Final Authors' Response to Referee Comments on "Hydrologic extremes – an intercomparison of multiple gridded statistical downscaling methods" by A. T. Werner and A. J. Cannon

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 (page 4 line 25). 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" (page 6 line 5).

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).” (page 6 line 25)

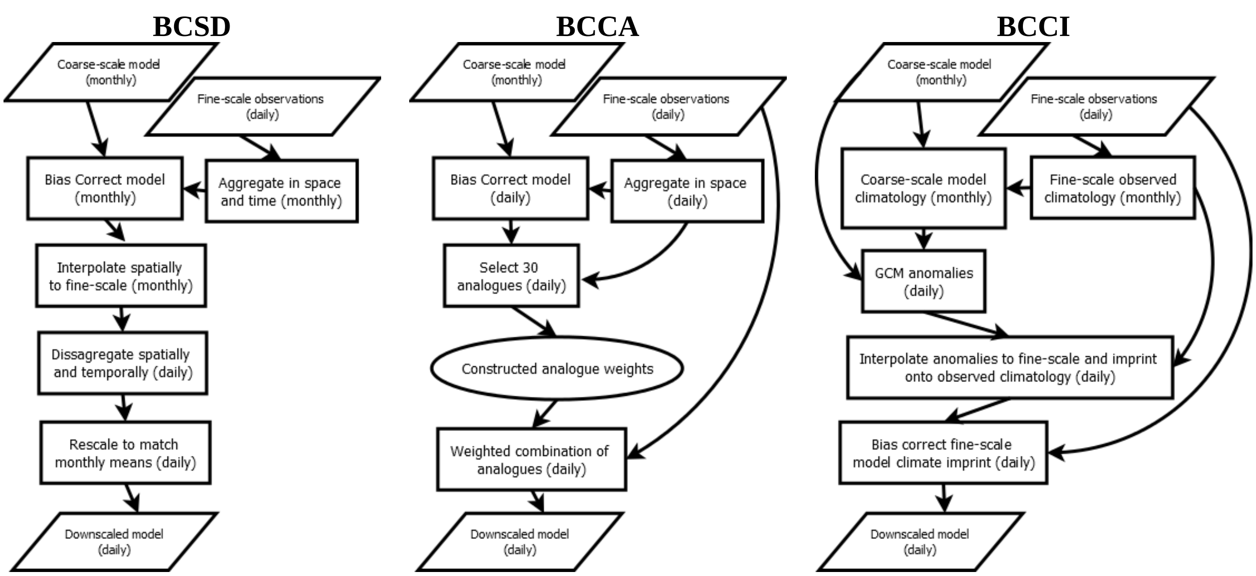
4) p. 6184, line 29, the text states that BCSD has not been tested with Tmax and Tmin, which is not correct. In its daily version it usually does explicitly include both Tmax and Tmin, by some method or another. See for example Thrasher et al. (HESS, 2012).

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.” (page 7 line 6)

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. (Page 48)



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).

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. (page 6 line 6)

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 km^2 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." (page 16 line 6)

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.)" (page 19 line 9)

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.” (page 15 line 32)

To clarify our application of Pearson's product moment correlation, which relies on the Student’s t distribution as the sampling distribution for the (transformed) correlation coefficient, the t test is robust even under violation of normality for sample sizes in this study (Edgell and Noon, 1984).

Edgell, S.E. and S.M. Noon, 1984. Effect of violation of normality on the t test of the correlation coefficient. Psychological Bulletin, 1984, Vol. 95, No. 3, 576-583.

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).” (page 15 line 23)

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 km^2 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.” (page 16 line 6)

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. (page 53 and page 55)

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. (page 48)

Hydrologic Extremes – An Intercomparison of Multiple Gridded Statistical Downscaling Methods

A. T. Werner and A. J. Cannon

Pacific Climate Impacts Consortium, Victoria, British Columbia, Canada

Correspondence to: A. T. Werner (wernera@uvic.com)

Abstract

Gridded statistical downscaling methods are the main means of preparing climate model data to drive distributed hydrological models. Past work on the validation of climate downscaling methods has focused on temperature and precipitation, with less attention paid to the ultimate outputs from hydrological models. Also, as attention shifts towards projections of extreme events, downscaling comparisons now commonly assess methods in terms of climate extremes, but hydrologic extremes are less well explored. Here, we test the ability of gridded downscaling models to replicate historical properties of climate and hydrologic extremes, as measured in terms of temporal sequencing (i.e., correlation tests) and distributional properties (i.e., tests for equality of probability distributions). Outputs from seven downscaling methods – bias correction constructed analogues (BCCA), double BCCA (DBCCA), BCCA with quantile mapping reordering (BCCAQ), bias correction spatial disaggregation (BCSD), BCSD using minimum/maximum temperature (BCSDX), climate imprint delta method (CI), and bias corrected CI (BCCI) – are used to drive the Variable Infiltration Capacity (VIC) model over the snow-dominated Peace River basin, British Columbia. Outputs are tested using split-sample validation on 26 climate extremes indices (ClimDEX) and two hydrologic extremes indices (3-day peak flow and 7-day peak flow). To characterize observational uncertainty, four atmospheric reanalyses are used as climate model surrogates and two gridded observational datasets are used as downscaling target data. The skill of the downscaling methods generally depended on reanalysis and gridded observational dataset. However, CI failed to reproduce the distribution and BCSD and BCSDX the timing of winter 7-day low flow events, regardless of reanalysis or observational dataset. Overall, DBCCA passed the greatest number of tests for the ClimDEX indices, while BCCAQ, which is designed to more accurately resolve event-scale spatial gradients, passed the greatest number

of tests for hydrologic extremes. Non-stationarity in the observational/reanalysis datasets complicated the evaluation of downscaling performance. Comparing temporal homogeneity and trends in climate indices and hydrological model outputs calculated from downscaled reanalyses and gridded observations was useful for diagnosing the reliability of the various historical datasets. We recommend that such analyses be conducted before such data are used to construct future hydro-climatic change scenarios.

1 Introduction

Water resources infrastructure is designed to accommodate hydrologic extremes such as floods and droughts ~~(Cunderlik and Ouarda, 2009; Cunderlik et al., 2004; Ouarda et al., 2006). The frequency and magnitude of extreme hydrologic events, such as floods and droughts have changed with climate and there is broad agreement that changes will continue with projected increases in greenhouse gases (IPCC, 2013). The direction and magnitude of change is not uniform across the globe, but regionally specific, distinguishable by hydrologic regime and by local changes to temperature and precipitation (Cunderlik and Ouarda, 2009; Monk et al., 2011; Sheffield et al., 2012; Stahl et al., 2010, 2012). For example, in Canada, floods in snowmelt dominated regimes decreased in magnitude while floods in rainfall fed regimes had no significant trend over 1974 to 2003 (Cunderlik and Ouarda, 2009). Conversely, Canadian annual low-flow indices showed spatially uniform decreases over 1970 to 2005 (Monk et al., 2011)~~(Cunderlik and Ouarda, 2009; Cunderlik et al., 2004; Ouarda et al., 2006). The frequency and magnitude of extreme hydrologic events, such as floods and droughts have changed with climate and there is broad agreement that changes will continue with projected increases in greenhouse gases (IPCC, 2013). The direction and magnitude of change is not uniform across the globe, but regionally specific, distinguishable by hydrologic regime and by local changes to temperature and precipitation (Cunderlik and Ouarda, 2009; Monk et al., 2011; Sheffield et al., 2012; Stahl et al., 2010, 2012). For example, in Canada, floods in snowmelt dominated regimes decreased in magnitude while floods in rainfall fed regimes had no significant trend over 1974 to 2003 (Cunderlik and Ouarda, 2009). Conversely, Canadian annual low-flow indices showed spatially uniform decreases over 1970 to 2005 (Monk et al., 2011). Thus, future changes in hydrologic extremes need to be estimated at regionally relevant resolutions (~10 km) and consider both temperature and precipitation effects.

~~Global climate models (GCMs) are one of our only tools for projecting the future climate, but they operate at scales too coarse (~100 km) for use in regional studies. Hence, before projecting changes in hydrologic extremes some intervening steps are required. Approaches to converting coarse scale GCM simulations to project changes to peak flows and low flows vary. Some examples include: direct downscaling of streamflow extremes by sparse Bayesian learning and multiple linear regression (Joshi et al., 2013); weather generators combined with hydrologic models (Cunderlik and Simonovic, 2007); regional frequency analysis of regional climate model (RCM) projections (Clavet-Gaumont et al., 2013); and, most commonly, statistical downscaling of GCM or RCM projections run through a physically based hydrologic model (Elsner et al., 2010a; Maurer et al., 2010; Schnorbus et al., 2014; Shrestha et al., 2012;~~ Global climate models (GCMs) are one of our only tools for projecting the future climate, but they operate at scales too coarse (~100 km) for use in regional studies. Hence, before projecting changes in hydrologic extremes some intervening steps are required. Approaches to converting coarse scale GCM simulations to project changes to peak flows and low flows vary. Some examples include: direct downscaling of streamflow extremes by sparse Bayesian learning and multiple linear regression (Joshi et al., 2013); weather generators combined with hydrologic models (Cunderlik and Simonovic, 2007); regional frequency analysis of regional climate model (RCM) projections (Clavet-Gaumont et al., 2013); and, most commonly, statistical downscaling of GCM or RCM projections run through a physically based hydrologic model (Elsner et al., 2010a; Maurer et al., 2010; Schnorbus et al., 2014; Shrestha et al., 2012; Bürger et al., 2011)~~. The uncertainty in hydrologic projections from GCMs is greater than that from emissions scenarios or model parameterizations (Bennett et al., 2012; Prudhomme and Davies, 2008) and all GCMs represent the climate imperfectly in different ways (Gleckler et al., 2008; Knutti et al., 2008); therefore to fully characterize the uncertainty in projected hydrological extremes an ensemble of GCMs is required.~~

~~Gridded statistical downscaling methods provide a computationally efficient and effective means of producing plausible hydro-climatology from a large ensemble of GCMs (Salathe et al., 2007; Salathé, 2005; Wood et al., 2004)~~. The uncertainty in hydrologic projections from GCMs is greater than that from emissions scenarios or model parameterizations (Bennett et al., 2012; Prudhomme and Davies, 2008) and all GCMs represent the climate imperfectly in different ways (Gleckler et al., 2008; Knutti et al., 2008); therefore to fully characterize the uncertainty in projected hydrological extremes an ensemble of GCMs is required.

