Summary:

The authors present a bias correction method for GCMs that relies on a quantile mapping approach and is based on the methods of Piani et al. (2010) and Hearter et al. (2011). Additionally, it preserves absolute rends in temperature and relative trends in other variables. The paper gives a very detailed technical description of the bias correction procedure and presents some evaluation results.

Major comments:

1) My major methodological point has already been raised by Reviewer 1 and John Smith: A major caveat of this study is that the authors only evaluate their bias correction method using the same calibration and evaluation period. In such a setup it is easy to achieve nearly perfect evaluation results regarding the distribution. However, since the bias correction is applied to future climate simulations, some kind of cross-validation, split-sample test, or any other evaluation that demonstrates how your method performs if calibration and evaluation periods are not the same has to be added. I want to add, that the current evaluation (same calibration and evaluation period) is in this specific case still important and shouldnt be removed from the study, since it clearly demonstrates substantial technical issues of the authors implementation of bias correction (see next comment).

A split sample sensitivity test is now added to the paper while the focus is still on the results based on the 40 year training period. Please also see our reply to John Smith.

We used the full 40 year sample for the calibration to avoid as much of a dependence on low frequency internal variability in the observed and simulated data set as possible. In the given context of bias correction we consider a split sample exercise as a sensitivity test regarding these internal variabilities (or trends): If the sample was split up and the bias correction was trained on the first part of the whole data set and tested on the other one, possible deviations between bias corrected simulation data and the observational data set would be due to the fact that mean values and variances do not change simultaneously in the observations and the simulated data set when switching from the training to the test period. These might e.g. be due to some internal variability on longer time scales or differences in forced trends. The bias correction should be as independent as possible of differences in the phases of internal variabilities. GCM experiments are not designed to provide agreement with observations regarding the timing of these internal processes. The dependency is decreased by using the longest calibration period that is available e.g. 40 years in our case. Therefore the results based on the 40 year calibration represent the core of our study. We have added the split sample exercise as a sensitivity analysis to the paper. It shows that in most regions the sensitivity to the calibration period is negligible.

2) A quite disturbing feature of the presented bias correction method is its weak performance. After bias correction you find, e.g., remaining deviations of more than 5K in q10-q50 ranges and remaining errors of comparable size in other evaluations. If you consider that in your evaluation setup (calibration period = evaluation period) a proper mapping would trivially lead to perfect evaluation results, these errors are very large. You correctly state that they are caused by the parametric fit (with only one parameter in case of temperature and 3 parameters in case of precipitation). Knowing that, why do you stick to this approach? Dont you expect to achieve better results by other (e.g. non-parametric) methods?. Please refer to, e.g., Gudmundsson et al. [2012] for an evaluation of different implementations of quantile mapping. If you stick to your approach, you have to discuss it compared to other published approaches and to argue why your approach is suitable for your application.

We agree with the reviewer that non-parametric methods may yield a better mapping of the distribution in the calibration phase. However, non-parametric fits are problematic when applied to future climate projections as they do not allow for extrapolation. Moreover, Gudmundsson et al. [2012] noted that the " success of the nonparametric transformations is likely related to their flexibility [but for these] methods with many degrees of freedom, over fitting may be a concern [and] it cannot be ruled out that the methods perform badly if the projected climatic conditions differ substantially from the calibration period." Moreover, most of the statistical approaches that are currently available, including methods based on parametric quantile mapping, are in principle capable to remove biases in simulation data (as Gudmundsson et al. also noted in their conclusions). In line with this finding, we argue that despite the imperfect performance for present-day conditions, a parametric approach still improves the representation of the historical climate substantially and is preferable for applications like ISI-MIP that rely on a wide range of future climate projections.

