Responses to Reviewer Comments on "Bias correction can modify climate model-simulated precipitation changes without adverse affect on the ensemble mean" by Maurer, E.P. and D.W. Pierce, Hydrol. Earth Syst. Sci. Discuss., 10, 11585-11611, 2013 www.hydrol-earth-syst-sci-discuss.net/10/11585/2013/doi:10.5194/hessd-10-11585-2013

We are grateful to the four anonymous reviewers for their careful consideration of this manuscript and for their helpful and insightful comments. Below we state how we address each comment in a revised manuscript. Original comments are in black; our responses are indented and in red type.

Anonymous Referee #1

Recently, problems have been detected in the use of quantile mapping for climate change simulations. In particular, several authors have shown that quantile mapping affects GCM trends. The authors of this paper address the question, whether the mapping actually deteriorates the change signal compared to observations. As such the topic is highly relevant. Also the authors show nicely the effect of quantile mapping on trends in their synthetic example. Yet I am still concerned about the setup of the study and the conclusions drawn.

Major comment: The authors compare AOGCM simulations (as far as I understand these are really coupled simulations, not driven with observed SST) for the US with observed data over two historical time periods and assess the effects of QM calibrated in the first period and then applied on the second period. These time periods are each 30 years long. It is well known that the climate of the US is strongly influenced by internal modes of climate variability, such as the PDO and the AMO. For instance, the AMO has a period of roughly 60 years and strongly controls the amount of precipitation over the US (e.g, Knight et al., GRL, 2006). The amplitude of this internal mode of variability is of the same order of magnitude than the observed climate change signal. See, e.g, Deser et al., Nat Clim Change, 2012, for the influence of internal variability on temperature in the US, a variable which has a much better signal to noise ratio. Thus the observed trend is only partly (if at all) a forced trend. As the GCMs are run in climate mode, their realisation of the AMO is not synchronised with the observed AMO, i.e., the 30 year long ups and downs will almost certainly not coincide with the observed ones. This has two important consequences:

1. the observed differences between observations and models are not biases, but a superposition of biases and random differences due to long term modes of climate variability.

2. there is no reason why the modelled trends should match the observed trends. The forced trends of course should, but not the overall trends which are a superposition of forced trends and random fluctuations. The defined index were a useful index if the forced signal were isolated, i.e., without internal climate variability. But in the current setting, it is not a useful measure. A perfect climate model might have a very bad index model, just because the realisation of random climate variability is out of phase with the observations, and a bad climate model could in principle have a good index value

because a bad forced trend superimposed by an out of phase realisation of climate model variability might by chance produce the observed trend.

Note again, that this is not an academic problem. Internal variability is a major source of uncertainty of precipitation projections and makes up about 30-50% of the total uncertainty even on time horizons of 60 years on a continental level (Hawkins and Sutton, Clim. Dynam., 2011). This problem has already been discussed in Maraun et al., Rev. Geophys., 2010 - the authors should be aware of it. In fact, they observe this problem for the simulations of the East Coast trends, where they found opposite trends in observations and half the models (p 11593, l 25).

I am not sure what conclusions should be drawn from this point. One which definitely has to be made is that GCM biases cannot easily be calculated and thus also not easily be removed (apart from the fact that bias correction in general works locally, but GCM circulation biases are non local, for a discussion see Eden et al., J Climate, 2012). My recommendation would be that the authors repeat the analysis with AMIP type simulations (i.e., atmosperhic models forced with observed SST to synchronise long term internal climate variability) or even better to use RCMs or nudged GCMs to avoid the erroneous correction of GCM circulation biases. But I see the point that the author's want to "correct" (coupled) GCM biases to finally provide bias corrected future simulations. Yet, again, GCM bias correction is not a simple task (and the fact that hundreds of studies have been published based on such corrections is not necessarily an indicator of quality). So far it has not been shown that GCM bias correction works in principle, it has just been applied. As the main point of the paper is about effects of quantile mapping, a compromise could be to point out all the problems listed above with proper references, and tune down the conclusions.

