Author responses to Interactive Discussion Comments on "Technical Note: The impact of spatial scale in bias correction of climate model output for hydrologic impact studies" by E. P. Maurer et al.

> *We are grateful for the detailed and helpful comments of four reviewers of this manuscript. We respond to each of the comments below, with original review comments in regular type, and our responses in indented red italics.*

RC C4694 - G. Pegram (Referee)

Review_ hess-2015-316_ Maurer_et_al

Downscaling sets of GCM rainfall and temperature data using quantile transforms from surrogates modelled by NCEP and NCAR time series, to a range of scales from 2 degrees down to 1/8th of a degree, are used to drive the SWAT hydrological model over 185 basins in the Western USA. The sensitivity of the basin responses is determined by comparing their modelled runoff, trained on a 30 year period, against observed data during a following 22 year validation period. It turns out, rather counter-intuitively, that the 'best' result is obtained by the 1/2 degree discretisation; coarser and finer division appears to add noise.

The attractive and valuable characteristics of this paper are that the authors do not try and sell their product, but let the data speak for themselves. The thoroughness, care and clear exposition in the paper is an exemplar. This is the first paper I have ever reviewed in over 40 years which needed not one correction. It is impeccable.

Publish as is!

> *Thank you for the encouraging comments.*

RC C4711

A codicil

This document deserves a higher status than 'Technical note', mainly because of the depth and breadth of the work that went into making it. The offering should be published as a full paper - it is more substantial than most.

> *If its length or other characteristics are, in the opinion of the editor, more fitting as a research article, we would have no objection to re-classifying it as such, though we have no strong opinions regarding this.*

RC C4830 - Anonymous Referee #2

General comments: This technical note illustrates the effect of the spatial scales at which quantile mapping bias correction is performed on coarse climate model data and subsequent hydrologic simulation. Quantile mapping method is most commonly used in the course of downscaling method in US hydrologic projection study, and the authors pointed out the fine spatial scale is tempting to use for the bias correction. However, the greater degree of non-stationarity of bias in higher resolution data (for certain location) have negatively impact on bias correction during validation period (therefore when applying to future climate scenario).

I think the manuscript is already publishable quality in terms of the paper presentation, science question they are asking and the results. However, I would like the authors to consider a few comments below.

Specific comments:
1. Page 10897, L8. The sentence "Sources of fine scale and coarse-scale variability. . ." sounds strange. Also, if you provide some examples of sources of variability at fine and coarse scale data, that helps readers to easily understand this whole paragraph (this part was the most difficult part to understand for me in the paper).

We appreciate this comment, identifying this poorly written sentence. The sentence has been modified and additional text added so the revised text reads: "Since variability at the coarse-scale (due to synoptic circulation, for example) and fine-scale (due to local topographic features, land-atmosphere interactions, etc.) have distinct sources, application of quantile mapping to simultaneously include spatial downscaling is arguably inappropriate. For example, Maraun (2013) highlights an example where a high large-scale precipitation value is translated by quantile mapping to high values at all points within the large-scale grid box, producing an erroneously large and uniform extent of an extreme event; fine-scale variability among the points is not replicated by the deterministic transformation of quantile mapping."

2. It looks like non-stationarity of bias is magnified in higher elevation to me (Fig3). If you agree, I think it would be great to add reason (or speculation that can be explored in the future). And do you think the illustrated impacts of the quantile mapping scale on precipitation bias are greater in complex terrain regions than flat region?