Gridded statistical downscaling methods provide a computationally efficient and effective means of producing plausible hydro-climatology from a large ensemble of GCMs (Salathe et al., 2007; Salathé, 2005; Wood et al., 2004). A number of studies have compared multiple statistical downscaling methods for use in climatological or hydrological projections. Maurer and Hidalgo (2008) compared constructed analogues (CA) and bias correction spatial disaggregation (BCSD) using the National Centers for Environmental Prediction / National Center for Atmospheric Research Reanalysis I (NCEP1) (Kalnay et al., 1996) as a surrogate GCM. Methods were comparable in producing precipitation and temperature at a monthly and seasonal level, but skilfully downscaled daily data depended on the ability of the climate model to show daily skill. Bürger et al. ~~(2012a)~~(2012a) compared five methods for their ability to represent climatic extremes including BCSD and expanded downscaling (XDS). The fixed diurnal temperature range in BCSD was seen as a shortcoming in Bürger et al. ~~(2012a)~~(2012a). XDS performed best passing 48% of single tests on average for 27 Climate Indices of Extremes (ClimDEX) with BCSD close behind passing 45% ~~(Bürger et al., 2012a)~~(Bürger et al., 2012a). Pierce et al. ~~(2013)~~(2013) found that projected increases in annual precipitation versus decreases in California were due to disagreements in the occurrence of the heaviest precipitation days (>60 mm day-1) amongst three dynamical and two statistical downscaling methods (BCCA and BCSD). Maurer et al. (2010) compared BCSD, BCCA and CA for their ability to reproduce hydrologic extremes. BCCA, when combined with the Variable Infiltration Capacity (VIC) model, consistently outperformed the other methods in simulating 3-day peak flow and 7-day low flow. BCCA is an improvement over CA because it includes bias correction and over BCSD because it includes daily GCM anomalies (Maurer et al., 2010). ~~An additional method described as Statistical Downscaling and Bias Correction (Ahmed et al., 2013)~~An additional method described as Statistical Downscaling and Bias Correction (Abatzoglou and Brown, 2012) and as Asynchronous Regression ~~(Gutmann et al., 2014)~~(Gutmann et al., 2014), both of which interpolate from the GCM to fine scale and then apply quantile mapping bias correction (i.e., basically reversing the steps of BCSD), was found to reproduce extreme precipitation events at the grid-scale, but overestimate them on aggregate scales ~~(Maraun, 2013).~~(Maraun, 2013). Studies to date have not assessed the strength of downscaling methods for use with climatic and hydrologic extremes concurrently.

The first generation National Centers for Environment Prediction / National Center for Atmospheric Research Reanalysis I (NCEP1) reanalysis (Kalnay et al., 1996) is often used as a surrogate GCM when testing downscaling techniques ~~(Bürger et al., 2012a; Gutmann et al., 2014; Maurer et al., 2010)~~(Bürger et al., 2012a; Gutmann et al., 2014; Maurer et al., 2010), primarily because of its long record length. Recently new reanalysis products have come on line bringing to light possible issues with NCEP1, such as a spurious pattern in precipitation fields at high latitudes ~~(Sheffield et al., 2012)~~(Sheffield et al., 2012), and lack of skill in producing daily air temperature at high-altitudes versus other reanalyses ~~(Hofer et al., 2012)~~(Hofer et al., 2012). Reanalyses differ due to variations in assimilated observational data, assimilation methods, representations of surface and boundary layer processes, physics packages, and dynamical cores, and the resulting uncertainty in output fields can be considerable, especially for climatic extremes ~~(Sillmann et al., 2013)~~(Sillmann et al., 2013). For instance, discrepancies between reanalyses for some climate extreme indices, such as frost days in some regions, are sometimes as large as the typical inter-model spread of the Coupled Model Intercomparison Project ensembles, ~~(Sillmann et al., 2013)~~(Sillmann et al., 2013). These differences arise because near surface temperature and precipitation extremes are calculated from variables that are relatively poorly constrained by observations in reanalyses. Additionally, non-stationarity exists in some reanalysis products because they amalgamate observational datasets from different sources over time (Donat et al., 2014). In the context of historical validation of downscaling methods, statistical downscaling methods may perform poorly simply because reanalysis outputs are not stationary over the calibration and validation periods (Maurer et al., 2013). All of these factors suggest that multiple reanalysis products should be used as GCM surrogates to ensure methods are not failing due to irreparable errors in reanalyses, and also to explore the variability in results due to reanalysis uncertainty.

Field Code Changed

~~Gridded climate observations underpin hydrologic projections. They are used to calibrate the downscaling technique and the hydrologic model, serving as targets and inputs respectively. Gridded observations are commonly evaluated via comparison with station observations (Hutchinson et al., 2009; Werner et al., 2015), intercomparison with other gridded observations (Eum et al., 2014) or by using them to drive a hydrologic model and comparing outputs to observed water balance fluxes and streamflow over large basins (Livneh et al., 2013; Maurer et al., 2002). We know that statistical downscaling methods perform poorly when non-stationarity occurs between the calibration and validation periods~~ Gridded climate observations underpin hydrologic projections. They are used to calibrate the downscaling technique and the hydrologic model, serving as targets and inputs respectively. Gridded observations are commonly evaluated via comparison with station observations (Hutchinson et al., 2009; Werner et al., 2015), intercomparison with other gridded observations (Eum et al., 2014) or by using them to drive a hydrologic model and comparing outputs to observed water balance fluxes and streamflow over large basins (Livneh et al., 2013; Maurer et al., 2002). We know that statistical downscaling methods perform poorly when non-stationarity occurs between the calibration and validation periods (Maurer et al., 2013)~~, but methods used to create gridded observations do not always ensure against non-stationarity and could also affect the success of downscaling methods. Furthermore, stationarity in mean annual precipitation and temperature does not dictate stationarity in climatic or hydrologic extremes. 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 BCSD (Bürger et al., 2012a; Salathé, 2005; Werner, 2011). The pros and cons of this extended calibration period have not been fully evaluated. This investigation will help the hydrologic modelling community build a better evaluation system for gridded observations to ensure their strength not only for projections of mean monthly changes over large basins ~100,000 km^2, but also to extremes in basins as small as 500 km^2.~~

~~When used to make climate change projections, distributed hydrologic models such as VIC are best driven with gridded daily data, which is usually produced via gridded statistical downscaling techniques such as BCSD, CA and BCCA, three gridded methods that have been tested to date. Applying BCSD using minimum and maximum~~, but we haven't evaluated how apparent non-stationarity caused by natural climate variability (Huang et al., 2014; Maraun, 2012) is amplified or diminished with methods used to create gridded observations, which could also affect the success of downscaling methods. Furthermore, stationarity in mean annual precipitation and temperature does not dictate stationarity in climatic or hydrologic extremes. 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., 2014; Themeßl et al., 2011). The pros and cons of this extended calibration period have not been fully evaluated. This investigation will help the hydrologic modelling community build a better evaluation system for gridded observations to ensure their strength not only for projections of mean monthly changes over large basins ~100,000 km^2, but also to extremes in basins as small as 500 km^2.

When used to make climate change projections, distributed hydrologic models such as VIC are best driven with gridded daily data, which is usually produced via gridded statistical downscaling techniques such as BCSD, CA and BCCA, three gridded methods that have been tested to date. 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).~~(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 is not the focus of this study. A few other methods have been developed recently that warrant investigation. These include double bias corrected constructed analogue (DBCCA), which is similar to BCCA but applies a second quantile mapping bias correction as a post-processing step to correct drizzle and other residual biases ~~-~~(Maurer et al., 2010). Additionally, the climate imprint delta method (CI) ~~(Hunter and Meentemeyer, 2005)~~(Hunter and Meentemeyer, 2005) and the “reverse” BCSD ~~(similar to SDBC in Ahmed et al., 2013; and AR in Gutmann et al., 2014)~~(similar to SDBC in Ahmed et al., 2013; and AR in Gutmann et al., 2014), which we refer to as bias corrected climate imprint (BCCI) due to its use of CI for interpolation, have not been explored for their applicability to hydrology. A recently developed hybrid of BCCA and BCCI, referred to as BCCAQ (Murdock et al., 2013, 2014) has the potential to be an improvement versus other gridded statistical downscaling techniques and has not been tested with hydrologic extremes. This work will also help to inform use of the resulting BCSD and hydrologic model output provided by the Pacific Climate Impacts Consortium (PCIC)[1]. Finally, PCIC also makes available Canada-wide downscaled climate change projections using both BCSD and BCCAQ methods[1]. This study provides the first rigorous intercomparison of these two methods.

The ClimDEX indices are recommended by the World Meteorological Organization Expert Team on Climate Change Detection and Indices (ETCCDI) ~~(Zhang et al., 2011)~~(Zhang et al., 2011) as a means of summarizing daily temperature and precipitation statistics focusing particularly on aspects of climate extremes. They have been developed to allow seamless

[1] http://www.pacificclimate.org/data

Formatted: German (Germany)

comparison of climate conditions on an international basis. There are many projects applying the ETCCDI indices to detect changes in extremes historically (e.g. Sillmann et al. 2013a), to project future changes (e.g. Sillmann et al. 2013b) and to provide future changes via data portals to allow local analysis (http://www.cccma.ec.gc.ca/data/climdex/). Two commonly investigated hydrologic extremes include 3-day peak flow, which represents potential flood conditions and 7-day low flow, which represents potential drought conditions (e.g. Maurer et al. 2010). Floods can be damaging to river and floodplain infrastructure, while droughts can be detrimental for human water use and aquatic habitat. We follow the framework developed by Bürger et al. (2012a), evaluating methods for their abilities in producing the temporal sequencing and distributional properties of climate indices and hydrologic extremes.

The objectives of this study are:

1) To compare several reanalyses in the study region against two gridded observation datasets.
2) To test the ability of BCCA, DBCCA, BCCI, CI, BCSD (mean temperature), BCSDX (minimum and maximum temperature) and BCCAQ downscaling techniques to simulate 26 ClimDEX indices using four reanalyses and two gridded observations.
3) To test the ability of BCCA, DBCCA, BCCI, CI, BCSD (mean temperature), BCSDX (minimum and maximum temperature) and BCCAQ downscaling techniques, when used to force the VIC hydrologic model, to simulate 3-day peak flow and 7-day low flow indices using four reanalyses and two gridded observations.
4) To learn more about the strengths and weaknesses of two gridded observations for use with hydrologic modelling.
5) To see if strength of a method to downscale for climate extremes relates to abilities for use with hydrologic extremes.

2 Study Area

The Peace River basin will be the focus of this work. The snow-dominated regime of this basin makes the findings of this work applicable to many mid-latitude areas. The Peace River is located in interior north-eastern BC and encompasses the 101,000 km^2 drainage area upstream of Taylor, BC (Figure 1). Elevations range from 400 m to 2800 m. The region is highly influenced by the Pacific Ocean and Arctic air masses. The region has a continental climate (Demarchi, 1996), with monthly average temperatures ranging from -12.0°C in

January to 12.3°C in July, averaging 0.2°C. Precipitation follows a seasonal pattern of summer maximum and spring minimum. The Peace River has a nival regime, with approximately 54% of the annual precipitation (440 mm) falling as snow (mostly during October–April) and 64% of the natural streamflow occurring during the freshet months of May–July. Low flows occur during the winter and early spring in head-water (INGEN) and downstream (BCGMS) basins (Figure 2). Due to the topographical complexity and strong climate gradients this region provides a stringent test of downscaling techniques. Additionally, the Peace River basin is the focus of two studies that explore uncertainty in hydrologic projections, one due to GCMs, emissions scenarios and parameter sets ~~(Bennett et al., 2012)~~(Bennett et al., 2012), the other due to statistically versus dynamically downscaled GCMs ~~(Shrestha et al., 2014a)~~(Shrestha et al., 2014a). This study provides a good complement to these by exploring new sources of uncertainty in the same basin.

