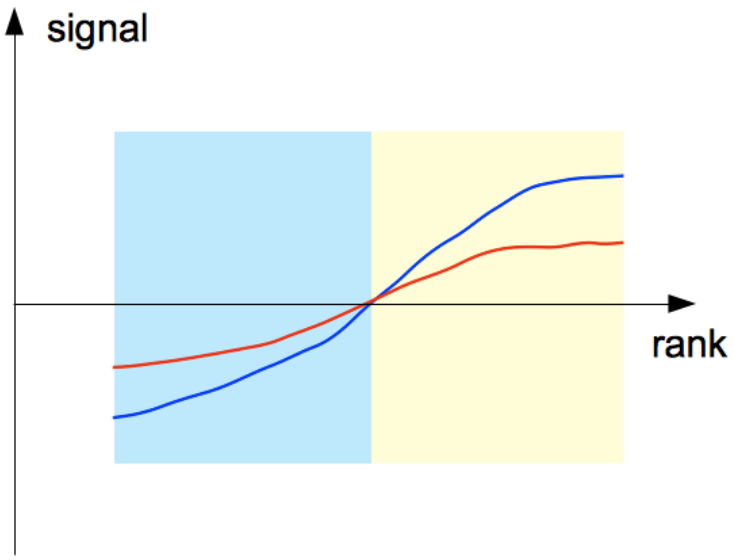
Review of "Is bias correction of Regional Climate Model (RCM) simulations possible for non-stationary conditions?"
by C. Teutschbein and J. Seibert.

The authors propose to use a split sample approach to test how bias correction methods perform under non-stationary conditions. The research question is definitely relevant, and only little has been published on this issue. Also, the manuscript contains some interesting aspects. Nevertheless, I am pretty sure that the authors cannot learn from their study what they intend to learn. Namely, they do not learn anything about the performance of bias correction methods under nonstationary conditions.

Main point

The main shortcoming of the study is that the authors do not clearly define what a bias is, and what nonstationarity of a bias means. In particular they do not distinguish between forced signals and internal climate variability. This imprecision has severe consequences for the design and interpretability of their study.
What is a bias, and what is a bias nonstationarity? A bias is by definition the systematic difference between the observed mean climate and the simulated mean climate, i.e., the difference between the expected observed signal and the expected simulated signal. This implies, and this is the important point, that a bias is by definition unaffected by realisations of internal climate variability. For a good estimate of a bias, one would need many realisations of both the model and observations, or at least a long series in a stationary climate. Consquently, nonstationarities of biases cannot be caused by internal climate variability, but only by changes in the mean climate itself, i.e., by changes in external forcings (for a discussion, see Maraun, Geophys Res Lett, 2012).



Consider the following example (see figure): assume a GCM/RCM simulation in a stationary climate (i.e., no nonstationarities!) that correctly simulates the mean climate (i.e., no bias of the mean!), but with a biased representation of internal climate variability. In fact, assume that the internal variability is only half as strong as in the real world.
The authors would take the observed data (e.g., temperature), sort them (blue line in the figure) and split the sample into the lower half for calibration (light blue), and the other half for validation (light yellow). The same procedure is carried out with the regional climate model (RCM) simulation. Because of the under-represented internal climate variability, the amplitude of the simulated signal is only half of the observed data (red line). Now in the calibration period, the RCM signal is higher than the observed signal, i.e., the calibration would estimate a positive bias in the mean climate. In the verification period, however, the RCM signal is lower than the observed signal. Consequently,

the authors would detect a negative bias in the mean climate, i.e., a bias nonstationarity. Yet by construction, there is no mean bias at all, and also no nonstationarity! The discrepancy instead arises from a wrong representation of internal variability, i.e., a bias in the second moment of the distribution of the considered variable.
To summarise: the chosen approach is not able to assess bias nonstationarities, basically because it cannot discriminate between changes in forcings (which would cause bias nonstationarities) and internal climate variability (which does not cause bias nonstationarities).

Further points

The statement that we are not able to check whether the stationarity assumption is actually true or not (p 12771, l 2) is not correct. In fact, in a pseudo reality, one can at least test whether the assumption is wrong. Two recent studies have addressed these questions, Maraun, Geophys Res Lett, 2012; and Räisänen and Räty, Clim Dynam, 2012. These studies should be cited.

Apart from the discussion above, the authors should also clearly discuss whether they address bias nonstationarities, or apparent bias nonstationarities caused by insufficient correction methods (for a discussion, see Maraun, Geophys. Res. Lett., 2012).

For precipitation, the climte change signal in this short period is very weak only. Especially here, the fundamental flaw discussed above will come into play, as one only looks at internal variability, not at any forced trends.

Please clearly explain how the bias correction methods have been applied! Did you apply them to the 4 seasons separately? Months? The whole year?

The term „linear scaling“ does not make sense for temperature (table 3). Scaling is by definition multiplicative, as it keeps the proportions of the scaled object. See Widmann et al., J Climate, 2003, and http://en.wiktionary.org/wiki/scale#Verb or any relevant mathematical textbook.
Also the term „statistical downscaling“ for quantile mapping (table 3) makes no sense. Statistical downscaling is the generic term for all statistical methods attempting to bridge the gap between large scales and local scales. Please avoid using misleading and wrong terminology!

Please clarify whether the example in Fig. 3 is artificial or not (what does it represent?). Also it makes no sense to add year from 1963 onwards, as the x-axis shows ranks, not years.

Sometimes the authors have to many self-citations. For instance, l 12769: "As Teutschbein and Seibert (2010) concluded, multi- model approaches (i.e. ensembles) have two advantages:" This conclusion has not been drawn by Teutschbein and Seibert for the first time. Here, the standard literature should be cited. Also regarding downscaling methods, the classical papers introducing PP and MOS (i.e., bias correction, e.g., Klein, BAMS, 1974) and relevant reviews (e.g., Maraun et al., Rev. Geophys., 2010) should be cited.

The statement „In this study, the differences between designed calibration and validation period were within a range of 18–36 % for precipitation (Fig. 4, left) and 0.86–1.75 C for tempera- ture (Fig. 4, right). These values represent a reasonable climate change signal that is likely to occur within this century (IPCC, 2007).“ (p 12772, l 16) is obviously misleading, as the likely temperature increase is much higher.

The term „ensure“ (p 12776, l 6) is overly optimistic and rather naive.