This is a helpful observation that where there is more fine-scale structure to climate that is not represented in the large-scale signal, there will generally be greater biases. We have examined this using relative changes in bias (dividing by mean precipitation), and the absolute values of the bias non-stationarity is comparable in most mountainous regions and the much less dramatic topography in the southeastern part of the domain. We have added a statement to the manuscript, in the paragraph discussing Figure 3: "Figure 3 shows the mountainous regions to have higher biases (and greater values for non-stationarity), which may be expected given greater local complexity of the terrain and thus more heterogeneity in the local precipitation that the bias correction is attempting to correct. However, the apparent higher non-stationarity in mountainous areas is also partially due to the greater precipitation at high elevations. Expressing bias as a relative change in bias (by dividing the bias at each grid cell by the mean observed precipitation) shows higher non-stationarity, and the amplification at some locations, to occur in some mountainous areas but also more broadly over much of the domain, including some prominent valleys such as California's Central Valley. The mechanisms driving the spatial variability in bias non-stationarity, and its amplification when bias correcting at finer scales, is reserved for future research."

The authors show CDF for bias corrected precipitation (and temperature) at one pixel which I believe is pixel where DeltaBias is large in Figure 3 (0.125 degree). If you select one pixel where DeltaBias is small, I would expect the CDF would be closer together even during validation period. I would consider showing "good pixel" and discussing the impact of quantile mapping scale differ from grid box to grid box.

While the chosen cell for Figure 4 was selected more or less randomly, there is indeed a wide variability in how bias correction performs at different scales at each grid cell. We address this comment by adding text to the end of the paragraph discussing Figure 4: "It should be noted that this stark of an example will not exist at every grid cell. Eden et al. (2012) suggest that model *errors due to unrepresented topographic effects on precipitation or inadequate climate model parameterization are most successfully corrected by quantile mapping, so where other small scale variability is less important there may be more successful removal of biases using quantile mapping at finer scales."*

3. Uncertainty of hydrologic simulations is always difficult to discuss due to hydrologic model uncertainty, and their parameters. I would include in the limitation (P 10904, L13-19) a sentence to state the result of the impact on runoff is also from one model.

We are in agreement with this assessment. We have added to the end of the sentence that begins "The spatial scale of the hydrological model, and its representation of sub-grid spatial variability, may also affect the results" the following: ", thus different parameterizations of the SWAT model or the use of other hydrology models would affect results (Ficklin and Barnhart, 2014; Maurer et al., 2010a). "

RC C4835 - Anonymous Referee #3

Review of "Technical Note: The impact of spatial scale in bias correction of climate model output for hydrologic impact studies" by E.P. Maurer, D.L. Ficklin, and W. Wang

This study investigated the impact of spatial scale at which the quantile-mapping bias correction is applied, on the streamflow produced by a hydrologic model. Daily precipitation and maximum and minimum surface air temperature from the NCEP/NCAR reanalysis were bias corrected against the gridded observations at spatial resolutions of 0.125, 0.25, 0.5, 1.0 and 2.0_. The bias-corrected variables were then interpolated to 0.125_ grid before inputting into the SWAT hydrology model to simulate streamflow across the Western United State. Skill was evaluated by comparing the simulated streamflow using bias-corrected reanalysis data with the gauge observations. It was found that while bias correction at the coarse resolution (2.0_) produced the least correspondence with observation-driven streamflow, increasing the spatial resolution to finer than 0.5_ for bias correction did not improve skill and even degraded skill.

This is a well-written paper that explored how bias correction at different spatial scale affects streamflow estimation skill. The methodology is robust and valid. I have only a few questions about the paper. While the bias correction performed at 1_ gives the best skill in streamflow estimation for a river basin with approximately 1.4_ spatial scale (i.e.the Sacramento River), 0.5_ is the optimal scale to apply bias correction for a smaller basin with 1/3_ spatial scale (i.e. the Tule River). The authors stated that there is no clear relationship between drainage areas and the skill (defined by the p values). I would suggest if the authors could show a p-value versus area plot.

Figures 8 and 9 do show p-value versus area, and these were motivated by the same thoughts expressed by the reviewer here.

The methodology involves a few times of interpolation. First, the observations and reanalysis data were interpolated (or downscaled) to the same resolution on which bias correction was applied. The corrected data were then interpolated to 0.125_ resolution before inputting into the SWAT model. How much uncertainty was introduced by the interpolation?