3 Methods

3.1 Gridded observations

Two daily, gridded observational datasets were available over the study area. The first was generated for BC for application with the Variable Infiltration Capacity (VIC) macro-scale distributed hydrologic model following the methods of Maurer et al. (2002) and Hamlet and Lettenmaier (2005). Daily gridded surfaces of minimum and maximum temperature and daily precipitation accumulation were produced at the spatial resolution of 1/16°, which is ~6 km^2 depending on latitude, for January 1950 to December 2005. Station data were contributed from multiple networks including those of Environment Canada, BC Ministry of Forests, Lands and Natural Resource Operations, BC Hydro, and the US National Weather Service Co-operative observer program, each with a varying range of quality control. Stations were interpolated to grids using the SYMAP inverse-distance weighting algorithm (Shepard, 1984). The raw gridded fields were temporally homogenized to remove interpolation artefacts introduced by using a temporally varying mix of stations and corrected for topographic effects using ClimateWNA, a 1961 to 1990 PRISM-based high-resolution climatology for western North America ~~(Daly et al., 1994; Wang et al., 2006)~~(Daly et al., 1994; Wang et al., 2006). This dataset is referred to as VIC Forcings.

The second dataset was created for all of Canada using the Australian National University Spline (ANUSPLIN) implementation of trivariate thin plate smoothing splines ~~(Hutchinson et~~

~~al., 2009)~~(Hutchinson et al., 2009). The Canada-wide ANUSPLIN observational dataset was created at a 1/12° grid spacing (~10 km) for daily minimum temperature, maximum temperature, and precipitation amounts for the period 1950-2010 by Hopkinson et al., ~~(2011)~~(2011) and McKenney et al., ~~(2011).~~(2011). Station data from Environment Canada observing sites were interpolated onto the high-resolution grid using the ANUSPLIN smoothing splines with elevation, longitude, and latitude as interpolation predictors. Precipitation occurrence and square-root transformed precipitation amounts were interpolated separately on each day, combined, and transformed back to original units. Observed station data were quality controlled and corrected for station relocation, changes in the definition of the climate day and trace precipitation amounts~~, and precipitation gauge under-catch~~.

3.2 Reanalyses

Four atmospheric reanalysis products were selected to span a range of complexity and spatial resolution. Chosen methods include NCEP1, European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis 40 (ERA40), ECMWF Re-Analysis Interim (ERAInt) and the National Oceanic and Atmospheric Administration – Cooperative Institute for Research in Environmental Sciences 20th Century Reanalysis V2 (20CR). NCEP1 is a popular reanalysis product applied in the validation of statistical downscaling techniques ~~(Bürger et al., 2012a; Maurer et al., 2010)~~(Bürger et al., 2012a; Maurer et al., 2010). It spans the period from 1948 to present, is ~1.9° in resolution and includes a wide range of observations assimilated from ships to satellite data (Kalnay et al., 1996). ERA40 is available from 1958 to 2002 and is archived at the coarsest resolution (2.5°) of the four products selected for this study. It was the first to assimilate satellite radiance data directly ~~(Uppala et al., 2005)~~(Uppala et al., 2005). ERAInt covers the satellite era from 1979 through to present. Data used here are archived at 1.5°, although the underlying forecast model runs at 0.75°. It has an improved atmospheric model and assimilation system over that used in ERA40 ~~(Dee et al., 2011)~~(Dee et al., 2011). The 20CR is one of the longest reanalysis records available, starting in 1871 and running to 2012. At 2° resolution it assimilates only surface observations of synoptic pressure, monthly sea surface temperature and sea ice distribution ~~(Compo et al., 2011)~~(Compo et al., 2011). Table 1 summarizes the availability of the gridded observations and reanalyses.

3.3 Downscaling techniques

Seven statistical approaches are selected based on their wide use and/or potential strength in downscaling coarse-scale models to gridded observations for representing extremes. BCSD has been applied across North America ~~(Maurer and Hidalgo, 2008; Salathé, 2005; Schnorbus et al., 2014; Wood et al., 2002, 2004)~~(Maurer and Hidalgo, 2008; Salathé, 2005; Schnorbus et al., 2014; Wood et al., 2002, 2004). Monthly minimum temperature, maximum temperature and precipitation from GCMs or reanalyses are bias corrected, using quantile mapping, against gridded observations aggregated to the large-scale model grid. Bias corrected, spatially disaggregated monthly data are temporally disaggregated to a daily time step via random sampling of historical months. Days in the selected month are rescaled (multiplicative for precipitation and additive for temperature) to match the bias corrected monthly precipitation and average temperature~~.~~ (Figure 3a). Two variations of BCSD are tested; one derives minimum and maximum temperature from mean temperature in the coarse-scale model by assuming a uniform monthly diurnal temperature range (BCSD), the other uses monthly minimum and maximum temperature directly from the large-scale model (BCSDX).

Two constructed analogue downscaling approaches are tested; BCCA and DBCCA (Maurer et al., 2010). BCCA bias corrects the large-scale temperature and precipitation using quantile mapping, as in BCSD, except on daily rather than monthly large-scale data. In the constructed analogue (CA) component, a library of observed daily coarse-resolution and corresponding high resolution climate patterns of the variable to be downscaled is built (Hidalgo et al., 2008). Daily data are downscaled by selecting 30 days from the coarse-scale library that have the closest similarity to a given simulated day; optimal weights are determined via ridge regression and the 30 corresponding fine-scale library patterns are combined using the same weights (Maurer et al., 2010). In the DBCCA technique, a second quantile mapping bias correction is then applied at the fine-scale to fix drizzle and other biases caused by the linear combination of daily fields in the CA step~~.~~ (Figure 3a).

Two climate imprint methods are tested; CI delta method ~~(Hunter and Meentemeyer, 2005)~~(Hunter and Meentemeyer, 2005) and bias corrected CI (BCCI), which applies quantile mapping to the interpolated series from CI~~.~~ (Figure 3a). For the imprint methods, long-term averages (i.e. 30 years) from the fine-scale data provide a “spatial imprint” that is used to represent environmental gradients. The ratio of a daily to average monthly values is multiplied by the fine-scale monthly values for a location to get the daily precipitation. This is similar for minimum and maximum temperature, except values are calculated as the difference between the monthly mean and the daily value ~~(Hunter and Meentemeyer, 2005).~~(Hunter and Meentemeyer, 2005).

While BCCI applies quantile mapping as a post-processing step to the interpolated fine-scale outputs from the CI method, BCCAQ is a post-processed version of BCCA where the final quantile mapping bias correction is based on BCCI. First, the BCCA and BCCI algorithms are run independently, and then BCCAQ corrects BCCA with BCCI. The daily BCCI outputs at each fine-scale grid point are reordered within a given month according to the daily BCCA ranks. Because the optimal weights used to combine the analogues in BCCA are derived on a day-by-day basis, without reference to the full historical dataset, the algorithm may be prone to "Huth's paradox", wherein models that are calibrated based on short-term variability may be biased and fail to produce realistic long-term trends (Benestad et al., 2008; Huth, 2004). Reordering data for each fine-scale grid point within a month effectively breaks the overly smooth representation of sub reanalysis-grid scale spatial variability inherited from BCCI ~~(Maraun, 2013)~~(Maraun, 2013) thereby resulting in a more accurate representation of event-scale spatial gradients; this also prevents the downscaled outputs from drifting too far from the BCCI long-term trend. Over longer time-scales, the spatial variability of BCCAQ converges to that of BCCI.

Statistical methods are calibrated from 1950 to 1990 for 20CR and NCEP1, and from 1958 to 1990 and from 1979 to 1990 for ERA40 and ERAInt, respectively (Table 2). Calibration periods were selected to include the longest overlapping record between the gridded observation and reanalyses to replicate the approach taken in Werner et al. (2011). Thus, the 20CR and NCEP1 reanalyses results will serve to evaluate the gridded observations and these two reanalyses, and also validate the calibration/validation approach taken with BCSD for a series of studies conducted in this region ~~(Bürger et al., 2012a, 2012b; Schnorbus et al., 2014; Shrestha et al., 2012; Werner et al., 2013).~~(Bürger et al., 2012a, 2012b; Schnorbus et al., 2014; Shrestha et al., 2012; Werner et al., 2013). The resulting modelling framework for these two gridded observations, four reanalysis products and seven gridded statistical downscaling techniques is displayed in Figure ~~3~~3b. All statistical downscaling methods use precipitation and temperature as predictors and predictands.

3.4 ClimDEX

ClimDEX is a common climate indices package that computes values for 27 core indices based on daily precipitation, minimum and maximum temperature (Karl et al., 1999; Peterson, 2005) and (http://etccdi.pacificclimate.org or http://www.clivar.org/panels-and-working-groups/etccdi/etccdi.php/). These indices describe events, such as the number of heavy precipitation days denoted as days where precipitation is greater than 10 mm or percentage of days when maximum temperature is greater than the 90th percentile. They do not usually represent the most extreme events conceivable, but instead represent "the more extreme aspects of climate," which are known to be relevant to a broad range of impact fields and are still statistically manageable so that they can be reliably estimated from current data for the present and future. ClimDEX has been adopted as a standard for extremes by the World Climate Research Programme (http://www.clivar.org/organization/extremes). Indices were computed from downscaled temperature and precipitation from seven statistical downscaling methods used with four reanalyses and two gridded observations for a total of 56 estimates of each index. The index of the annual count when daily minimum temperature is > 20ºC, tropical nights (*tr*), was dropped for this analysis because this temperature threshold is not exceeded in the Peace River basin. See Table 1 in Bürger et al. ~~(2012a)~~(2012a) for a description of indices explored in this study.

3.5 Hydrologic Modelling

Hydrologic projections for the Peace River basin are derived using the Variable Infiltration Capacity (VIC) model ~~(Liang et al., 1994, 1996)~~(Liang et al., 1994, 1996). The VIC model is a spatially distributed macro-scale hydrologic model that was originally developed as a soil-vegetation atmosphere transfer scheme for general circulation models. It has been used to evaluate climate change impacts on global river systems (Nijssen et al., 2001) and in the mountainous western United States and BC ~~(Elsner et al., 2010b; Hamlet and Lettenmaier, 2005, 2007; Schnorbus et al., 2014; Shrestha et al., 2012)~~(Elsner et al., 2010b; Hamlet and Lettenmaier, 2005, 2007; Schnorbus et al., 2014; Shrestha et al., 2012). Its spatially distributed nature makes it suitable for capturing regional variation in the hydrologic cycle due to topographic, physiographic, and climatic controls. The VIC model is also process-based allowing for a more plausible extrapolation of hydrologic processes into future climate regimes ~~(Leavesley, 1994).~~(Leavesley, 1994). The VIC model is applied at a resolution of 1/16° (approximately 27-31 km^2, depending upon latitude) and run at a daily time step (onehour time step for the snow model). Surface routing between grid cells is done using the linearized Saint-Venant equations ~~(Lohmann et al., 1996)~~(Lohmann et al., 1996).

The Finlay River above Akie River, Ingenika River above Swannell River, Parsnip River above Misinchinka River, and Peace River above Pine River sub-basins of the Peace River were calibrated to observations from Water Survey of Canada (Figure 1). Peace River at Bennett Dam was calibrated to naturalized flow provided by BC Hydro. The sub-basins range in drainage area from 4,200 km^2 to 83,900 km^2 and from a minimum elevation of 392 m to a maximum of 2799 m (Table 3). All selected basins had strong calibration results over 1990 to 1995 for both the VIC Forcings and ANUSPLIN gridded observations based on the Nash-Sutcliff Efficiency score ~~(Nash and Sutcliffe, 1970)~~(Nash and Sutcliffe, 1970), the Nash-Sutcliff Efficiency score of the log-transformed discharge and the percent volume bias error (Table 4). Nash-Sutcliff Efficiency score values improved, Nash-Sutcliff Efficiency score of the log-transformed discharge stayed roughly the same and percent volume bias error differences became larger in magnitude in the 1985 to 1989 split-sample validation period, negative in VIC Forcings and positive in ANUSPLIN.

