Authors present a new bias correction methodology for daily precipitation and temperature data, based on an already applied to the ISI-MIP methodology. The proposed methodology belongs to the parametric transformation methods. The manuscript is generally presented in a clear and understandable way. It is also nice to use a single timeseries to present the correction procedure clearly. However, there are still some major points of criticism.

1) On the initial submission, it was clear that the major caveat of the methodology was the lack of cross validation. Anonymous Referee #1 (AR#1) pointed out this problem, asking for a major revision. Authors claim that 40 years of observations are not enough to calibrate/validate their methodology. It is generally true that the larger the observational dataset, the higher the certainty level of the calibration. However, the fact that 40 years of data are not enough is not supported by the literature (e.g. the split sample calibration – validation scheme has been tested successfully in Piani et al, (2010), using 10 years' time slices – this is half the available data). Upon AR#1 request authors agreed to the need of a calibration-validation experiment (in fact they made an honest effort to add a calibration validation period). Having said that, one should expect that the entire evaluation of the methodology would be based on a calibration validation context. More specifically, the comparison to the WATCH methodology should be done by applying the WATCH methodology to the same periods of calibration-validation.

Thank you very much for these valuable comments! We are sorry for the confusion we may have created with regard to the feasibility of a split sample calibration given a 40 years data set. We hope we can clarify our position with regard to a split sample calibration:

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): When the sample is split up and the bias correction is 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 second period of the observational data set may be expected due to internal variability on longer (e.g. decadal) time scales not being represented entirely realistically in the global climate model, or to 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 general the

sensitivity to the calibration period is small. We compare our results to the WATCH data to highlight that a conservation of the long term trend is not generally ensured within other bias correction methods. To underline this major difference between both approaches we used the same calibration period as originally used in the WATCH project for the WaterMIP exercise.

2) Authors should elaborate on their results further. The difference between the past and future trends of Fig. 7 and the remaining bias of Fig. 8 is not an adequate metric for the method evaluation. Assuming that authors will add the validation section, they should provide comparisons for every aspects of P and T they elaborate with, i.e. mean, standard deviation, number of wet days between corrected GCM P and T and the WFD P and T.

We are sorry that there seems to be a misunderstanding here. In Fig. 7, we do not show the difference between the past and future trends. What is shown is the difference between the trend in the uncorrected and corrected data, where the trend is defined via the comparison of the mean values over a past and a future period. This is to illustrate the basic feature of the new method, which is to preserve the trend.

In Fig. 8, we show the remaining bias over the reference period in order to demonstrate limitations of the bias correction, as the distributions of observational and bias corrected data should, by construction, match over this time span. With regards to the statistical properties we think that showing the anomaly (difference to the observational data) before and after the bias correction captures the crucial information and is a more readily interpretable illustration than showing the mean of corrected and observational data separately. Moreover, we decided to show the interpercentile ranges rather than the standard deviation as they provide not only information on the width but also on the skewness of the distribution.

Moreover, in the amended version of the manuscript we add a map of the number of dry days.

3) After the drizzle day correction, (page 63), the dry days' precipitation is evenly distributed in the wet days. This is a "fine tune" that serves for the need to keep the mean precipitation between WFD and corrected GCM consistent, but alters the climate signal in a fairly arbitrary way that is not supported adequately on arguments.

We agree that redistributing the dry day precipitation evenly to the wet days is somewhat arbitrary, but in our special case it is preferable to a simple truncation of the drizzle as it is common in other approaches (e.g., described by Piani et al. in 2010). Setting the low precipitation values to zero without accounting for the lost amount

would modify the monthly mean and, thus, the climate signal on monthly and longer timescales. As our approach is explicitly designed to preserve the long-term trend, we decided for a pragmatic redistribution of the amount of water. The chosen additive approach also preserves the variability of the daily precipitation intensity as it shifts the whole intensity distribution for wet days. We consider the uniform distribution for the redistribution of the drizzle as the most basic choice.

4) In Figure 9, 10, the lower and upper 10%-ile results are not presented. Especially the upper percentile of the daily precipitation is of great importance, since it carries a great proportion of the total precipitation. In many semi-arid areas, the upper 10 percentile may carry half the precipitation that the area receives. Finally the extreme precipitation, also included in the upper 10 percentile, is also an aspect that should not be ignored.

