Hydrol. Earth Syst. Sci. Discuss., 12, C3784–C3795, 2015
www.hydrol-earth-syst-sci-discuss.net/12/C3784/2015/
© Author(s) 2015. This work is distributed under
the Creative Commons Attribute 3.0 License.

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Hydrologic extremes – an intercomparison of multiple gridded statistical downscaling methods" *by* A. T. Werner and A. J. Cannon

A. T. Werner and A. J. Cannon

wernera@uvic.ca

Received and published: 18 September 2015

We thank both Anonymous Referees for their valuable comments and suggestions that greatly improved the quality of this manuscript.

Referee #1

1) p. 6183, line 4, other perhaps better references for SDBC would be Hwang and Graham (HESS, 2013) and Abatzoglou and Brown (J. Climate, 2012).

Thank you for pointing out that the Ahmed et al., 2013 reference was not the primary reference for the SDBC method. We have replaced it with Abatzoglou and Brown

2012 as given by Ahmed et al., 2013 and Hwang and Graham 2013 for this method. Additionally, the Hwang and Graham 2013 reference is highly relevant for this area of study. Thank you for bringing it to our attention.

2) p. 6184, line 14, it might be noted that the presence of internal variability ensures that non-stationarity will always exist between two different time periods in any data set of observed meteorology. There is no way to "ensure against" it.

In this sentence we are trying to make the point that not all gridded-observations are created in the same way or with attention to temporal inhomogeneity caused by stations dropping in and out over time. This has implications for the success of downscaling methods. To better articulate this we modified the sentence as follows "We know that statistical downscaling methods perform poorly when non-stationarity occurs between the calibration and validation periods (Maurer et al., 2013), but we haven't evaluated how apparent non-stationarity caused by natural climate variability (Maraun, 2012; Huang et al., 2014) is amplified or diminished with methods used to create gridded observations, which could also affect the success of downscaling methods".

3) p. 6184, line 18, it is stated that "previous studies have included as many years as possible in the calibration" of downscaling. I do not believe that is the case – there is a balance between including enough years so internal variability does not dominate differences between different periods, but short enough where trends in the data do not introduce erroneous variability.

The Referee is correct in stating that there are studies that have tried to strike a balance with the number of years used in calibration to represent natural variability, but to avoid trends. However, there are studies that have included the full length of record, such as Bürger et al. 2012a, Salathé 2005 and Werner 2011, as listed in the manuscript. This is more common when applying the monthly BCSD method, as was done in these studies, because the daily events are resampled from the historic record and short records would constrain the number of samples available for temporal disaggregation.

Additionally, there are other studies, such as Huang et al. 2012 and Themeßl et al. 2012, which state that a long calibration period is required when bias correcting data for use with extremes. One objective of this study was to "to learn more about the strengths and weaknesses of two gridded observations for use with hydrologic modelling". This long calibration period also assisted with comparing VIC Forcings and ANUSPLIN for their full length of record.

The passage was adjusted as follows to better reflect these points "Not all, but some previous studies have included as many years as possible in the calibration, with the goal of maximizing the available historical record available for resampling in the temporal disaggregation step applied in BCSD (Bürger et al., 2012a; Salathé, 2005; Werner, 2011). This approach is also supported by other studies that found bias correction is more robust for larger samples from longer time series, especially for extremes, such as flood events (Huang et al. 2012; Themeßl et al. 2012)."

4) p. 6184, line 29, the text states that BCSD has not been tested with Tmax and Tmin, which is not correct. In its daily version it usually does explicitly include both Tmax and Tmin, by some method or another. See for example Thrasher et al. (HESS, 2012).

Thank you for your comment. We are evaluating the monthly BCSD instead of daily because the monthly version has been used in many hydrologic modelling studies. It is true that BCSD has been tested with Tmin and Tmax in its daily version. However, using Tmin and Tmax with the monthly version has not been tested, to the authors' knowledge. We have inserted the Thrasher reference to highlight what is known about the effect of daily BCSD on the diurnal temperature range. The sentence will be modified and an additional sentence will follow.

"Applying BCSD using minimum and maximum monthly temperature instead of mean monthly temperature has not been tested and may correct some issues with diurnal temperature range (Bürger et al., 2012a). It is important to note that the effect of BCSD on daily temperature range (DTR) when used with daily data and ways to ensure

minimum temperature is less than maximum temperature has been tested by Thrasher et al. (2012) and was not the focus of this study."

