example would not suffice. In fact, the reason for this short article was that - during the review process of the Maraun et al. (2017) paper we realised that the situation was even worse than laid out in that paper. There, the key starting point for the discussion was that cross validation may not be able to identify a nonsense bias correction. This is the false negative case in the manuscript at hand. Here we show additionally that there is another problematic case: that a sensible bias correction may be rejected (false positive). This case is not contained in the Maraun et al. (2017) example. Of course, it makes sense to furthermore show the true positive and true negative cases as well.
We also do not believe that the exaggerated examples are limited in applicability/transferability to more realistic cases: the true negative case is a realistic case where a well performing climate model is successfully bias corrected - here, the case is not at all exaggerated. Similarly the false positive case, where the bias correction is sensible, but the residual bias does not vanish, is far from exaggerated: this is exactly the case we would like to highlight with this paper.

The other two cases are chosen to display wrong applications of bias correction in a convincing case. To avoid any discussion about the sensibility of bias correction in one or the other situation, we decided to take examples where anybody would agree. In fact, the false negative case - the correction does not make sense, but it is not rejected - is of a very similar character as the example chosen in Maraun et al. (2017). One may actually argue that the correction of temperature against precipitation from two different regions in that paper is even more exaggerated than the example here (where the same variables are chosen, but different locations). In a real application, of course, the problem will not be as obvious as constructed in our examples. Here, the user of BC has to carefully assess whether the bias correction makes sense at all. This discussion, however, is not the main focus of our manuscript.

Therefore we kept the examples in the text. We added, however, some explanations and brief discussions on the transferability of these examples to real applications (page 6, line 11-15; page 7, line 5-8).

> *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.*

We believe a specific discussion of the contexts in which cross validation may make sense (including a figure) would go well beyond the scope of our manuscript. As indicated, this even depends on the way cross validation is carried out (in the "statisticians way" or in the "atmospheric

scientists way"), with several subtleties. A thorough discussion could easily distract the reader from our main point. We added, however, a short and rather general discussion of the issues influencing the sensibility of a cross validation. This discussion covers all the issues raised by the reviewer, but in a general way only (page 8, line 13-22).

> *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.*

The reviewer is of course right that this is an entirely open question. We have had a discussion if this issue in the Maraun et al. (2017) paper, where we highlight the fact that BC has to be accompanied by a thorough evaluation of non-corrected features (in particular temporal and spatial), by a process-based evaluation of the underlying climate model (in terms of location biases, relevant feedbacks etc.), and by reasoning about representativeness and trend modifications. In the conclusions, we had a very short discussion of this topic. We have slightly extended this discussion to accommodate for the reviewer's comment (page 9).