There are several daily streamflow metrics that are useful for water resources design and management, which are also ecologically relevant ~~(Monk et al., 2011; Richter et al., 1996; Shrestha et al., 2014b)~~(Monk et al., 2011; Richter et al., 1996; Shrestha et al., 2014b). A recent intercomparison of statistical downscaling techniques for use with daily streamflow investigated the hydrologic extremes 3-day peak flow and 7-day low flow (Maurer et al., 2010). To build on that study we investigate the strength of seven downscaling methods for the same metrics using 3-day peak flow to represent flood and 7-day low flow, drought. Two low flow periods are investigated because the lowest discharge takes place in the months of October to April in sub-basins of the Peace River (Figure 2) and summer low-flows (July to September) are of interest to agriculture and ecology. Hydrologic models can have low flows in different seasons than observations due to their poor parameterization of baseflow conditions and because calibration approaches favour good performance for peak flow ~~(Najafi et al., 2011)~~(Najafi et al., 2011). This issue can be exaggerated by downscaling approaches (Shrestha et al., 2014b). Thus, narrowing the window over which low flows are accessed is important to prevent low flows in one season being compared to low flows in another. Peak flows are analyzed between May and July.

Field Code Changed

3.6 Statistical Tests

The seven statistical downscaling methods vary in their approach, which can result in differing strengths and weakness. We chose our statistical tests to fully evaluate these downscaling techniques for the climate and hydrologic results and to follow the framework of Bürger et al. ~~(2012a)~~(2012a). The time period for calibration of the downscaling techniques was selected to match Bürger et al. ~~(2012a) (pre-1991 depending on the availability of the reanalyses).~~(2012a) (pre-1991 depending on the availability of the reanalyses). Longer calibration periods available for NCEP1 and 20CR were also seen as favourable when applying bias correction based downscaling methods, especially when working with extremes (Huang et al., 2014; Themeßl et al., 2011) and assisted with evaluating the two gridded observations. Validation was set to 1991-2005 to accommodate the overlap of available reanalyses, gridded observations and observed streamflow records. ERA40 is an exception with the last full year of available record for 2001. Validation results for ERA40 are provided for 1991-2001.

Two statistical tests are applied to the ClimDEX results over the Peace River basin: the Kolmogorov-Smirnov (KS) test and the test for Pearson's correlation. The KS test is used to see how well the distribution of climate indices for the statistically downscaled reanalyses match the distribution of those calculated from the gridded observations used as downscaling targets. ~~Pearson's correlation is used to test the correlation between the annual climate indices for the statistically downscaled reanalyses and the associated gridded observation. The KS and correlation tests are conducted on each of the grid cells in the Peace River basin and then the Walker field significance test (Wilks, 2006) is applied to see if the KS test and Pearson's correlation results are significant over the basin as a whole~~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). 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.

~~The same two tests~~The 101,000 km^2 Peace River basin is represented by 3975 grid cells at the 1/16° resolution used to run the VIC hydrologic model. The KS test and Pearson's correlation are evaluated 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 field significance test (Storch, 1982) 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.

The KS test and the test for Pearson's correlation were applied on the 3-day peak flow and 7-day low flow in winter and summer for hydrologic data from the five sub-basins of the Peace River. In this case, the KS test is used to test how well the distribution of the hydrologic extremes created by driving the VIC model with the statistically downscaled reanalyses matched those derived from driving the VIC model with the two gridded observations. Pearson's correlation is used to test the ~~correlation~~temporal correspondence between the hydrologic extremes created by driving VIC with downscaled reanalyses versus gridded observations.

4 Results

4.1 Gridded observations and Reanalyses

Four reanalyses (NCEP1, ERA40, ERAInt and 20CR) are compared to two gridded observations (VIC Forcings and ANUSPLIN) over the Peace River basin. Daily precipitation, minimum temperature and maximum temperature are converted to total monthly precipitation and average monthly temperatures over the 1950 to 2005. Average minimum and maximum temperatures in ANUSPLIN and VIC Forcings are similar from year to year in most months (Figure 4 and Figure 5). However, prior to 1970, ANUSPLIN can be up to five degrees cooler than the VIC Forcings and reanalyses from December to February. Precipitation totals are similar from year to year for all months in the two gridded observations, except October when precipitation difference can be up to 50 mm (Figure 6). This could be because there is greater station coverage in the VIC Forcings and an elevation adjustment is made with ClimateWNA. Differences in these two products resulting from these factors might be more apparent in the shoulder season.

There is a warm bias in minimum temperature in 20CR and ERA40 from May to November and a cool bias in NCEP1 from March to October relative to gridded observations (Figure 4). The biases in NCEP1 tend to be greater over part of the record in some months, such as from 1970 to ~1995 in June. ERAInt is closest to gridded observations for minimum temperature, but is only available after 1979. Some of the patterns seen in minimum temperature are repeated in maximum temperature (Figure 5). NCEP1 values are noticeably cooler than observations and other reanalyses in May, June, July, September and October in some years. In April, maximum temperature in 20CR and NCEP1 are close to each other and roughly five degrees less than the other reanalyses and gridded observations. Maximum temperatures for ERA40 and ERAInt are closest to gridded observations from year to year in all months. Monthly precipitation in the NCEP1 and ERA40 reanalyses has similar magnitudes and variability as the gridded observations (Figure 6). ERAInt is close to observations in the fall and winter months, but has higher precipitation values in March through August. 20CR stands apart from the other reanalyses and both gridded observations with consistently larger precipitation amounts, roughly twice the magnitude as observations in September through April. However, sequencing of events is similar between 20CR and observations.

This confirms that near surface temperature and precipitation values from the selected reanalyses have different characteristics due to their different resolutions, model physics and contributing data in the Peace River basin. The two gridded observations also displayed some dissimilarity in time. Differences between these four reanalyses in this particular region should act as a stringent test of the downscaling techniques applied. However, we expect that the time-dependent differences between gridded observations and NCEP1 for minimum and maximum temperature, and precipitation, will reduce the success rate of any of the downscaling techniques (Maurer et al., 2013). Nevertheless, we carry NCEP1 through the analysis to quantify the impacts of using a potentially flawed reanalysis and also to evaluate VIC Forcings and ANUSPLIN over their full record (1950 to 2005) with two reanalyses (NCEP1 and 20CR).

4.2 Impact of downscaling approach and reanalyses on ClimDEX results

Downscaled minimum temperature, maximum temperature and precipitation from seven gridded downscaling methods, two gridded observations and four reanalyses were used to generate 26 ClimDEX indices. Results were compared to the indices generated from the respective gridded observations at their native resolution (VIC Forcings (~6 km) and ANUSPLIN (~10 km)) for their ability to match the timing (Pearson's correlation) and distribution (KS test) of values over the Peace River basin using the Walker field significance test ~~(Wilks, 2006)~~(Wilks, 2006).

In the calibration (1950-1990) and validation (1991-2005) periods the VIC Forcings and ANUSPLIN dataset are similar for most temperature based indices and show some large differences for precipitation based indices (Table 5). Namely, PRCPTOT, annual total wet day precipitation (> 1 mm), in ANUSPLIN is 18% and 21% less than VIC Forcings in the calibration and validation periods, respectively. The events on a given day are larger in VIC Forcings than ANUSPLIN as shown by the higher R95p, RX1day, RX5day, R10mm and R20mm values. Between the validation and the calibration period PRCPTOT increases more in VIC Forcings than in ANUSPLIN. The increase in VIC Forcings comes from an increase in precipitation days (R1mm) rather than an increase in intensity. Magnitudes of the larger precipitation events actually decrease for VIC Forcings while they increase for ANUSPLIN, although these events are still larger in VIC Forcings than ANUSPLIN in the validation period. The percentage of cool nights decrease and the duration of warm spells increase somewhat equally for both gridded observations. However, increases in the percentage of warm days and warm nights, and decreases in the percentage of cool days and duration of cold spells, are greater in ANUSPLIN than VIC Forcings, which suggests that the warming signal in ANUSPLIN is stronger. Statistically significant increases in annual minimum temperatures were found by Rodenhuis et al. (2009) in this region. Differing trends in climate extremes are common in gridded observations due differences in stations, interpolation techniques and potential corrections for temporal inhomogeneity. Donat et al. ~~(2014)~~(2014)

found that decadal trends in maximum 5 day precipitation amounts (Rx5day) over 1979-2008 ranged from -15 to 5 mm/decade in the Peace River basin region depending on the gridded observations they studied. VIC Forcings included a monthly temporal adjustment to increase homogeneity (Hamlet and Lettenmaier, 2005), while ANUSPLIN did not. Additionally, stations were allowed to drop in and out on a daily bases in ANUSPLIN, whereas stations had to be available for a minimum of one year of consecutive days and five years over the record to be included in VIC Forcings. Hence, trends in some climate extremes differ for these gridded observations and may or may not match those of "reality" and/or reanalyses.

Irrespective of downscaling method or reanalysis, those methods calibrated and validated against the ANUSPLIN gridded observations were more successful ~~versus~~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.) The largest differences in the number of tests passed primarily occur for precipitation based indices where ANUSPLIN passes more than VIC Forcings. VIC Forcings passes 29 more tests than ANUSPLIN for DTR (Table 7). This result is not unexpected because the differences between the calibration and validation period are precipitation related in VIC Forcings and temperature related in ANUSPLIN (Table 5). Step changes in daily temperature range (DTR) from 1950 to 2005 are apparent in ANUSPLIN (Figure 8). DTR is a strong driver of snow pack generation and melt and errors in simulating realistic DTR could affect hydrologic modelling results.

The sequencing of precipitation indices, such as CWD, PRCPTOT, R10mm, R20mm, R95p, R99p, Rx1day, Rx5day and SDII, is most difficult to replicate for all methods, especially under VIC Forcings. VIC Forcings has a higher station density than ANUSPLIN because it includes stations from BC Hydro, the BC Ministry of Forests Lands and Natural Resource Operations, and the Ministry of Environment's BC River Forecast Centre Snow Survey Network in addition to those from Environment Canada (Werner et al., 2015). The BC Hydro network provided a large number of stations in the Peace River basin, most of which were not available until the 1980s (Werner et al., 2015). The increase in the number of stations after

1980 in the VIC Forcings likely resulted in more complex spatial patterns in precipitation, despite the monthly temporal adjustment, because it is designed to maintain spatial variability (Hamlet and Lettenmaier, 2005). Increased spatial variability in the validation period, coupled with a different interpolation method in VIC Forcings, could have made precipitation patterns harder to replicate with downscaling. If we are going to rely on these datasets to investigate changes to extreme climate and hydrology we should develop a way to maintain temporal and spatial homogeneity for daily values, while allowing datasets to reflect natural trends. Minimizing homogeneity problems throughout the record is favourable when using gridded observations to calibrate statistical downscaling methods ~~(Gutmann et al., 2014; Livneh et al., 2013; Maurer et al., 2002)~~(Gutmann et al., 2014; Livneh et al., 2013; Maurer et al., 2002).