5) p. 6189-6190, a summary table (or maybe a graphical flowchart?) of the downscaling methods would be helpful, especially if it showed their relationships. The many acronyms were at times difficult to follow.

We have prepared the following diagram that includes a summary table, which will be added to the manuscript. We depict the BCSD, BCCA and BCCI methods in full and summarize the augmentations made to these methods to arrive at the remaining methods in the table. It was too cumbersome to include diagrams of all seven methods and it was deemed unnecessary given the relatively few added or exempted steps from one version to the other.

(See Figure 3a - Attached)

Figure 3a. Diagram of the Bias Corrected Spatial Disaggregation (BCSD), Bias Corrected Constructed Analogues (BCCA) and Bias Corrected Climate Imprint (BCCI) downscaling methods and a summary of adjustments made to these methods to create BCSD with monthly minimum and maximum temperature (BCSDX), Double BCCA (DBCCA), Climate Imprint (CI) and BCCA corrected to BCCI (BCCAQ).

6) p. 6191, line 3 and Table 2, the use of different calibration periods for different reanalysis products is problematic, as noted in my comment 2 above and later in the manuscript (p. 6199, p. 6201). Was the motivation simply to include a long period for calibration? If so, that would also cause issues (as noted in my comment 3 above). Maybe adding one additional downscaling demonstration with NCEP1 using just 1979- might help show how important that decision was for the results.

The motivation was three-fold 1) to follow the approach taken by Burger et al. 2012 where as many years as available were used; 2) to replicate the selection process for Werner et al. 2011, etc. as well to measure the potential consequences of this longer

calibration period and 3) to do this with two gridded-observations to test the potential trends in the gridded observations and their influence on the downscaling. Also, we were able to downscale two reanalyses over the long period (20CR and NCEP1) and demonstrated that the problems still persist with 20CR even though it is not documented to have the same problems with non-stationarity as NCEP1. Furthermore, Guttmann et al., 2014 showed that selecting the post-1979 period still led to issues with the downscaling related to non-stationarity. We have made some adjustments to the introduction in response to your third comment that better explain the rationale for the different calibration periods.

7) p. 6196, line 16, elaborate a little on what the Walker field significance test is and how it is applied.

We've added significantly more detail on the Walker field significance test:

"The 101,000 km2 Peace River basin is represented by 3975 grid cells at the 1/16° resolution used to run the VIC hydrologic model. The KS and correlation tests are conducted on each of the grid cells in the Peace River basin for each climate index. Statistical significance of the KS test and Pearson's correlation results over the basin as a whole is measured using a field significance test; the Walker field significance test (Wilks, 2006), where the evaluation of field significance is done by using the minimum local p value as the global test statistic. The Walker field significance test was selected because it is relatively insensitive to correlations among local tests allowing global tests based on data exhibiting both spatial and temporal correlation to be conducted. Temporal and spatial correlation between climate indices grids would require a cumbersome procedure to address correctly with conventional resampling tests. Walker's test, can be seen as being closely related to the conventional (von Storch, 1982) field significance test based on counting significant local results, except that Walker's test statistic is the smallest of the K local p values, rather than the number of K local tests that are significant at some level."

8) p. 6197, line 20, BCCA is shown to perform better with one observed data set. It also seems that for all other downscaling methods the two observed data sets are fairly consistent.

Figure 7 and Figure 9 were flipped 180° in the publishing process. At the time it seemed like it wouldn't interfere with understanding, but I see now that they are easier to interpret when presented the other way. There are more dark grey boxes in the column for ANUSPLIN than the VIC Forcings, which means that ANUSPLIN passes more tests than VIC Forcings (we also added the meaning of dark versus light boxes to the figure caption). We can confirm this with Table 8, which corresponds most directly to Figure 7 and Figure 9 out of all of the tables because it gives the count of number of tests passed by each combination of observation, downscaling method and reanalysis. Comparing the number of tests passed for ANUSPLIN versus VIC Forcings for NCEP1; more tests are passed for BCCA, DBCCA, CI and BCSD for ANUSPLIN than VIC Forcings, vice versa for BCSDX and BCCAQ and the same for BCCI; for ERA40 and ERAInt, considerably more tests are passed with ANUSPLIN than VIC Forcings for all downscaling methods. It is only with 20CR that more tests are passed with VIC Forcings than ANUSPLIN for all downscaling methods except CI that passes the same amount in both. Thus, for Pearson's correlation, the vast majority of downscaling methods passed more tests for NCEP1, ERA40 and ERAInt reanalyses under ANUSPLIN than VIC Forcings. In the case of the KS test, the ANUSPLIN based comparisons of downscaled based simulations versus gridded observations always pass more tests than those based on VIC Forcings.