This is an interesting point, though beyond the scope of this investigation. We have added to the paragraph on limitations of the study, at the end of section 3, the following: "Also, this study focused on biases at different scales for output from the BCSD process as it is typically applied. We did not assess the influence of each step in the BCSD process (as shown in Figure 1) on the biases, though this could be a fruitful avenue for future research."

Next, there is evidence of change in the spatial distribution of rainfall, at least in the case of extremes that are often important for hydrologic simulations. An example of this is presented in (LI, J., SHARMA, A., JOHNSON, F. & EVANS, J. 2015. Evaluating the effect of climate change on areal eduction factors using regional climate model projections. Journal of Hydrology, 528, 419–434.). I believe, for completeness of this evaluation, this needs to be acknowledged in the presentation being made here, along with a discussion of how the inputs being used in this study correspond with the type of changes it suggests.

This is a helpful insight, and we are grateful for the reference. At the end of the paragraph (p. 10897 line 13-18 in the discussion paper) discussing quantile mapping and spatial scale of storm structure, we have added: "A further consideration, when applying quantile mapping to future precipitation projections, is that the relationship between the spatial scale of fine- and coarse-scale precipitation may change in ways that could affect extreme runoff projections (Li et al., 2015)."

The conclusion that bias correction at scale of 0.5_ produced best skill in streamflow estimation is based on comparing the SWAT output with gauge observations. Have the authors taken into account the bias caused by using SWAT model (i.e. the discrepancies between the SWAT output based on observed data and gauge observations)? If this bias is also non-stationary, how will it interact with the bias in climate model simulations? For example, if the bias of the SWAT model for the validation period is larger than that for the calibration period, whereas the bias of climate model at 0.5_ resolution for the validation period is smaller than that for the calibration period, then these two kinds of bias can cancel out with each other for the validation period. This will lead to the conclusion that performing bias correction at 0.5_ resolution is optimal in terms of streamflow skill. But if other hydrological model is used, the same conclusion may not hold.

It is certainly correct that different parameterizations or calibrations of the hydrology model, or the use of another model, would also add to uncertainties. This is similar to the last comment by Reviewer 2 above. Please refer to that prior response.

This brings me to a question about the importance of the quantile mapping approach the authors use here. This approach scales precipitation inputs differently for difference cells that are adjacent to each other. Consequently, one can expect the original spatial dependence structure of the atmospheric variables that are used in downscaling are corrected to different extents, forcing a mis-match of their modelled dependence structure and causing implications in the quality of the downscaled simulations obtained. I personally do not like quantile mapping as a bias correction alternative for this reason and would rather trust an approach that corrects the GCM field ensuring the multivariate dependence is intact. There are approaches that do this (MEHROTRA, R. & SHARMA, A. 2015. Correcting for systematic biases in multiple raw GCM variables across a range of timescales. Journal of Hydrology, 520, 214-223.). There are also the Nested Bias Correction approaches that address biases across multiple temporal scales which are of great relevance when simulating flows or soil moisture for agricultural applications. I believe the authors need to broaden their discussion about these issues and the impact of the quantile mapping bias correction they have used.

> *We appreciate this additional reference related to developing improved bias correction schemes. As this is a recurring comment (see comment of reviewer 4 as well), we have expanded this discussion in the revised manuscript to include the following: "In addition to those noted above, there are other known shortcomings of quantile mapping, some of which have been accommodated by modifying or augmenting quantile mapping or by developing alternative statistical procedures. For example, where it is desired to maintain a joint distribution of multiple variables through bias correction, as opposed to individual variable downscaling as used here, joint downscaling methods have been developed (Abatzoglou and Brown, 2012; Mehrotra and Sharma, 2015;Zhang and Georgakakos, 2012). The probability transformations in quantile mapping are incapable of correcting for GCM biases in low frequency variability, and autoregressive and spectral transformations have been developed to accommodate these biases where important (Mehrotra and Sharma, 2012;Pierce et al., 2015). While we recognize the deficiencies in quantile mapping, as discussed for statistical bias correction in general by Ehret et al. (2012), and there is the promise of recent advances in bias correction, it remains that quantile mapping is widely used and generally effective at removing biases (Gudmundsson et al., 2012), even in the presence of some non-stationarity (Lafon et al., 2012;Maurer et al., 2013). Our aim in this study is not to advocate for a specific downscaling method, but to understand a specific aspect of this widely used method."*