> Response 1.1: We are very appreciative of this valuable and detailed comment. This highlighted some shortcomings of the manuscript, which we have revised to address this issue. While specific responses and modifications to the paper are included in more detail below, in general our revisions have addressed this in two ways: 1) as suggested above we replace the original analysis (based on unconstrained historical GCM runs) with a new analysis using an ensemble of AMIP model output, to apply this process to a set of model runs in which the natural variability is more closely tied to observations; 2) the introduction and interpretation of results are much more clear about how quantile mapping used as a bias correction is blind to the sources of the 'bias' and attempts to correct both differences due to natural variability and systematic errors in a forced model response equally and we discuss the implications of this for applications to future projections.
>
> The new AMIP-based experiment is first mentioned in the Introduction: ".. this study uses model output contributed as part of the Atmospheric Model Intercomparison Project (AMIP) experiment. In these AMIP model runs the simulated natural variability is more closely tied to observations, since observed sea surface temperatures and sea ice are imposed on the atmospheric model, with the same greenhouse gas concentrations as the historical simulations. This provides a test where the effects of unsynchronized low frequency natural variability

between the models is diminished relative to unconstrained historic runs. The improved representation of trends in AMIP-simulated precipitation, as compared to unconstrained historical runs, has been demonstrated (Hoerling et al., 2010)." The two historical periods have been replaced by two periods from the AMIP simulations (which begin in 1979), as mentioned in the revised Methods section, fourth paragraph: "The period 1979-1993 is used to train the QM, which is then applied to 1994-2005. The difference in precipitation between 1994-2005 and 1979-1993 is assessed both before and after bias correction." All Figures have been re-created to reflect the new results with this ensemble and the different periods. As a final result, the new Figure 9 is shown here:

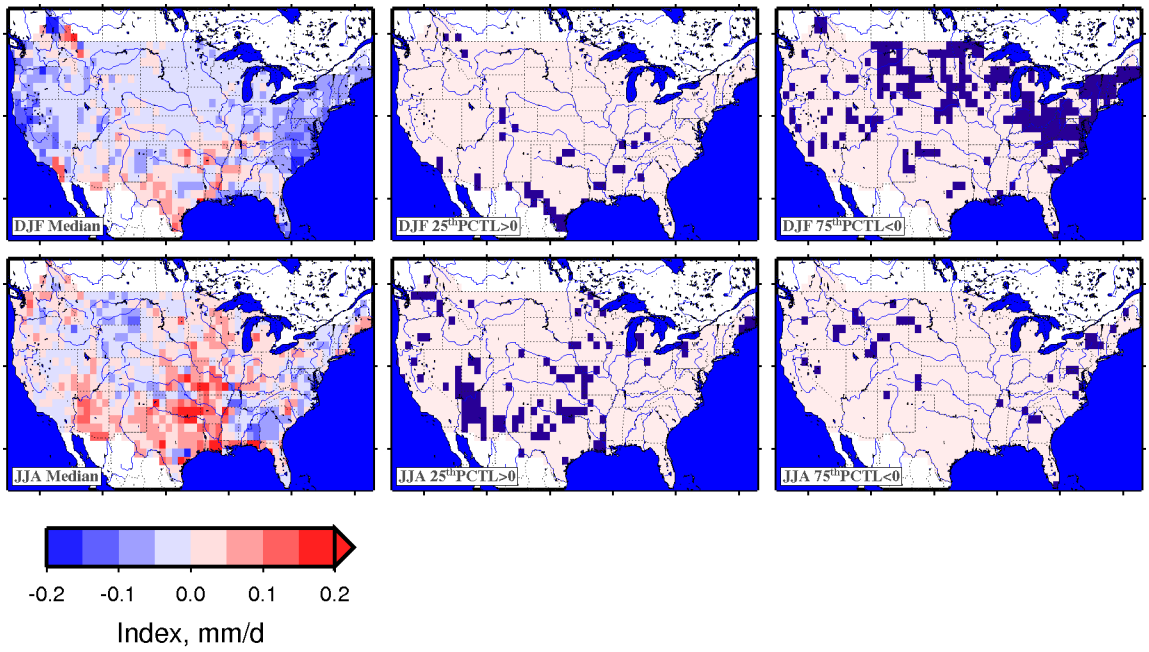

Figure 1 - For DJF and JJA, the ensemble median TM index value (left panels), the locations of grid cells (dark rectangles) where the 25th percentile TM index value exceeds 0 (center panels), and the grid cells where the 75th percentile value is less than 0.