To help the reader we have adjusted the sentence here to refer to Table 8 sooner and clarify things overall.

"Irrespective of downscaling method or reanalysis, those methods calibrated and validated against the ANUSPLIN gridded observations were more successful vs. those based on VIC Forcings overall (Table 6) although there were some cases where VIC Forcings passed more tests than ANUSPLIN (Table 8). For example, under the BCCA

method, precipitation amounts on extremely wet days (R95p) for all reanalyses based on VIC Forcing failed the Walker field significance test for the Pearson's correlation while those for ANUSPLIN passed (Figure 7). (Note: time series shown are averages of all of the VIC Forcings or ANUSPLIN cells in the Peace basin, while the significance of results was based on the Walker field significance of the correlation tested on each grid cell in the basin.)"

Referee #2

General comments:

In their study the ability of seven different statistical downscaling methods is analysed to replicate properties of climate and hydrologic extreme indices. For this purpose, the authors are using different statistical tests and a split-sample validation approach. Four different reanalyses products are used as climate surrogates, which are downscaled to two gridded observational data sets. This is an interesting study, which is of certainly useful for the readers of HESS.

Overall, the quality of the paper is good, however, I think that mainly the methodological section needs significant improvement before publication. Since this may also require a repetition of statistical tests and rewriting of parts of the results, I recommend "major revisions".

Major issues: The description of the statistical tests (section 3.6) needs improvement: 1) The test for "Pearson's correlation" is mentioned (L6, page 6194), however, I do not understand at all. I guess the authors mean the test of significance for the calculated Pearson correlation coefficients. If this assumption is correct performed the author's will have to clarify how this is done. Additionally, this is procedure assumes that the variables follow a normal distribution. I doubt this is true for the extreme indices this study focuses on.

Thank you for pointing out that more detail is required. We have added the following:

"Pearson's correlation is used to test the temporal correspondence between the annual climate indices for the statistically downscaled reanalyses and the associated gridded observation. Pearson's product moment correlation coefficient is used to measure the linear correlation between climate indices from downscaled reanalyses and indices from observations. If the p-value was < 0.05 the downscaled and observed samples were not linearly correlated."

2) Similarly for the KS-test. The authors should at least provide information about the hypothesis, which are tested, the level of significance under consideration, etc.

We have included the details you quite rightfully requested.

"The KS test is a nonparametric test of the equality of continuous one-dimensional probability distributions. Here, it is used to compare two samples, namely annual climate indices for the statistically downscaled reanalyses and the associated gridded observation. The KS test statistic is used to quantify the distance between empirical distribution functions of these two samples. The null hypothesis is that the two samples are drawn from the same distribution and is rejected if p-value < 0.05. The distributions considered under the null hypothesis have to be continuous distributions but are otherwise unrestricted. While some of the climate indices are not strictly continuous (e.g., frost days, etc.), asymptotic critical values may still be used in the presence of a small number of ties (Janssen 1994)."

3) The Walker field significance test. I have no knowledge about this test, and I think the authors should give much more details about the test than a single reference only. It seems that this test is only rarely applied in hydrology and climatology.

We have expanded the description from a sentence to a paragraph.

"The 101,000 km2 Peace River basin is represented by 3975 grid cells at the 1/16° resolution used to run the VIC hydrologic model. The KS and correlation tests are conducted on each of the grid cells in the Peace River basin for each climate index.

Statistical significance of the KS test and Pearson's correlation results over the basin as a whole is measured using a field significance test; the Walker field significance test (Wilks, 2006), where the evaluation of field significance is done by using the minimum local p value as the global test statistic. The Walker field significance test was selected because it is relatively insensitive to correlations among local tests allowing global tests based on data exhibiting both spatial and temporal correlation to be conducted. Temporal and spatial correlation between climate indices grids would require a cumbersome procedure to address correctly with conventional resampling tests. Walker's test, can be seen as being closely related to the conventional (von Storch 1982) field significance test based on counting significant local results, except that Walker's test statistic is the smallest of the K local p values, rather than the number of K local tests that are significant at some level."