3) Now a quite fundamental issue: The trend-preserving nature of the presented method is achieved by mapping anomalies instead of absolute values. As a consequence, a certain correction value is not attached to a certain absolute value of the corrected variable, as it is the case in "traditional" quantile mapping. One usually argues that a climate model can be expected to have a typical temperature error at, say, -10°C daily mean temperature and another (potentially different) typical error at +25°C. Those temperatures are obviously related to different weather situations and it can be expected that models feature typical errors related to each weather situation. This concept is of course a wild simplification, but at least it roughly explains why quantile mapping can be successful when it is applied to future simulations. However, in your application, a specific correction value is attached only to a temperature anomaly, i.e. to different absolute temperatures and consequently to different weather situations. Why

should the same correction value be appropriate for different weather situations? How can you argue that? This problem is particularly severe when you apply your method to future simulations, where similar anomalies can be related to quite different absolute values. I cant judge how severe this problem is, but since it is a fundamental conceptual issue, it should be analyzed and discussed in your study.

We thank the reviewer for this important hint. Generally the variations of the monthly mean temperatures are not expected to be that large as our correction is done on a month specific basis. However, there might remain a dependence of the daily variances on the monthly mean values we do not account for by our correction method. To quantify this effect we have plotted the monthly mean values against the variances of the associated daily data (for the bias-corrected and for the observed data) and used a linear model to describe the dependence. The following text was added to the manuscript to highlight this issue:

"In the following section we present a method to adjust the daily variability of the residual temperature and the normalised precipitation data. Note that this means that a specific correction value will refer to an anomaly. Thus, a given correction value for temperature might be related for example to different absolute temperatures and consequently to different weather situations with potential systematic differences in variability. Generally the variations of the monthly mean temperatures are not expected to be very large as the correction is done on a month specific basis. Nevertheless, a small dependence of the daily variances on the monthly mean values remains which is neglected at this point. Assuming a linear dependence slopes different from zero are obtained particularly for temperature in some mid- and high-latitude regions (cf. Supplement Fig. 1). The related R-squared values, however, suggest that there is no significant linear dependence between mean and variance for temperature in most cases."

4) The troubles described above follow from the aim to preserve absolute or relative trends. As you correctly note, it is currently not so clear whether preserving trends desirable or not (P51, L25, Ehret et al. [2012]). I want to add that there is strong indication from Christensen et al. [2008] and Boberg and Christensen [2012], and even stronger indication from unpublished work, that certain bias correction methods (including quantile mapping) may modify trends in a way that can be regarded as improvement. You dont argue clear enough why the trends and relative trends should be preserved in your application. Your arguments on P54, at least in their present form, are not convincing. Please elaborate on why it is so important for impact models to have consistent temperature trends over land and over sea. And if so, why is it then acceptable to have inconsistent trends (= consistent relative trends) in all other variables

than temperature? Otherwise, also taking into account the arguments in the previous paragraphs, I dont see how the presented method is a step forward, compared to what has already been published.

First of all we like to clarify that we do not claim that the conservation of the trend is a general goal for all impact modeling applications, but it is desirable within the framework of the ISI-MIP and possible follow-up projects or similar applications.

In particular a modification of the (local and) global temperature trend would in the end mean that the climate sensitivity of the model is modified which we believe cannot be justified based on a 40 year observational data series. Moreover, in ISI-MIP the impacts of climate change and the related uncertainty shall be quantified for different levels of global warming. This issue is of particular relevance e.g. for decision makers wishing to better quantify possible consequences of specific temperature targets. Since the observational data set, and thus also the bias-corrected data set, only includes temperature over land it is usually not consistent with the global mean temperature derived from the uncorrected simulation data. Hence, a method that preserves the absolute trend in temperature is desired for such kind of applications.

Generally, also regarding the other variables we are in favor of this transparent approach that provides some control over the GCM features that are preserved. The choice to preserve the relative trend rather than the absolute one for other climate variables is due to the positivity constraints of these variables. For this reasons other methods also applied correction factors instead of additive constants to correct precipitation data (for example Ines and Hansen 2006). For some applications e.g the one discussed in Boberg and Christensen 2012 it might be desirable to modify the trend of the GCM, however, this introduces a new level of uncertainty at the larger timescales and is thus problematic for impact assessments like the one intended by ISI-MIP.

5) You provide a bias correction method for quite a bunch of meteorological variables, but evaluate only temperature and precipitation. Please also show results for the other variables, at least as supplementary material. This is very important, since your data are used by impact modelers in ISI-MIP and plenty of impact-publications will be based on it. This requires a complete evaluation in order to enable the subsequent studies to rely on a well described basis.