The revised results change in details, but not dramatically in the overall findings: "The median TM values (left panels) tend to lie close to zero, and neither degraded (TM>0) or improved (TM<0) values dominate the picture for either DJF or JJA. The center panels highlight regions where 75% of the GCMs show a degraded change in precipitation (relative to the observed change) due to the BC process. These cases constitute 4.3% of the grid cells for DJF and 13.0% of the grid cells for JJA. The right panels show the grid cells where 75% of the GCMs show improved correspondence with the observed change after BC. These cover 26.2% and 4.5% of the domain for DJF and JJA, respectively."

Furthermore, the revised manuscript should be much more clear that our focus is on the impact of the tendency of the bias correction to change trends, and whether that results in a systematic

change, for better or worse, in the correspondence of the bias corrected precipitation changes to observed changes. As the proportion of the variance due to forced versus internal variability changes in the future, these conclusions may need to be revisited.

To temper the conclusions to align more precisely with what was found, the abstract now ends with the following statement: "While not representative of a future where natural precipitation variability is much smaller than that due to external forcing, these results suggest that at least for the next several decades the influence of quantile mapping on trends does not degrade projected differences."

The following points definitely need to be mentioned:

-biases are systematic differences in the physics of a model, i.e., in forced signals, not random realisations.

-it is difficult to estimate GCM biases in presence of internal modes of variability such as the AMO.

Response 1.2: These first two points are now discussed in the last two paragraphs of the introduction of the revised manuscript, most of which has been added to make this point clear.

-GCM bias correction is therefore also difficult. Here the East coast example might be shown.

Response 1.3: We do not assess the effectiveness of GCM bias correction in this paper. The example in the Results section where Figure 6 is discussed (with the new AMIP-based results, it is no longer the East Coast that is discussed but the Pacific Northwest -- much of the domain has trends that align better with observed trends, presumably due to using AMIP simulations), shows only that trends between two periods do not correspond with observed trends in many GCM runs. The bias correction is not intended to have any effect on that. This has been clarified in the text of the revised paper, specifically where Figure 6 is discussed, we added "It should be emphasized that the BC only adjusts the quantiles of the GCM to match those of observations within a 15-year training period -- there is no attempt to match trends, either within the 15-year training period or over longer periods. Thus, any trends are inherited directly from the GCM, though the QM can, as discussed above modify these."

-one should really define which biases to correct, see Eden et al, J Climate, 2012 (note that they call internal climate variability errors, which is at least misleading; they mean uncertainty; personal communication with the authors).

Response 1.4: In the revised paper we discuss the different sources of variability. We do not, however, attempt to separate the different sources. The end of the Introduction section now includes the statements: "While historic GCM simulations include the climatic response to forcings such as changes in atmospheric greenhouse gas concentrations, solar variability, etc., they are unsynchronized with historic natural variability (Eden et al., 2012). This natural, or internal, variability of precipitation can be dominant even at time scales as long as 50 years

(Deser et al., 2012; Maraun et al., 2010), and may play a substantial role in GCM variability in future projections through the mid-21st century (Hawkins and Sutton, 2011). Thus, the differences in a regional precipitation change between two periods in a GCM historic simulation compared to the observed change result from both GCM biases in sensitivity to external forcing and the fact that natural variability is not synchronized with the observed record. Only the former represents a bias in the GCM. To lessen this effect, this study uses model output contributed as part of the Atmospheric Model Intercomparison Project (AMIP) experiment."

The revised Introduction continues with the following: "In this study we do not attempt to separate the different sources of variability, applying a QM bias correction as it is typically done, where the QM recognizes the difference between a simulated and observed variable (calling it 'bias'), but is blind to the source of the difference. As the sources of this aggregate 'bias' change in the future, for example, when the precipitation trends forced by increased atmospheric greenhouse gas concentrations dominate regional precipitation variability, it is conceivable that the effect of QM on the GCM trends may change. It is also possible that the relative importance of different mechanisms driving regional precipitation (e.g., large-scale circulation, orographic enhancement, convective storms) will change in the future (Cloke et al., 2013; Maraun et al., 2010), altering the GCM biases and ultimately the effect of QM on trends. Thus, the findings from this experiment should be limited to the historic period and the next few decades, when natural precipitation variability constitutes a similar proportion of the variability as over the most recent three decades.