Lines 1-3, Page 10899: How were the bias corrected anomalies calculated? Were the anomalies calculated by removing monthly mean or daily mean from the bias-corrected daily time series?

> *Following equation 1, we have modified the text to add the following clarification: "where F is the CDF, at each X° grid cell, for the calibration period, x is a daily value of precipitation or temperature, with the CDF developed for a moving window of ±15 days from the day pertaining to x. The subscripts indicate large-scale model data or observations (obs). After the quantile mapping bias correction, precipitation and temperature values are expressed as anomalies relative to the climatological mean for the moving window, using a difference for temperature and a fraction for precipitation. These anomalies are interpolated from the large scale to the final 1/8° grid and applied to climatological values to obtain final daily downscaled data."*

RC C5079 – Review of Technical Note 'The impact of spatial scale in bias correction of climate model output for hydrologic impact studies' by E.P. Maurer et al.

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 investigate the influence of the spatial resolution at which a bias correction of climate model output (daily precipitation and max and min temperature) is performed. The bias correction (quantile mapping) is set up in a calibration period (1969-1989) and applied in a validation period (1990-2011). Input data are daily observations interpolated on 1/8° grid and reanalysis data on 1.9° grid over several large river basins in the Western United States. The bias correction is performed on data mapped to resolutions ranging from 2.0° to 0.125°. Both the observed and bias corrected data are then further downscaled to 0.125° resolution and fed to hydrological models.

The observed and bias corrected data are compared in various ways: The cdf's for the three meteorological parameters are compared for a single grid cell, the river flow cdfs of daily streamflow are discussed for two gauges, and a test statistic (Mann-Whitney U test for agreement of observation- and reanalysis-based flow cdf median) of annual 3-day high flow and annual 7-day low flow for 185 gauges is discussed for the validation period.

The main findings are that i) there is no clear 'best bias correction resolution' for the meteorological variables. Instead, results vary with variable, location and quantile; ii) with respect to the hydrological output, differences in the quantiles of observation- and reanalysis-driven streamflow are considerable, with a (weak) agreement optimum for bias-correction at 0.5° resolution.

The authors conclude that bias-correcting at 0.5° may indicate an optimum between detail of the observations exploited and nonstationary effects between calibration and validation period minimized due to spatial averaging.

3. Overall ranking

The work is ranked 'Major revision'.

4. Evaluation

For the reasons detailed in Ehret et al. (2012), I still argue that bias correction of climate model projections for further application (e.g. in hydrological models) is highly questionable. However, as the study discussed here rather evaluates than applies a bias correction method, I set these reservations aside for now.

> *In response to comments by reviewer 3, some of these concerns are reflected in the revised manuscript as noted above, and the Ehret et al. reference is included.*

The study is generally short but thoroughly conducted (some minor objections are given below), what strikes me is the weak agreement of the cdf's for river flow (Figures 6 and 7) for the extreme high and low flows and for the agreement of the cdf's in general (Figures 8 and 9). For high flows, the difference between the reference and the bias-corrected reanalysis data is roughly that between a once-in-10 and a once-in-100 year flood (compare discharge from reference at P=0.99 and P from 2.0° bias corrected reanalysis at P = 0.9. Admittedly, from a 20-year data set, extrapolations to 100-year floods are uncertain). For low flow cases this is even worse. As the data used here are reanalysis data, which, as the authors state (10899/6pp) should outperform predictions from free climate model runs, this makes the applicability of climate model output to address questions of hydrological extremes doubtful. Even if this is not the core topic of the manuscript, it should be discussed to put results into perspective.