We thank the reviewer for this suggestion, and have added maps of the longterm mean and interquantile ranges also for the other variables in the Supplements (Supplement Fig. 6).

Considering these points, I cannot suggest the paper to be published in its present form, but it should be considered for publication after major revisions. Specific comments: Abstract: Results (e.g. your main evaluation results) are missing in the abstract. Abstract: Please streamline. E.g., you mention twice that you present a trend preserving bias correction method once is sufficient. You could, e.g., apply the following structure: Describe all introductory information (general information about bias correction and ISI-MIP) in the first paragraph, followed by a paragraph describing the method (its trend preserving nature and all information that is currently in the last paragraph), and finish with results (which are so far missing).

We are sorry that we missed to state the main evaluation results in the abstract. An paragraph on that is added in the amended version of the manuscript. We also improved the structure of the abstract.

P51, L14: Bias correction via scaling (applying a "multiplicative constant") does obviously not conserve trends, as you describe in sect. 3.1. Please correct the sentence.

We thank the reviewer for noting this inconsistency. We rewrote the text to clarify the statement.

P52, L13: You refer to bias correcftion also as downscaling tool. However, the downscaling ability of most bias correction methods (including quantile mapping) has limitations, particularly regarding variability, as described by Maraun [2013]. Please mention these limitations (see also comment by D. Maraun).

We agree that the ability of bias correction to downscale data is very limited. Downscaling is not the main intent of the proposed methodology, as also noted in the reply to D. Maraun. We rewrote the introductory part of our manuscript and included Supplement Figs. 4 and 5 (mentioned in the evaluation section) to point out limitations of ISI-MIP with regard to this point.

P52, L19: What do you mean with "more detailed altitude-stratified" apart from what you already said in the previous paragraph? Isnt that exactly what downscaling does? Please clarify.

We like to point out here, that in the first paragraph we refer to the adjustment at the horizontal scale, while the second refers to the vertical scale. However, we agree that both issues are not independent and are associated with downscaling.

P54, L5, Why exactly is it essential to ensure consistency between global mean temperature change and bias corrected temperature change (i.e. to preserve the trend)? And why is it the essential conserve relative trends in the other meteorological elements? Please argue in more detail. The trend conservation is the key development of your method and it should be clearly explained why it is advantageous or necessary. (See also general comments)

As argued above, the consistency between the uncorrected and bias-corrected global mean temperature is not a purely scientific issue, but allows to address the society relevant question which impacts can be expected at which level of global warming. Temperature is the key variable that frames the public discussion and political decision-making on climate change. Furthermore, also regarding the other variables the long-term trend is the key signal provided by the GCM. Hence, it is desired to preserve this quantity in a precisely defined way (e.g., in relative terms as offered by ISIMIP) in order to keep the range of climate projections comparable to the original GCM simulations. Our choice to preserve the relative instead of the absolute trend for the variables – except temperature – has pragmatic reasons (positivity constraints) as described above.

Chapter 2: Please add a sub-section for the model data (as you did for the observations).

We thank the reviewer for this suggestion and added another subsection in section 2.

P59, L19: Again the issue of variability. Bias correction in the form presented here cannot improve the temporal structure of the time series. You might get a the variance closer to the observations due to "blowing up" the time series (effect of bias), but the sequence of e.g. precipitation events is not improved. See Maraun (2013) and the comment of D. Maraun for much sharper arguments regarding this topic.

We agree that bias-correction does in general not improve the temporal structure of the time series. This might be possible by applying a sequence of corrections at different time scales (as described in the conclusions and future work section), but it is not the main intent of the proposed methodology.

P60, L10: "we adjust the residual distribution of the GCM to that of the WFD using a parametric quantile mapping (cf. Eq. 7)." It is unclear why you refer to Eq. 7 here. It doesnt describe quantile mapping.

The reference to Eq. 7 was includes as it describes the residuals (not the quantile mapping). We agree that the formulation might be misleading here and rewrote this sentence.

P60, L10: "Since temperature is well described by a normal distribution, a linear fit is sufficient.” Please explain clearer, that you mean a linear fit to the transfer function. The current formulation is a bit confusing. P60, L11: You state: “Since temperature is well described by a normal distribution, a linear fit is sufficient.” As you show later in your study, the linear fit causes quite some error. It is therefore inappropriate to call it sufficient (see also general comments).