It should also be emphasized that this study does not examine the effectiveness of QM at reducing differences between observed and GCM simulated precipitation, but only its effect on mean precipitation changes over multi-decadal time scales. This experiment examines whether there are coherent changes to the simulated precipitation changes induced by QM, and if so, whether they might have a tendency to improve or degrade the projected changes."

-in particular bias correction works locally (e.g., convective parameterisation errors could be corrected), but cannot shift, e.g., the storm tracks.

-currently it has not been shown that GCM bias correction really works (as it has been shown for RCM bias correction, e.g, Maraun, GRL, 2012).

Response 1.5: These last two major comments also pertain to the effectiveness of QM to remove biases in some projected period. This study was not designed to assess that, but was not clear enough in stating this. In addition to the prior responses, which should help clarify this, the abstract now includes the statement "The effectiveness of the bias correction is not assessed, only its effect on precipitation differences." The first paragraph of the conclusions now includes: "It is emphasized here that this study includes no assessment of the effectiveness of quantile mapping at reducing biases, but only its effect on precipitation trends."

Minor comments:

in the title it should be effect, not affect.

Response 1.6: Corrected.

page 11586, line 26ff: "in any downscaling...". This statement is wrong. Most statistical downscaling approaches are perfect prog, i.e., by construction they do not correct GCM biases. Please state this!

Response 1.7: The second sentence of the introduction now includes a statement contrasting perfect prog and MOS in this regard. In particular, the revised text states "While "perfect-prognosis" downscaling estimates fine scale projections by assuming the predictors are realistically-simulated (Eden et al., 2012), any "model output statistics" (MOS, Glahn and Lowry, 1972) approach by design includes some form of bias correction to remove the time-invariant GCM biases, allowing the signal, or change, simulated by the GCM to be isolated to some degree from the systematic errors. This is critical in applications such as hydrology, where runoff is a non-linear function of precipitation, and so highly sensitive to model biases."

page 11595, line 26ff: please rewrite the following five sentences. They all start with "we". This is tiring.

Response 1.8: This has been changed.

page 11588, line 27: this does not hold for rare extremes. There, parametric distributions are needed to constrain the mapping. This is, however, difficult to validate because of the rareness. Please add "moderate"

Response 1.9: This was a result from the cited reference (Gudmundsson et al.), not our own claim. In any case, it is not essential to the discussion and we have removed the phrase "for both means and extremes."

Eq 2: use a different name than just "index". It carries no information!

Response 1.10: It is now called the trend modification (TM) index.

page 11591, l 1ff: this effect has been shown in Maraun, J Climate, 2013. Please cite.

Response 1.11: The citation is now included at that point in the text.

page 11591, l10: not M-M, but M=M

Response 1.12: Corrected.

Eq. 3: again, use a different name

Response 1.13: This was changed to bias-correction ratio (BCR).

Anonymous Referee #2

There is an increasing need to better interpret and utilize the outputs of climate models to support impact studies. The quantile mapping method has been widely used as a means to improve the correspondence of simulated climate patterns and trends with observed variability and changes. This paper aims to answer whether quantile mapping tends to improve or degrade the performance of a multi-GCM ensemble in reproducing observed changes in precipitations trends over the conterminous United States. The results suggest that quantile mapping modifies simulated precipitation trends and that this effect is model-specific and spatially heterogeneous, which is consistent with some recent studies on this issue. Overall, this paper is well written and organized. Methods are clearly described. Findings are informative and valuable to impact studies regarding water resources management. However, the value and potential impact of this work could be improved through addressing the following issues.