Considering results for all downscaling methods and both gridded observations, results based on ERAInt had the highest score of all four reanalyses for the Pearson's correlation and KS tests combined (Table 6). ERAInt results matched sequencing of events most often according to the Pearson's correlation test (Figure 7; Table 8) and ERA40 results matched distributions most often according to the KS test (Figure 9; Table 8). ERAInt passed the correlation test for both gridded observations for the number of heavy precipitation days (R10mm) when the other reanalyses did not (Figure 7). ERA40 and ERAInt monthly average minimum and maximum temperature and total precipitation matched those of the gridded observations most closely (see Section 4.1). ERAInt is the highest resolution (1.5°) and both ERAInt and ERA40 excluded 1950-1958 in their calibration when NCEP1 and 20CR did not (Table 2), which may have avoided potential problems with the gridded observations caused by lower station availability earlier in the record and with reanalysis data from the pre-satellite era (1979-) and before the expansion and standardization of global radiosonde network (1958-). Results for SDII for VIC Forcings and ANUSPLIN under all seven downscaling methods show large differences between gridded observations and downscaled NCEP1 prior to 1958 (Figure 10). Gutmann et al. ~~(2014)~~(2014) tested four downscaling methods with NCEP1 focusing on the period containing satellite microwave and infrared atmospheric soundings (1979-) and still found temporal instabilities in NCEP1 contributed to failure in downscaling techniques for some metrics. Root mean square error in sea level pressure decreases from 1950 to 2008 strongly in NCEP1, somewhat in ERA40 and minimally in 20CR (see Figure 10 in ~~(Compo et al., 2011)~~(Compo et al., 2011)). Assimilating only surface pressure reports and using observed monthly sea-surface temperature and sea-ice distributions as boundary conditions to create 20CR has resulted in a more temporally consistent product. However, it still has improved over time. Changes in 20CR in combination with changes in the gridded observations over 1950-2005 has resulted in fewer passed tests for 20CR than ERA40 or ERAInt. Thus, choice of reanalysis, calibration period and the gridded observation dataset can influence the measured success of the downscaling approach being tested, irrespective of the method's inherent strengths and weaknesses.

The highest ranked downscaling method based on the combined results for field significance of Pearson's correlation and KS test for all gridded observations, reanalyses and ClimDEX indices was DBCCA (Table 6). It tied for highest rank with CI for correlation, while BCCAQ superseded all other methods for distribution. Bias remains in results of the BCCA method for precipitation due to the linear combination of fine-scale analogues and uncorrected "drizzle" and related biases (Guttmann et al., 2014). All downscaling methods, except CI, include a quantile mapping bias correction step and are expected to do well in matching distributions with their respective gridded observation. All methods except CI pass 86% or more of the tests for distribution (KS test), while CI passes 78%. The correlation of DTR was a problem for all the downscaling methods and both gridded observations (Figure 8) and for distribution based on ANUSPLIN (except BCCAQ) but not when based on VIC Forcings. BCCAQ in combination with ANUSPLIN matched DTR distributions for ERAInt, ERA40 and 20CR when all other methods failed, which points to the success of its approach of post-processing BCCA with a final quantile mapping bias correction based on BCCI. As mentioned above DTR is an important driver in snowpack. Additionally it plays a key role in evaporation ~~(Sheffield et al., 2012)~~(Sheffield et al., 2012). Rates of evaporation are an important component of projecting future water availability and drought ~~(Sherwood and Fu, 2014).~~(Sherwood and Fu, 2014). Therefore, accurately downscaling DTR should be a priority. Including minimum and maximum monthly temperature predictors in BCSDX did not improve the correlation of DTR as was hypothesized in previous studies (Bürger et al., 2012a).

Field Code Changed

4.3 Impact of downscaling approach and reanalyses on hydrologic extremes

The previous section shows how raw reanalyses and observations differ in the Peace River basin and how downscaled reanalyses can differ in their representation of climate extremes when calibrated to one gridded observation versus another. NCEP1 has routinely been used to compare the performance of statistical downscaling methods in terms of climate and hydrologic extremes (e.g. Bürger et al. ~~(2012a)~~(2012a) and Maurer et al. (2010)). We thus continue our comparison of multiple gridded observations, reanalyses and downscaling techniques for hydrologic extremes. Results are compared for 15 years from 1991 to 2005 (inclusive) for the five sub-basins, except for ERA40 (11 years; 1991 to 2001). We evaluate methods for their ability to replicate the timing (Pearson's correlation) and distribution (KS test) of the 3-day peak flow, 7-day low flow in summer and 7-day low flow in winter.

Irrespective of reanalysis or downscaling method, VIC hydrologic model simulations based on the VIC Forcings gridded observations passed 8% more tests than those based on the ANUSPLIN gridded observations (Table 9), whereas for the ClimDEX indices ANUSPLIN passed 7% more tests than VIC Forcings (Table 6). The difference in the number of test passed is not great. Therefore, the success of the downscaling methods does not depend strongly on which of the gridded observations is applied overall. However, the greater number of test passed for hydrologic modelling with the VIC Forcings gridded observations could relate to VIC Forcings being created at the native resolution of the VIC hydrologic model (1/16°) whereas the ANUSPLIN data was created at 1/12° and remapped to 1/16° using bilinear interpolation. Additionally, a larger precipitation bias correction was required during calibration with the ANUSPLIN data than the VIC Forcings data suggesting that ANUSPLIN precipitation is less representative than VIC Forcings. Out of the two statistical tests and three metrics the only case where ANUSPLIN passed more tests than VIC Forcings was for correlation in summer 7-day low flow (Table 10), especially when driven with NCEP1 and 20CR downscaled via BCCA and DBCCA. Similar results were found for ANUSPLIN and BCCA and DBCCA with the ClimDEX indices (Section 4.2). This suggests that there is potential for ClimDEX results to act as predictor of hydrologic extremes.

When considering results regardless of gridded observation or downscaling technique the number of tests past under ERA40 was the highest overall (Table 9). Additionally, the number of tests passed for the Pearson's correlation and the KS test were both highest for ERA40. The truncated validation period for ERA40, 1990-2001 versus 1990-2005 for other reanalyses, could have avoided some challenging hydrologic extreme events in 2002-2005. However, ERAInt, which was validated over 1990-2005, passed nearly the same number of tests as ERA40. Thus, the shorter calibration period in ERA40 and ERAInt avoids step changes in the gridded observations and reanalyses prior to 1958. Peculiarities with the gridded observations were apparent from 1950 to 1958 for the monthly average minimum and maximum temperature (Figure 4 and Figure 5) and for the DTR and SDII ClimDEX indices

(Figure 8 and Figure 10). Avoiding these years could have reduced artefacts in the downscaled products and hydrologic model results. Nevertheless, many studies have demonstrated that ERA40 and ERAInt are superior products versus NCEP1 ~~(Donat et al., 2014; Ma et al., 2008, 2009; Sillmann et al., 2013)~~(Donat et al., 2014; Ma et al., 2008, 2009; Sillmann et al., 2013). In our own analysis ERA40 and ERAInt have similar timing and magnitude in minimum and maximum temperature and precipitation (Figures 4, 5 and 6) as the gridded observations when NCEP1 and 20CR do not. These results confirm that downscaling methods will succeed when applied to reanalyses that have correct timing, magnitude and trends such as ERA40 and ERAInt more so than when applied to reanalyses such as NCEP1 and 20CR that have irregular step changes ~~(Maraun, 2013).~~(Maraun, 2013). We should be able to assume that although the biases in GCMs will be greater than those found in reanalyses they are consistent over time. The strength of downscaling methods when downscaling ERA40 and ERAInt versus NCEP1 and 20CR was also found with the ClimDEX indices.

The BCCAQ method was the best overall performer for the three hydrologic extremes. It was the best method according to Pearson's correlation and tied for second place with DBCCA and BCCI, after BCSD and BCSDX, for the KS test. BCSD and BCSDX passed the fewest number of tests for correlation, while CI passed the fewest for distribution. In the case of ClimDEX, BCCAQ ranked third after BCCA and BCCI. The strength of the BCCAQ method when tested in terms of basin-wide hydrologic modelling and hydrologic extremes, rather than in terms of ClimDEX indices at individual grid cells, comes from the maintenance of daily spatial patterns resulting from the combination of BCCA and BCCI methods. Event-scale spatial gradients and magnitudes are preserved by reordering the BCCI outputs based on the rank order structure from BCCA. In effect, this removes the overly smooth representation of sub reanalysis-grid scale variability from BCCI (Maraun, 2013) and largely corrects remnant biases in magnitude from BCCA (Guttmann et al., 2014). Spatial covariability is much more relevant in hydrologic modelling than the comparison of climate indices between products on a grid cell to grid cell basis. This method is also better at maintaining long-term trends, which might explain failed tests in some of the sub-basins when downscaling NCEP1 and 20CR, which, as shown earlier, exhibit inhomogeneities between calibration and validation periods. BCCAQ could be failing for the "right reason" when the trend in VIC Forcings or ANUSPLIN for a given metric is opposite that in NCEP1 or 20CR. BCCAQ is the only method to pass the Pearson's correlation and KS test in all five sub-basins when downscaling ERA40 or ERAInt to VIC Forcings or ANUSPLIN for all three hydrologic extremes. BCCAQ has overcome some of the challenges of BCCA that Maurer et al. (2010) would not have been able to find using NCEP1 alone as surrogate GCM. It is also more successful than the BCCI method, which is analogous to the Statistical Downscaling and Bias Correction (SDBC) method in Ahmed et al. ~~(2013) and Asynchronous Regression (AR) in Gutmann et al.~~(2013) and Asynchronous Regression (AR) in Gutmann et al. (2014), by avoiding overestimates of extreme events at aggregate scales (Maraun, 2013).

Field Code Changed

The BCSD methods pass the most tests for distribution for all basins and reanalyses, while they fail more tests than any other downscaling method for correlation due to their reliance on random sampling of historical months when temporally disaggregating from the monthly to daily time step (Table 6). Thus, these methods will get the frequency and magnitude of events correct, but will get the timing of when these events occur wrong. Again including the minimum and maximum temperature from the large-scale model (reanalysis) does not improve the number of tests passed with BCSDX versus BCSD. For 3-day peak flow (Table 11; Figure 11) and 7-day low flow in summer (Table 10; Figure 12) these methods pass the majority of tests for correlation. Very few tests are passed for correlation in seven-day low flow in winter (Table 12; Figure 13). Winter low flows are challenging to monitor and to model. There could be ice on the river causing the stage-discharge relationships to be incorrect. Also, as mentioned models are not parametrized or calibrated to best represent base flow. However, BCSD and BCSDX have more trouble than any of the other downscaling methods. Due to the resampling of daily events from the historical gridded observations there can be precipitation occurring in combination with temperatures warm enough to generate runoff (Figure 14). This is because of the stochastic resampling of the historical precipitation, but is also related to temperature since runoff is occurring when conditions should be near freezing. Additionally, the random selection of months from the historical record can lead to large discontinuities across month boundaries, such as in December to January (Figure 14). This is when it is important to get daily events from the GCM or reanalyses (e.g., as in the CI, BCCI, BCCA, DBCCA and BCCAQ methods). As calibrated, the VIC model is known to have limited performance for low flows and additional errors were suspected to have been contributed by BCSD in downscaled 20C3M GCM results ~~(Shrestha et al., 2014b)~~(Shrestha et al., 2014b). Some sharp spikes on the rising limb of the hydrograph suggest rain-on-snow events caused by the downscaling-driven results that are not displayed in the runs based on gridded observations. The CI method is the closest to the delta method that we have investigated. The median and ranges for CI are much lower for winter 7-day low flow (not shown). The poorer performance of the CI method for the KS test is due to the lack of quantile mapping bias correction in this method.