We thank the reviewer for pointing this out and changed the formulation accordingly. “In general tempature values are considered to follow a normal distribution. This means the distribution is expected to be well described by only two moments (mean and standard deviaton). For that reason a linear fit is considered an appropriate approximation in most cases and has thus been chosen to map the simulated to the observational temperature values.” Furthermore, please note that the errors mentioned later in the manuscript are not caused by the linear fit, but are deviations in higher moments of the simulated and observed temperature distributions which cannot be corrected by the linear transfer function.

P70, L15, Eq. 29: It has been already discussed in section 3 why the method preserves temperature trends. There is no need to repeat that here. P71: Same for precipitation

We agree with the reviewer and rewrote this paragraph.

Figures 7,8,9, and 10: For easier reading, please clearly specify in the figure captions what exactly you show. You do this mostly in the text of the paper, but it would be helpful for the reader also having it in the figure caption. In particular, you should clearly state what you mean with “deviation” (I assume you mean: model observation) (fig8), or with “differences in trend” (fig 7) (I assume it is the difference between the trend of the corrected and the uncorrected model).

We thank the reviewer for pointing this out and rewrote the figure captions.

Figure 8 and its discussion: After bias correction you find remaining deviations of more than 5K in case of temperature. This is large, if you consider that in your evaluation setup (calibration period = evaluation period) a proper mapping would trivially lead to perfect evaluation results. You correctly state that the remaining deviations from the observations are caused your parametric fit (with only one parameter in case of temperature and 3 parameters in case of precipitation) of the transfer function. Knowing that, why do you stick to this approach? (See also general comments).

We decided to stick to the selected parametric fits, since they have been proven a useful description of the functional dependence be-

tween GCM and WFD distribution in the WaterMIP studies. Moreover, a parametric approach allows extrapolation to values which haven't been observed so far and can be expected to be rather robust (e.g., with regards to outliers) because of the low number of free parameters.

Figure 9 and its discussion: You show quite large differences between the basic and the extended version. What consequences does this have for the impact modelers? Please clearly quantify the total errors in the data that has been delivered to ISI-MIP (not only the differences to the enhanced version).

Figure 9 (bottom panels) illustrates that differences exist mainly in some tropical regions. The upper panels show the total values of the interpercentile ranges for the basic and the extended version. As discussed in the manuscript, in those regions where larger differences occur the variability is closer to that of the GCM than suggested by the observational data during the reference period. This means that in the tropical regions assessments of extreme events must consider larger uncertainties.

Equations in general: The paper contains a large amount of equations (30), but some of them hardly contain helpful information in addition to what is already very clear from the text, or are more or less duplicates (e.g. Eq. 3 and Eq. 22). Please consider removing some of the not so important equations.

We thank the reviewer for his comment. We believe that the equations are necessary to support the understanding of the main algorithmic steps. However, we rewrote some equations to increase readability.

References:
Boberg, F. and J. H. Christensen (2012), Overestimation of Mediterranean summer temperature projections due to model deficiencies, Nature Climate Change, 2(6), 433- 436, doi:10.1038/NCLIMATE1454.
Christensen, J. H., F. Boberg, O. B. Christensen, and P. Lucas-Picher (2008), On the need for bias correction of regional climate change projections of temperature and precipitation, Geophys. Res. Lett., 35(20), L20709, doi:10.1029/2008GL035694.
Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., and Liebert, J.: HESS Opinions Should we apply bias correction to global and regional climate model data?, Hydrol. Earth Syst. Sci., 16, 33913404, doi:10.5194/hess-16-3391-2012, 2012. 50, 51, 53. Gudmundsson, L., Bremnes, J. B., Haugen, J. E., and Engen Skaugen, T.: Technical Note: Downscaling RCM precipitation to the station scale using quantile mapping a comparison of methods, Hydrol. Earth Syst. Sci. Discuss., 9, 6185-6201, doi:10.5194/hessd-9-6185-2012, 2012.
Maraun, D., Bias correction, quantile mapping and downscaling: revisiting the inflation issue. J. Climate, Online First, 2013.