1. It is helpful to use a hypothetical case prior to the real-world case to illustrate the general effect of quantile mapping. As this paper has a clear focus on the performance of a GCM ensemble, I would suggest enhancing the hypothetical case to reflect the cumulative effects of bias corrections of two or more models, i.e., how would the modified trends of individual models amplify or counteract each other in an ensemble context?

> How an ensemble of simulations might affect the trends is included more explicitly in the revised paper. The paragraph following equation 3 in the revised paper includes "From this simple synthetic demonstration, it can be inferred that, if there were a preponderance of GCMs with biases in variance in the same direction, the net effect of QM on the simulated difference between eras could be systematically in one direction, even with random biases in the mean." The following paragraph, in discussing Figure 3, notes that this does appear to be the case with the ensemble of GCMs used in this study.

2. I agree on the merits of quantile mapping in matching observed and simulated precipitation patterns. The authors have made a successful effort to explain the effectiveness of quantile mapping from a statistical perspective. What is lack here is a better discussion of the underlying physical processes associated with the imperfect behaviors of the involved GCMs. It is good to know that quantile mapping comes with a price. However, as shown in Cloke et al. (2013), it would be also important to understand in what circumstances we need to apply this kind of transformations instead of using alternative methods.

> While as noted above, this study did not aim to make any assessment of the skill or applicability of quantile mapping bias correction, the potential for changing mechanisms to drive different responses of the bias correction as far as trend modification is a possibility that deserves to be mentioned. We have added this (and a citation to the interesting Cloke et al. reference) to the revised manuscript (see response 1.4 to reviewer 1).

3. Extreme rainfall events are among the most challenging components in the evaluations of hydrological impacts of climate change. Although the effect of bias correction on extreme events is

listed as a direction of future efforts, I suggest that authors provide more insights into this issue in the present framework. This may highlight the motivation of this work and enhance its value to a broader range of researchers.

> The motivation of the current research and the focus for future efforts is more clearly stated in the last paragraph of the conclusions in the revised paper: "These findings are limited to the extent of this study, namely seasonal mean precipitation for the observed periods used here. This focus was motivated by the observation of changes in mean precipitation produced by quantile mapping. Since changes in the magnitude of extreme precipitation events are important for assessing many impacts to society, future efforts will examine the effect of quantile mapping bias correction on trends in extreme events. Quantile mapping can have different effects at the tails of distributions (Li et al., 2010), and changes in the projected trends in extreme events due to quantile mapping have not been explored."

Reference: Cloke, H. L., Wetterhall, F., He, Y., Freer, J. E. and Pappenberger, F. (2013) Modelling climate impact on floods with ensemble climate projections. Quarterly Journal of the Royal Meteorological Society, 139 pp. 282-297. doi: 10.1002/qj.1998.

Anonymous Referee #3

The paper tests the effect of quantile matching on GCMs over the United States. Although the paper deals with an important issue and is nicely presented, the point that bias correction has to be applied with care is known and discussed in literature already. Also, the setup of the study is questionable, as also pointed out by a previous reviewer.

Major comments

1. Applying bias correction to coupled models with no forcing from SSTs or atmospheric nudging will correct not only for systematic errors but also for internal climate variability. This was mentioned by a previous reviewer, but this has to be more carefully addressed. Quantile mapping cannot be applied in this way unless you can show that the discrepancy between GCM output and observations are only due to model structure errors and not external forcing or internal climate variability. As applied in this study, the corrected output from individual GCMs will have different and conflicting results, but with a large enough sample it will not affect the mean, which is also what the study shows. This is however to be expected, and is not novel. The advice to always use an ensemble of models is good practice, and it is of course worth repeating.

> The aim of this study was not to determine the skill of quantile mapping bias correction, but only its effect on trends. The reviewer is correct in that the presence of internal climate variability in the unconstrained historical GCM simulations would complicate an analysis of bias correction skill. We have replaced the original analysis using unconstrained historical GCM runs with AMIP simulations that are more closely tied to observations. Please see Responses 1.1, 1.3, 1.4 and 1.5 to Reviewer 1 above.

2. Bias correction, or model output statistics (MOS), is usually performed in smaller case studies where quality of precipitation is well known, and the performance of the climate models can be carefully assessed. Using it on such a large area as the US is questionable since, and if used for impact studies on this level I would suggest not using bias correction, or employing a simpler one, like the delta approach. It is not clear what the advantage is to employ it over such a large area.