The analysis of precipitation extremes (or events in the outer range of the distribution) in the bias corrected data is, of course, an crucial topic. We do not show percentiles here at all, but only use the interpercentile ranges as a measure to evaluate the adjustment of the width and the skewness of the distribution. Nevertheless, we agree that the probleme of extreme events must be addressed at some point. Therefore, we added the following paragraph:

"We focus in our sensitivity study on the range between between the 10% and the 90% quantile. For this central range bias correction methods are expected to perform well, while the correction in the outer ranges of the distribution is typically worse, since there are less events (Maraun 2013). In general bias correction methods tend to exaggerate extreme events, since the limited number of data points prohibits a robust analysis of the relationship between observations and simulations, potentially resulting in an overestimation of these events. In addition, the extreme events always cover the whole gridbox area, i.e. their spatial extent is typically too large. However, since we introduced an upper bound for the bias-corrected values in ISIMIP, the impact of this effect is not arbitrarily large. On the global scale the bias-corrected variables show good agreement with the observational data even in the tails of the distribution (cf. Supplement Fig. 4)."

5) An arbitrary upper bound of daily precipitation of 400mm/day is introduced to avoid single extremes. How this threshold was chosen? The WFD dataset has a maximum daily precipitation value of 724 mm/day and several daily values over 400mm/day. This suggests that the 400mm/day threshold not only limits the observational dataset itself, but also the probable "new extremes" of the future corrected precipitation, creating the false assumption that the maximum daily precipitation cannot exceed this threshold.

The threshold was chosen for pragmatic reasons as it turned out that several state-of-the-art impact models have problems to process very large values of daily precipitation. As the rare extremes of this magnitude turned out to provide problems with regard to the impact simulations, we decided to set the cut-off value for the daily precipitation to 400 mm/day. This truncation, however, affects only few grid cells and less than 1% of the days. On average an amount of approximately 1000 mm/year is globally lost by truncation, i.e. summed up for all 67420 grid boxes less than 3 mm/day are truncated.

6) If authors will to compare their results to another bias correction technique's results, it is of course very welcome. However it would be better to compare results for the entire year, not just a calendar month (April).

The bias correction is done separately for each month. Thus, we decided to consider single months for the comparison. Results for the other months are very similar and therefore omitted. The comparison to the WATCH approach was mainly done to highlight the fact that the trend is not necessarily preserved in the original method. That also looks similar for the other months and the April data are shown as an illustration of the effect.

Minor points:

Page 64 – lines 1 to 5: The correction procedure is subject to constrains. This means that there are grid cells that were not corrected at all. This should be defined in some way in Figure 7, 9, 10.

Please excuse the confusion. All grid cells were indeed corrected. We distinguish three cases: (i) A correction of longterm mean and nonlinear transfer function for the variability correction, (ii) a correction of longterm mean and linear transfer function for the variability correction, and (iii) only a correction of the longterm mean. Following the very helpful hint global maps showing which approach is applied at which grid point have been included into the Supplement (Supplement Fig. 2). In addition, the maps indicate regions where the correction factor for the longterm mean was truncated – this happened exclusively in regions where the correction was already restricted to a correction of the longterm mean.

Page 58 – line 14: The threshold of 10 for the multiplicative factor is defined vaguely.

This is a pragmatic choice. As noted above we included a Figure in the Supplement which shows that this rule is only applied in re-

gions which are considerably dry in the particular month.

Page 58 – line 19-20: Authors state that a possible reason that multiplicative factor C can take unrealistic values is that the assumption that "model and observations are well described by the same type of distribution (e.g. gamma distribution) does not hold". How did this conclusion occur?

Sorry for the confusion. Of course in our approach the gamma distribution is only relevant for the daily data while the correction factor is applied to the monthly data. This was meant rather as a general statement that wrong model assumptions may affect a parametric bias correction method as the one used for ISIMIP. We agree that the sentence is misleading and deleted it.

Page 61 – line 1 to 5: Authors state that gamma function well approximates the GCM and WFD precipitation, and that gamma function is not defined at zero. I could not understand why this is the reason that frequency and intensity should be separately corrected.

The actual distribution of the daily precipitation is described by some point mass at zero (probability of dry days) and a continuous distribution of precipitation intensities > 0 that have been shown to be well approximated by a gamma distribution that is only defined for values > 0. Thus the actual distribution can be described as a mixture of the 0-1 distribution describing the probability that a day is wet (1) or dry (0) and the gamma distribution describing the intensity of the precipitation at wet days. Both are corrected separately as described in the text.

Page 64 – line 4: The cut-off value of 80 wet days is stated to be motivated by sensitivity studies performed in WaterMIP. A proper references needed though.

This choice was based on personal communications and some initial tests with cut-off values of 40, 80 and 160 wet days, which did not change the maps of the chosen fitting type (nonlinear, linear or only monthly).

Figure 8, 9: the precipitation values should be presented in a more common unit i.e mm/day, mm/year.

We thank the reviewer for these helpful comment. We modified the illustration accordingly.