5 Conclusions

We have tested the applicability of seven techniques for downscaling coarse-scale climate models in terms of ClimDEX indices and hydrologic extremes. The seven approaches investigated include several methods commonly used in hydrologic modelling. Some of these had been explored before (i.e. BCSD and BCCA), but not using multiple reanalyses. Choice of reanalysis was found to affect the number of tests passed for a given downscaling technique. Downscaling methods were more successful under ERA40 or ERAInt than they were under NCEP1 or 20CR. The quality of reanalyses and gridded observations changed over the calibration period due to changes in availability of satellite/radiosonde data and station observations. NCEP1, the reanalysis used as a surrogate GCM in many previous downscaling intercomparisons, had an obviously erroneous step change in temperature over the Peace River basin. Between the calibration and validation period, changes in ClimDEX indices were greater for precipitation with VIC Forcings but greater for temperature with ANUSPLIN. Thus, trends in ClimDEX indices differed in these gridded observations. ANUSPLIN passed 5% more tests than VIC Forcings, mostly for precipitation related ClimDEX indices. Through this work we learned a lot about these gridded observations and discovered evaluation procedures that will be useful for future studies.

BCSDX, DBCCA and BCCAQ downscaling methods had not been evaluated in terms of ClimDEX indices and hydrologic extremes before now. The BCSDX method included minimum and maximum temperature from the reanalyses instead of mean as is done in BCSD, but this did not improve its ability to resolve temperature indices, such as diurnal temperature range, or hydrologic extremes. DBCCA was an improvement over BCCA and passed the greatest number of tests for the ClimDEX indices. The double bias correction proved to reduce some of the drizzle and remnant bias in precipitation amounts found in BCCA. The BCCAQ method, which combines BCCA and BCCI, performed well in terms of number of tests passed for the ClimDEX indices, but it really shone for use with modelling hydrologic extremes. In this context, it exceeded all other methods. BCCAQ provides a more accurate representation of event-scale spatial gradients, removing the overly smooth representation of sub reanalysis-grid scale variability inherited from BCCI and correcting biases from BCCA. These attributes are important for simulating the climate events that occur over a basin that drive runoff. All methods passed correlation and distribution tests for 3-day peak flow and 7-day low flow in summer for the majority of sub-basins and reanalyses. BCSD and BCSDX failed all or most correlation tests and CI failed all or most distribution tests for 7-day low flow in winter. Based on results from this study, use of a daily downscaling method, such as BCCAQ, in conjunction with a rigorously constructed and validated observational dataset, is recommended to supplement the existing hydrologic modelling efforts at PCIC and improve projections of hydrologic extremes.

We can build on this work to develop tools that predict changes to hydrologic extremes from changes in climate extremes without the direct application of a hydrologic model. Similar emulations have been made by drawing on the relationship between GCMs and hydrologic model projections ~~(Schnorbus and Cannon, 2014)~~(Schnorbus and Cannon, 2014) and by identifying relationships between GCMs and RCMs ~~(Li et al., 2011)~~(Li et al., 2011). The next step is to identify which of the 26 ClimDEX indices are predictors of 3-day peak flow and 7-day low flow and avoid those downscaling methods that simulate them poorly.

Acknowledgements

We thank David Bronaugh for developing and making available the R package “climdex.pcic” that expedited this work. Dr Belaid Moa’s assistance with the Compute Canada/WestGrid/University Systems is also appreciated. Hailey Ekstrand provided useful assistance with hypsometric information for the five sub-basins of the Peace River basin upstream of Taylor. Thoughtful and thorough reviews were provided by Markus Schnorbus and Dr Francis Zwiers of PCIC that greatly improved this work. We appreciate the easy access to reanalyses provided by reanalysis.org and thank the centres that contributed NCEP1, 20CR, ERA40 and ERAInt. We also thank Dan McKenney of Natural Resources Canada for sharing the ANUSPLIN gridded observational product for Canada and Dr Katrina Bennett for constructing the VIC Forcings gridded observational product for British Columbia.

References

Abatzoglou, J.T., and Brown, T.J. (2012). A comparison of statistical downscaling methods suited for wildfire applications. Int. J. Climatol. *32*, 772–780.

Ahmed, K.F., Wang, G., Silander, J., Wilson, A.M., Allen, J.M., Horton, R., and Anyah, R. (2013). Statistical downscaling and bias correction of climate model outputs for climate change impact assessment in the U.S. northeast. Glob. Planet. Change *100*, 320–332.

Benestad, B.E., Hanssen-Bauer, I., and Chen, D. (2008). Chapter 8 in: Emperical-Statistical Downscaling (Singapore: World Scientific).

Bennett, K.E., Werner, A.T., and Schnorbus, M. (2012). Uncertainties in Hydrologic and Climate Change Impact Analyses in Headwater Basins of British Columbia. J. Clim. *25*, 5711–5730.

Bürger, G., Schulla, J., and Werner, A.T. (2011). Estimates of future flow, including extremes, of the Columbia River headwaters. Water Resour. Res. *47*, W10520.

Bürger, G., Murdock, T.Q., Werner, A.T., Sobie, S.R., and Cannon, A.J. (2012a). Downscaling extremes - an intercomparison of multiple statistical methods for present climate. J. Clim. *25*, 4366–4388.

Bürger, G., Murdock, T.Q., Werner, A.T., Sobie, S.R., and Cannon, A.J. (2012b). Downscaling extremes - an intercomparison of multiple methods for future climate. J. Clim.

Clavet-Gaumont, J., Sushama, L., Khaliq, M.N., Huziy, O., and Roy, R. (2013). Canadian RCM projected changes to high flows for Québec watersheds using regional frequency analysis. Int. J. Climatol. *33*, 2940–2955.

Compo, G.P., Whitaker, J.S., Sardeshmukh, P.D., Matsui, N., Allan, R.J., Yin, X., Gleason, B.E., Vose, R.S., Rutledge, G., Bessemoulin, P., et al. (2011). The Twentieth Century Reanalysis Project. Q. J. R. Meteorol. Soc. *137*, 1–28.

Cunderlik, J.M., and Ouarda, T.B.M.J. (2009). Trends in the timing and magnitude of floods in Canada. J. Hydrol. *375*, 471–480.

Cunderlik, J.M., and Simonovic, S.P. (2007). Inverse flood risk modelling under changing climatic conditions. Hydrol. Process. *21*, 563–577.

Cunderlik, J.M., Ouarda, T.B.M.J., and Bobée, B. (2004). On the objective identification of flood seasons. Water Resour. Res. *40*, W01520.

Daly, C., Neilson, R.P., and Phillips, D.L. (1994). A statistical-topographic model for mapping climatological precipitation over mountainous terrain. J. Appl. Meteorol. *33*, 140–158.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., et al. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. *137*, 553–597.

Demarchi, D.A. (1996). An introduction to the ecoregions of British Columbia.

Donat, M.G., Sillmann, J., Wild, S., Alexander, L.V., Lippmann, T., and Zwiers, F.W. (2014). Consistency of Temperature and Precipitation Extremes across Various Global Gridded In Situ and Reanalysis Datasets. J. Clim. *27*, 5019–5035.

Elsner, M.M., Cuo, L., Voisin, N., Deems, J.S., Hamlet, A.F., Vano, J.A., Mickelson, K.E.B., Lee, S.-Y., and Lettenmaier, D.P. (2010a). Implications of 21st century climate change for the hydrology of Washington State. Clim. Change *102*, 225–260.

Elsner, M.M., Cuo, L., Voisin, N., Deems, J.S., Hamlet, A.F., Vano, J.A., Mickelson, K.E.B., Lee, S.-Y., and Lettenmaier, D.P. (2010b). Implications of 21st century climate change for the hydrology of Washington State. Clim. Change *102*, 225–260.

Eum, H.-I., Dibike, Y., Prowse, T., and Bonsal, B. (2014). Inter-comparison of high-resolution gridded climate data sets and their implication on hydrological model simulation over the Athabasca Watershed, Canada. Hydrol. Process. *28*, 4250–4271.

Gleckler, P.J., Taylor, K.E., and Doutriaux, C. (2008). Performance metrics for climate models. J. Geophys. Res.-Atmospheres *113*, D06104–D06104.

Gutmann, E., Pruitt, T., Clark, M.P., Brekke, L., Arnold, J.R., Raff, D.A., and Rasmussen, R.M. (2014). An intercomparison of statistical downscaling methods used for water resource assessments in the United States. Water Resour. Res. *50*, 7167–7186.

Hamlet, A.F., and Lettenmaier, D.P. (2005). Production of Temporally Consistent Gridded Precipitation and Temperature Fields for the Continental United States. J. Hydrometeorol. *6*, 330–336.

Hamlet, A.F., and Lettenmaier, D.P. (2007). Effects of 20th century warming and climate variability on flood risk in the western U.S. Water Resour. Res. *43*.

Hidalgo, H.G., Dettinger, M.D., and Cayan, D.R. (2008). Downscaling with constructed analogues: daily precipitation and temperature fields over the United States.

Hofer, M., Marzeion, B., and Mölg, T. (2012). Comparing the skill of different reanalyses and their ensembles as predictors for daily air temperature on a glaciated mountain (Peru). Clim. Dyn. *39*, 1969–1980.

Hopkinson, R.F., McKenney, D.W., Milewska, E.J., Hutchinson, M.F., Papadopol, P., and Vincent, L.A. (2011). Impact of Aligning Climatological Day on Gridding Daily Maximum–Minimum Temperature and Precipitation over Canada. J. Appl. Meteorol. Climatol. *50*, 1654–1665.

Huang, S., Krysanova, V., and Hattermann, F.F. (2014). Does bias correction increase reliability of flood projections under climate change? A case study of large rivers in Germany. Int. J. Climatol. *34*, 3780–3800.

Hunter, R.D., and Meentemeyer, R.K. (2005). Climatologically Aided Mapping of Daily Precipitation and Temperature. J. Appl. Meteorol. *44*, 1501–1510.

Hutchinson, M.F., McKenney, D.W., Lawrence, K., Pedlar, J.H., Hopkinson, R.F., Milewska, E., and Papadopol, P. (2009). Development and Testing of Canada-Wide Interpolated Spatial Models of Daily Minimum–Maximum Temperature and Precipitation for 1961–2003. J. Appl. Meteorol. Climatol. *48*, 725–741.

Huth, R. (2004). Sensitivity of Local Daily Temperature Change Estimates to the Selection of Downscaling Models and Predictors. J. Clim. *17*, 640–652.

Janssen, A. (1994). Two-sample goodness-of-fit tests when ties are present. J. Stat. Plan. Inference *39*, 399–424.

Joshi, D., St-Hilaire, A., Daigle, A., and Ouarda, T.B.M.J. (2013). Databased comparison of Sparse Bayesian Learning and Multiple Linear Regression for statistical downscaling of low flow indices. J. Hydrol. *488*, 136–149.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., et al. (1996). The NCEP/NCAR 40-Year Reanalysis Project. Bull Amer Meteor Soc *77*, 437–471.