4) Likewise, the presentation of the results of the tests confuses me, i.e. mainly Table 6 – 12, Figure 7 and 9. In the captions of the tables, the "number of test passed" are mentioned, or "similarity in the distributions" is mentioned. Since the tests are not explained in detail in the methodological section, I have difficulties to follow. I also doubt that the number of tests passed is a good indicator, and I am wondering if the grid cells that passed the tests are somehow clustered in space, depending e.g. on the terrain.

As mentioned above the statistical tests have been discussed in much greater detail in the methods section. This background should better support the results presented in Tables 6 – 12, Figure 7 and 9. The number of tests passed in the case of the climate indices is based on the Walker field significance test. We have chosen the Walker field significance test because it avoids problems of the conventional counting test, which typically includes many false rejections of local null hypotheses among the nominally significant local tests. The Walker test tends to identify only the most significant local tests. The convention of using the number of tests passed as an indicator of success of a method is adapted from Burger et al. 2012.

5) What do you mean by similar distributions? Is it the same family of a distribution with slightly different parameters?

The KS test statistic quantifies the distance between the empirical distribution functions of these two samples (one based on downscaled results and the other based on gridded-observations). The null hypothesis is that the two samples are drawn from the same distribution and is rejected if the p value < 0.05.

6) In Figure 7, you can obviously have dark and light grey boxes, but what does it mean?

That's quite an oversight on our part. Thank you for catching this. We have added a key sentence to the captions for Figure 7 and 9.

Figure 7. Field significant correlations based on the Walker field significance test over the Peace River basin between ClimDEX indices for downscaled reanalysis versus target gridded observation, VIC Forcings (left) and ANUSPLIN (right), by downscaling method for 1991-2005 (1991-2001 ERA40). Dark grey boxes indicate statistically significant correlations.

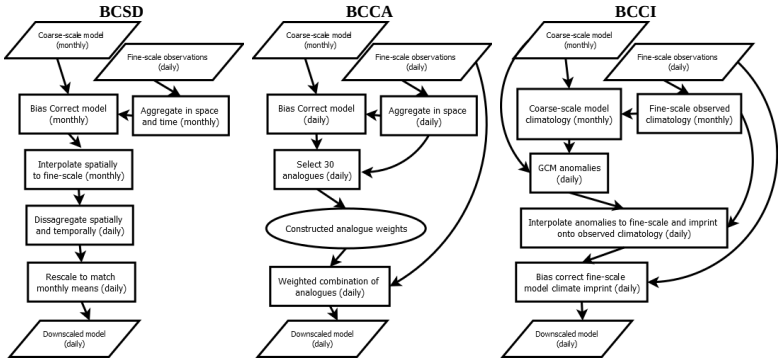
Figure 9. Field significant similarities of distributions based on the Walker field significance test over the Peace River basin between ClimDEX indices for downscaled reanalysis versus target gridded observation, VIC Forcings (left) and ANUSPLIN (right), by downscaling method for 1991-2005 (1991-2001 ERA40). Dark grey boxes indicate statistically significant similar distributions.

Minor issues: 1) Reading the abstract is quite difficult due to the abbreviations, which are mostly quite similar (line 13-15). I suggest leaving out the abbreviations in the abstract. A table explaining the methods in brief at the beginning of the methods and including a list of the abbreviations would be very helpful for the reading process.

In reviewing other publications that compare downscaling methods it appears standard to provide abbreviations in the abstract. Thus, we will continue to follow that custom.

We have included a diagram and summary table explaining the downscaling methods, which should make related acronyms easier to follow.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 12, 6179, 2015.



BCSDX Same as **BCSD** except quantile mapping of monthly minimum and maximum temperature, versus monthly mean temperature.
DBCCA Same as **BCCA** except there is an extra quantile correction at the fine-scale to get rid of drizzle and other biases caused by combining patterns from 30 days.
CI Same as **BCCI** except without bias correction. A form of delta-method.
BCCAQ Daily **BCCI** outputs at each fine-scale grid point are reordered within a given month according to the daily **BCCA** ranks.

Figure 3a. Diagram of the Bias Corrected Spatial Disaggregation (BCSD), Bias Corrected Constructed Analogues (BCCA) and Bias Corrected Climate Imprint (BCCI) downscaling methods and a summary of adjustments made to these methods to create BCSD with monthly minimum and maximum temperature (BCSDX), Double BCCA (DBCCA), Climate Imprint (CI) and BCCA corrected to BCCI (BCCAQ).

Fig. 1.