> *A disclaimer to this effect has been added to the end of the paragraph discussing Figure 7. It reads: "As with Figure 6, Figure 7 shows worse performance of bias correction in many cases at the high and low extremes compared to the center of the distribution, as would be expected with*

fewer observations for defining the driving precipitation and temperature CDFs in the relatively short calibration period. Thus, while quantile mapping generally reduces the biases compared to using raw GCM output, significant biases may remain, especially at the tails of the distributions. If streamflows produced using bias corrected and downscaled GCM output are to be used for analysis of extreme events, it may be desirable to use a further bias correction (such as quantile mapping of simulated streamflows to match observed streamflows), as has been done for water resources system operations and seasonal forecasting (Snover et al., 2003;Yuan and Wood, 2012) to ensure downscaled streamflows are comparable to observations at all quantiles."

Secondly, as the authors correctly state in the text (10895/14pp) quantile mapping bias correction affects both stationary and instationary parts of model-observation discrepancies, which means that it has the potential to worsen results in a validation period, which can partly be attributed to removing the physical coherence between the model output fields. If this procedure is generally accepted in the climate change community, I wonder why bias correction is not applied 'end-of-the-pipe', i.e. directly to the river flow, if this is the variable of interest. The advantage is that the meteorological fields remain coherent, and the undesirable step of disaggregating meteorological input to the scale of hydrological subbasins can be omitted (as the results are re-aggregated by the convolution of river flow anyways). And if application of bias correction to output of meteorological models is justified, why should it not be so for river flow? The only justification that comes to my mind to do the one (bias correction of climate model output) and leave the other (bias correction of hydrological model output) is to argue that the atmosphere-landsurface interface is one of weak interaction. It would be interesting to see whether bias correcting river flow yields better results in a validation period than correcting the meteorological drivers. I realize that this point is not directly related to the scope of this paper, but I wonder if the authors can comment on it.

This is a helpful observation, and one that has inspired past work to this effect, which is cited in response to the prior comment. It is interesting to note that Yuan and Wood compare quantile mapping only streamflow, only meteorology, and both, and find the latter most effective.

Some specific points

• 10898/14: Please give some more information on the underlying observations (number of stations etc.)

This has been added to the revised manuscript at the beginning of Section 2: "The Livneh et al. data use approximately 20,000 sites with daily meteorological records to define their field."

• 10898/25pp: Please clarify whether the cdf's are determined individually for each grid cell or aggregated for larger regions or the entire domain.

This was added in the response to the last comment of Reviewer 3 above.

• 10900/14pp: Please state in which period and with which forcing the hydrological model calibration was done.

This additional detail has been added to the noted paragraph: "All SWAT models were calibrated and validated, at the 185 sites, during the 1950-2005 time period, though *because observations were not complete at all sites some gauges did not encompass the entire period."*

• 10901/1pp and Figure 3: This does not tell us much about the bias in the extremes. Rather show differences for extreme rainfall and length of dry spells, which better corresponds to the hydrological statistics discussed later in the text.

The revised manuscript includes a statement explaining that a similar figure to Figure 3, but for annual maximum daily precipitation, showed a nearly identical pattern. We have added a statement to the revised manuscript, at the end of the paragraph discussing Figure 3 to this effect:" A similar plot to Figure 3, but for annual maximum precipitation, showed comparable patterns and characteristics."

Your sincerely,

Uwe Ehret

Ehret, U., E. Zehe, V. Wulfmeyer, K. Warrach-Sagi and J. Liebert (2012): HESS Opinions "Should we apply bias correction to global and regional climate model data?". Hydrol. Earth Syst. Sci. 16 (9), 3391-3404, 10.5194/hess-16-3391-2012.