Karl, T.R., Nicholls, N., and Ghazi, A. (1999). CLIVAR/GCOS/WMO workshop on indices and indicators for climate extremes: Workshop summary. Climatic Change *42*, 3–7.

Knutti, R., Allen, M.R., Friedlingstein, P., Gregory, J.M., Hegerl, G.C., Meehl, G.A., Meinshausen, M., Murphy, J.M., Plattner, G.-K., Raper, S.C.B., et al. (2008). A Review of Uncertainties in Global Temperature Projections over the Twenty-First Century. J. Clim. *21*, 2651–2663.

Leavesley, G.H. (1994). Modeling the effects of climate change on water resources - a review. Clim. Change *28*, 159–177.

Li, G., Zhang, X., Zwiers, F., and Wen, Q.H. (2011). Quantification of Uncertainty in High-Resolution Temperature Scenarios for North America. J. Clim. *25*, 3373–3389.

Liang, X., Lettenmaier, D.P., Wood, E.F., and Burges, S.J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J. Geophys. Res. Atmospheres *99*, 14415–14428.

Liang, X., Wood, E.F., and Lettenmaier, D.P. (1996). Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification. Glob. Planet Change *13*, 195–206.

Livneh, B., Rosenberg, E.A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K.M., Maurer, E.P., and Lettenmaier, D.P. (2013). A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States: Update and Extensions*. J. Clim. *26*, 9384–9392.

Lohmann, D., Nolte-Holube, R., and Raschke, E. (1996). A large-scale horizontal routing model to be coupled to land surface parametrization schemes. Tellus A *48*, 708–721.

Ma, L., Zhang, T., Li, Q., Frauenfeld, O.W., and Qin, D. (2008). Evaluation of ERA-40, NCEP-1, and NCEP-2 reanalysis air temperatures with ground-based measurements in China. J. Geophys. Res. Atmospheres *113*, D15115.

Ma, L., Zhang, T., Frauenfeld, O.W., Ye, B., Yang, D., and Qin, D. (2009). Evaluation of precipitation from the ERA-40, NCEP-1, and NCEP-2 Reanalyses and CMAP-1, CMAP-2, and GPCP-2 with ground-based measurements in China. J. Geophys. Res. Atmospheres *114*, D09105.

Maraun, D. (2012). Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. Geophys. Res. Lett. *39*, L06706.

Maraun, D. (2013). Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue. J. Clim. *26*, 2137–2143.

Maurer, E.P., and Hidalgo, H.G. (2008). Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods. Hydrol Earth Syst Sci *12*, 551–563.

Maurer, E.P., Wood, A.W., Adam, J.C., Lettenmaier, D.P., and Nijssen, B. (2002). A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States. J Clim. *15*, 3237–3251.

Maurer, E.P., Hidalgo, H.G., Das, T., Dettinger, M.D., and Cayan, D.R. (2010). The utility of daily large-scale climate data in the assessment of climate change impacts on daily streamflow in California. Hydrol Earth Syst Sci *14*, 1125–1138.

Maurer, E.P., Das, T., and Cayan, D.R. (2013). Errors in climate model daily precipitation and temperature output: time invariance and implications for bias correction. Hydrol. Earth Syst. Sci. *17*, 2147–2159.

McKenney, D.W., Hutchinson, M.F., Papadopol, P., Lawrence, K., Pedlar, J., Campbell, K., Milewska, E., Hopkinson, R.F., Price, D., and Owen, T. (2011). Customized Spatial Climate Models for North America. Bull. Am. Meteorol. Soc. *92*, 1611–1622.

Monk, W.A., Peters, D.L., Allen Curry, R., and Baird, D.J. (2011). Quantifying trends in indicator hydroecological variables for regime-based groups of Canadian rivers. Hydrol. Process. *25*, 3086–3100.

Murdock, T.Q., Cannon, A.J., and Sobie, S.R. (2013). New statistical downscaling for Canada.

Murdock, T.Q., Cannon, A.J., and Sobie, S.R. (2014). Statistical downscaling of future climate projections for North America. (Victoria, BC, Canada: Report on Contract No: KM040-131148/A. Prepared for Environment Canada, Pacific Climate Impacts Consortium).

Najafi, M.R., Moradkhani, H., and Jung, I.W. (2011). Assessing the uncertainties of hydrologic model selection in climate change impact studies. Hydrol. Process. *25*, 2814–2826.

Nash, J.E., and Sutcliffe, J.V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. *10*, 282–290.

Nijssen, B., Schnur, R., and Lettenmaier, D.P. (2001). Global retrospective estimation of soil moisture using the variable infiltration capacity land surface model, 1980-93. J. Clim. *14*, 1790–1808.

Ouarda, T.B.M.J., Cunderlik, J.M., St-Hilaire, A., Barbet, M., Bruneau, P., and Bobée, B. (2006). Data-based comparison of seasonality-based regional flood frequency methods. J. Hydrol. *330*, 329–339.

Peterson, T. C. (2005). Climate Change Indices. WMO Bull. *54*, 83–86.

Pierce, D.W., Cayan, D.R., Das, T., Maurer, E.P., Miller, N.L., Bao, Y., Kanamitsu, M., Yoshimura, K., Snyder, M.A., Sloan, L.C., et al. (2013). The Key Role of Heavy Precipitation Events in Climate Model Disagreements of Future Annual Precipitation Changes in California. J. Clim. *26*, 5879–5896.

Prudhomme, C., and Davies, H. (2008). Assessing uncertainties in climate change impact analyses on the river flow regimes in the UK. Part 2: future climate. Clim. Change *93*, 197–222.

Richter, B.D., Baumgartner, J.V., Powell, J., and Braun, D.P. (1996). A Method for Assessing Hydrologic Alteration within Ecosystems. Conserv. Biol. *10*, 1163–1174.

Rodenhuis, D., Bennett, K., Werner, A., Murdock, T.Q., and Bronaugh, D. (2009). Hydro-climatology and Future Climate Impacts in British Columbia, revised 2009 (Victoria, BC: Pacific Climate Impacts Consortium, University of Victoria).

Salathé, E.P. (2005). Downscaling simulations of future global climate with application to hydrologic modelling. Int. J. Climatol. *25*, 419–436.

Salathe, E.P., Mote, P.W., and Wiley, M.W. (2007). Review of scenario selection and downscaling methods for the assessment of climate change impacts on hydrology in the United States pacific northwest. Int. J. Climatol. *27*, 1611–1621.

Schnorbus, M.A., and Cannon, A.J. (2014). Statistical emulation of streamflow projections from a distributed hydrological model: Application to CMIP3 and CMIP5 climate projections for British Columbia, Canada. Water Resour. Res. *50*, 8907–8926.

Schnorbus, M., Werner, A., and Bennett, K. (2014). Impacts of climate change in three hydrologic regimes in British Columbia, Canada. Hydrol. Process. *28*, 1170–1189.

Sheffield, J., Wood, E.F., and Roderick, M.L. (2012). Little change in global drought over the past 60 years. Nature *491*, 435–438.

Sherwood, S., and Fu, Q. (2014). A Drier Future? Science *343*, 737–739.

Shrestha, R.R., Schnorbus, M.A., Werner, A.T., and Berland, A.J. (2012). Modelling spatial and temporal variability of hydrologic impacts of climate change in the Fraser River basin, British Columbia, Canada. Hydrol. Process. *26*, 1840–1860.

Shrestha, R.R., Schnorbus, M.A., Werner, A.T., and Zwiers, F.W. (2014a). Evaluating Hydroclimatic Change Signals from Statistically and Dynamically Downscaled GCMs and Hydrologic Models. J. Hydrometeorol. *15*, 844–860.

Shrestha, R.R., Peters, D.L., and Schnorbus, M.A. (2014b). Evaluating the ability of a hydrologic model to replicate hydro-ecologically relevant indicators. Hydrol. Process. *28*, 4294–4310.

Sillmann, J., Kharin, V.V., Zhang, X., Zwiers, F.W., and Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. J. Geophys. Res. Atmospheres *118*, 1716–1733.

Stahl, K., Hisdal, H., Hannaford, J., Tallaksen, L., Van Lanen, H., Sauquet, E., Demuth, S., Fendekova, M., and Jordar, J. (2010). Streamflow trends in Europe: evidence from a dataset of near-natural catchments. Hydrol. Earth Syst. Sci. *14*, p. 2367 – p. 2382.

Stahl, K., Tallaksen, L.M., Hannaford, J., and van Lanen, H.A.J. (2012). Filling the white space on maps of European runoff trends: estimates from a multi-model ensemble. Hydrol Earth Syst Sci *16*, 2035–2047.

Storch, H.V. (1982). A Remark on Chervin-Schneider's Algorithm to Test Significance of Climate Experiments with GCM's. J. Atmospheric Sci. *39*, 187–189.

Themeßl, M.J., Gobiet, A., and Heinrich, G. (2011). Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal. Clim. Change *112*, 449–468.

Thrasher, B., Maurer, E.P., McKellar, C., and Duffy, P.B. (2012). Technical Note: Bias correcting climate model simulated daily temperature extremes with quantile mapping. Hydrol Earth Syst Sci *16*, 3309–3314.

Uppala, S.M., KÅllberg, P.W., Simmons, A.J., Andrae, U., Bechtold, V.D.C., Fiorino, M., Gibson, J.K., Haseler, J., Hernandez, A., Kelly, G.A., et al. (2005). The ERA-40 re-analysis. Q. J. R. Meteorol. Soc. *131*, 2961–3012.

Wang, T., Hamann, A., Spittlehouse, D.L., and Aitken, S.N. (2006). Development of scale-free climate data for Western Canada for use in resource management. Int. J. Climatol. *26*, 383–397.

Werner, A.T. (2011). BCSD Downscaled Transient Climate Projections for Eight Select GCMs over British Columbia, Canada (Victoria, BC: Pacific Climate Impacts Consortium, University of Victoria).

Werner, A.T., Schnorbus, M.A., Shrestha, R.R., and Eckstrand, H.D. (2013). Spatial and Temporal Change in the Hydro-Climatology of the Canadian Portion of the Columbia River Basin under Multiple Emissions Scenarios. Atmosphere-Ocean *51*, 357–379.

Werner, A.T., Nienaber, P., Schnorbus, M.A., and Bronaugh, D. (2015). A Cross Validation of the VIC Forcings Gridded-Observations for British Columbia (Victoria, BC, Canada: Pacific Climate Impacts Consortium, University of Victoria).

Wilks, D.S. (2006). On "Field Significance" and the False Discovery Rate. J. Appl. Meteorol. Climatol. *45*, 1181–1189.

Wood, A.W., Maurer, E.P., Kumar, A., and Lettenmaier, D.P. (2002). Long-range experimental hydrologic forecasting for the eastern United States. J Geophys Res *107*, 4429.

Wood, A.W., Leung, L.R., Sridhar, V., and Lettenmaier, D.P. (2004). Hydrologic Implications of Dynamical and Statistical Approaches to Downscaling Climate Model Outputs. Clim. Change *62*, 189–216.

Zhang, X., Alexander, L., Hegerl, G.C., Jones, P., Tank, A.K., Peterson, T.C., Trewin, B., and Zwiers, F.W. (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. Wiley Interdiscip. Rev. Clim. Change *2*, 851–870.

Formatted: Bibliography

Table 1. Availability of gridded observations and reanalyses.