> Quantile mapping is not applied to the entire domain, but, as now stated in the revised text: "QM is then applied (independently) to each 1° grid cell in the domain." In this way, the size of the domain is irrelevant to the bias correction performed at each grid cell.

3. The authors raise concern on the misuse of MOS for impact studies, but a simple way of avoiding this is to always use raw climate model output together with corrected output to assess the possible climate impacts along with the effect of such a correction. Ideally, the methods of correction should be more than one since they have different characteristics. This would provide a more robust assessment of the uncertainties in the modelling chain as well, and account for the effect of MOS.

> This point is also noted in the Cloke et al. reference, added at the suggestion of Reviewer 2. To emphasize this, the following has been added to the conclusions: "Similar to the suggestions by others (Cloke et al., 2013), it may be prudent for practitioners to examine the projected trends in raw GCM output as well as in bias corrected output, to be completely transparent as to the effects of bias correction on trends."

4. The language in the paper is quite casual, and although this is a style I was sometimes distracted by it, especially the overuse of “we”. For me it is obvious that the authors have carried out the research, and this does not have to be pointed out constantly.

> This was also noted by reviewer 1 (see response 1.8), and has been corrected.

Minor comments.

1. P11586, L5. You use General Circulation Model as the abbreviation GCM, but I think in this context Global Climate Model or perhaps Earth System Model (ESM) is more correct, since it is coupled models. You also point this out later in the text. General Circulation Models is a generic term, and in this case the models are AOGCMs, correct?

> Correct. Global climate models has been used instead.

2. P11588, L14. “quantile for” should be “quantile of”

> Corrected.

Anonymous Referee #4

The core of the study is fundamentally wrong. As already pointed out by reviewers 1 and 3 it is likely that much, if not most, of the observed and the simulated precipitation changes between the analyzed

periods (1916 – 1945 and 1976 – 2005) are caused by random variability rather than by climate forcings. If this is the case a comparison of simulated and observed differences makes no sense at all when standard forced coupled atmosphere ocean GCM simulations are used, which is the case in this study. I find it quite surprising that there is not a single comment on this point in the paper.

> The reviewer is correct in that natural variability is largely the cause of the discrepancy between differences in observed and GCM-simulated trends in the historic unconstrained GCM runs. However with the new experiment using AMIP GCM runs, more closely tied to observations, this effect should be lessened. In addition, the presence of natural variability affecting the simulated precipitation trend would not necessarily result in the effect of quantile mapping on trends tending to zero, which should be more clearly stated in the revised paper (see response to comment 1 by reviewer 2 above). Please see the detailed responses to reviewer 1 numbered 1.1, 1.3, 1.4 and 1.5.

A key conclusion is that the difference between the raw and the bias-corrected ensemble median change is small compared to the observed change. This is a meaningless statement if the simulated changes are random. With an increasing number of GCM simulations considered the median of the raw changes will approach zero, and so will the median of the bias corrected changes, and in turn their difference. This difference can thus in principle be made arbitrarily small if only a large enough number of simulations is considered, and comparing it with the observed change does not provide any useful information.

> This comment is correct in that if the changes were completely random (and centered on zero) they would, of course, tend toward zero in the mean with a large ensemble. It does not follow, however, that the effect of the quantile mapping bias correction would necessarily tend to zero. Please see our responses above, and especially response to Reviewer 2, comment 1.

The paper does contain some useful thoughts about the effect of quantile mapping on simulated differences (eqn. 3 and related text) but this material on its own is in my opinion not enough to justify publication.

> The revised manuscript is much more clear in the questions that are addressed, and the limitations of the findings, which seemed to be a source of confusion with several of the reviewers.

If another version is submitted the authors should distinguish between MOS and PP downscaling (as already pointed out by other reviewers), and also define the term ‘bias’; at the moment it is used with varying meanings in the text.

> Regarding MOS and PP, please see Response 1.7 to reviewer 1, above. The term 'bias' is discussed in greater detail in the revised manuscript; please see Response 1.1 to reviewer 1.