Reanalysis Product	***Start***	***End***	***Resolution***	***Reference***
NCEP1	1948	present	~1.9°	Kalnay et al. 1996
20CR	1871	2011	2°	Compo et al. 2011
ERA40	1958	2001	2.5°	Uppala et al. 2005
ERAInt	1979	present	1.5°	Dee et al. 2011
Gridded Observation	***Start***	***End***	***Resolution***	***Reference***
VIC Forcings	1950	2005	~ 6 km	Schnorbus et al. 2014
ANUSPLIN	1950	2005	~ 10 km	Hutchinson et al. 2009

Table 2. Calibration and validation periods for downscaling methods by reanalyses.

Reanalysis Product	*Calibration*	*No. Years*	*Validation*	*No. Years*
NCEP1	1950-1990	41	1991-2005	15
20CR	1950-1990	41	1991-2005	15
ERA40	1958-1990	33	1991-2001	11
ERAInt	1979-1990	12	1991-2005	15

Table 3. Metadata for five select sub-basins of the Peace River basin.

Basin	*Water Survey of Canada ID*	*Drainage Area (km^2)*	*Elevation (m)*		
			mean	*min*	*max*
BCGMS	---	72,078			
FINAK	07EA005	16,000	1452	693	2799
INGEN	07EA004	4,200	1503	674	2289
PARMS	07EE007	4,900	1128	645	2343
PEAPN	07FA004	83,900	1126	392	2799

Table 4. Calibration and validation statistics for five select sub-basins of the Peace River basin under the under VIC Forcings and ANUSPLIN gridded observational datasets including the Nash-Sutcliff Efficiency score (*NS*), the Nash-Sutcliff Efficiency score of the log-transformed discharge (*LNS*) and the percent volume bias error (*%VB*).

	VIC Forcings						**ANUSPLIN**					
Basin	*Calibration 1990-1995*			*Validation 1985-1989*			*Calibration 1990-1995*			*Validation 1985-1989*		
	NS	*LNS*	*%VB*	*NS*	*LNS*	*%VB*	*NS*	*LNS*	*%VB*	*NS*	*LNS*	*%VB*
BCGMS	0.64	0.81	-1	0.75	0.83	-12	0.72	0.82	3	0.82	0.84	3
FINAK	0.66	0.85	0	0.83	0.88	-14	0.76	0.81	11	0.73	0.81	30
INGEN	0.76	0.82	0	0.82	0.78	-15	0.69	0.83	10	0.72	0.85	26
PARMS	0.78	0.71	0	0.81	0.66	-9	0.78	0.62	10	0.75	0.63	8
PEAPN	0.65	0.79	-2	0.76	0.87	-10	0.71	0.80	2	0.82	0.85	2

Table 5. Mean annual ClimDEX values for VIC Forcings and ANUSPLIN averaged over the Peace River basin.

	Calibration (1950-1990)		Validation (1991-2005)			
Index	**VIC Forcings**	**ANUSPLIN**	**VIC Forcings**	**ANUSPLIN**	***Units***	**Indicator Name**
cdd	20	19	18	19	*Days*	Consecutive dry days
csdi	5	9	5	6	*Days*	Cold spell duration
cwd	9	10	11	12	*Days*	Consecutive wet days
dtr	11	11	10.6	10.3	*°C*	Diurnal T range
fd	239	238	233	230	*Days*	Frost days
gsl	136	131	140	138	*Days*	Growing season
id	109	122	102	106	*Days*	Ice days
prcptot	703	578	742	585	*mm*	Annual total wet-day
r1mm	133	142	150	153	*Days*	Precipitation days
r10mm	17	8	17	8	*Days*	Heavy prec. days
r20mm	4	1	4	1	*Days*	Very heavy prec.
r95p	145	97	142	100	*mm*	Very wet days
r99p	42	28	38	32	*mm*	Extremely wet days
rx1day	32	22	31	23	*mm*	Max 1-day prec.
rx5day	63	46	64	46	*mm*	Max 5-day prec.
sdii	5	4	5	4	$mm\ day^{-1}$	Simple daily intense
su	7	6	7	7	*Days*	Summer days
tn10p	11	13	7	8	*%*	Cool nights
tn90p	10	9	12	14	*%*	Warm nights
tnn	-37	-41	-35.5	-37.6	*°C*	Min Monthly Tn
tnx	11	11	11.5	11.8	*°C*	Max Monthly Tn
tx10p	11	11	9	8	*%*	Cool days
tx90p	10	10	11	14	*%*	Warm days
txn	-27	-29	-24.9	-25.8	*°C*	Min Monthly Tx
txx	27	27	27.9	27.4	*°C*	Max Monthly Tx
wsdi	4	5	8	12	*Days*	Warm spell duration

Table 6. Summary of number of tests passed for Pearson's correlations and similarity in distributions (KS test) based on the Walker field significance test between ClimDEX indices for downscaled reanalyses versus target gridded observation over the Peace River basin for 1991-2005 (1991-2001 ERA40), summarized by gridded observation, reanalysis and downscaling method. Max indicates maximum possible tests to pass in that category.

Gridded Observation	**Pearson's correlation**	**KS test**	**Combined**
VIC Force	367	578	945
ANUSPLIN	388	628	1016
Max	**728**	**728**	**1456**
Reanalyses	**Pearson's correlation**	**KS test**	**Combined**
NCEP1	159	284	443
20CR	147	287	434
ERA40	201	340	541
ERAInt	248	295	543
Max	**364**	**364**	**728**
Downscaling Method	**Pearson's correlation**	**KS test**	**Combined**
BCCA	130	171	301
DBCCA	139	174	313
BCCI	131	176	307
CI	139	154	293
BCSD	56	175	231
BCSDX	48	173	221
BCCAQ	112	183	295
Max	**208**	**208**	**416**

Table 7. Number of tests passed for each ClimDEX indices for VIC Forcings and ANUSPLIN for 1991-2005 (1991-2001 ERA40).

	VIC Forcings	ANUSPLIN	Difference
cdd	48	44	4
csdi	54	54	0
cwd	19	31	-12
dtr	32	3	29
fd	51	48	3
gsl	54	52	2
id	55	47	8
prcptot	24	33	-9
r10mm	28	31	-3
r1mm	24	36	-12
r20mm	26	42	-16
r95p	11	28	-17
r99p	24	41	-17
rx1day	14	35	-21
rx5day	30	33	-3
sdii	2	15	-13
su	51	50	1
tn10p	52	52	0
tn90p	48	43	5
tnn	42	39	3
tnx	30	32	-2
tx10p	52	52	0
tx90p	50	50	0
txn	43	44	-1
txx	41	42	-1
wsdi	40	39	1

Table 8. Summary of number of tests passed for Pearson's correlations and similarity in distributions (KS test) based on the Walker field significance test between ClimDEX indices for downscaled reanalyses versus target gridded observation over the Peace River basin for 1991-2005 (1991-2001) for reanalysis (ERA40) versus downscaling method for each gridded observation.

		Pearson's correlation					**KS test**					
		NCEP1	**20CR**	**ERA40**	**ERAInt**	***Sub***	**NCEP1**	**20CR**	**ERA40**	**ERAInt**	***Sub***	**Total**
VIC Forcings	**BCCA**	14	14	14	17	*59*	19	21	24	18	*82*	141
	DBCCA	15	14	15	18	*62*	20	22	24	18	*84*	146
	BCCI	14	14	16	20	*64*	20	21	24	22	*87*	151
	CI	13	14	17	22	*66*	16	14	24	18	*72*	138
	BCSD	4	6	6	12	*28*	20	20	24	20	*84*	112
	BCSDX	4	5	7	11	*27*	20	20	24	20	*84*	111
	BCCAQ	15	13	14	19	*61*	20	21	24	20	*85*	146
	Subtotal	79	*80*	*89*	*119*		*135*	*139*	*168*	*136*		
ANUSPLIN	**BCCA**	17	11	23	20	*71*	22	23	24	20	*89*	160
	DBCCA	17	13	23	24	*77*	21	20	24	25	*90*	167
	BCCI	14	12	18	23	*67*	21	21	24	23	*89*	156
	CI	15	14	20	24	*73*	15	19	24	24	*82*	155
	BCSD	5	4	8	11	*28*	24	21	25	21	*91*	119
	BCSDX	3	3	5	10	*21*	24	20	25	20	*89*	110
	BCCAQ	9	10	15	17	*51*	22	24	26	26	*98*	149
	Subtotal	80	*67*	*112*	*129*		*149*	*148*	*172*	*159*		

Table 9. Summary of number of tests passed for Pearson's correlations and similarity in distributions (KS test) based on the Walker field significance test between hydrologic extremes for downscaled reanalyses versus target gridded observation over the Peace basin for 1991-2005 (1991-2001 ERA40), summarized by gridded observation, reanalysis and downscaling method. Max indicates maximum possible tests to pass in that category.

Gridded Observation	**Pearson's correlation**	**KS test**	**Combined**
VIC Force	309	404	713
ANUSPLIN	310	350	660
Max	**420**	**420**	**840**
Reanalyses	**Pearson's correlation**	**KS test**	**Combined**
NCEP1	135	188	323
20CR	125	181	306
ERA40	180	196	376
ERAInt	179	189	368
Max	**210**	**210**	**420**
Downscaling Method	**Pearson's correlation**	**KS test**	**Combined**
BCCA	102	96	198
DBCCA	104	111	215
BCCI	107	111	218
CI	99	87	186
BCSD	49	119	168
BCSDX	48	119	167
BCCAQ	110	111	221
Max	**120**	**120**	**240**

Table 10. Number of basins where downscaled results were significantly correlated and distributions were not significantly different than those from the VIC Forcings gridded observations derived 3-day peak flow by downscaling method / reanalysis combinations for 1991-2005 (1991-2001 ERA40).

		Pearson's correlation					**KS test**					
		NCEP1	**20CR**	**ERA40**	**ERAInt**	***Sub***	**NCEP1**	**20CR**	**ERA40**	**ERAInt**	***Sub***	**Total**
VIC Forcings	**BCCA**	2	2	5	5	14	5	5	5	1	16	30
	DBCCA	1	3	5	5	14	5	5	5	5	20	34
	BCCI	2	5	5	5	17	5	5	5	5	20	37
	CI	5	2	5	5	17	5	5	5	5	20	37
	BCSD	3	2	4	2	11	5	5	5	5	20	31
	BCSDX	3	3	4	2	12	5	5	5	5	20	32
	BCCAQ	3	5	5	5	18	5	5	5	5	20	38
	Subtotal	19	22	33	29	**103**	35	35	35	31	**136**	
ANUSPLIN	**BCCA**	5	0	5	5	15	4	4	4	1	13	28
	DBCCA	5	1	5	5	16	5	2	4	5	16	32
	BCCI	5	2	4	5	16	5	2	4	5	16	32
	CI	4	0	5	5	14	5	2	4	5	16	30
	BCSD	2	0	3	3	8	5	5	5	5	20	28
	BCSDX	2	0	3	3	8	5	5	4	5	19	27
	BCCAQ	5	3	5	5	18	5	2	5	5	17	35
	Subtotal	28	6	30	31	**95**	34	22	30	31	**117**	

Table 11. Number of basins where downscaled results were significantly correlated and distributions were not significantly different than those from the VIC Forcings gridded observations derived summer 7-day low flow for downscaling method / reanalysis combinations for 1991-2005 (1991-2001 ERA40).

		Pearson's correlation					**KS test**					
		NCEP1	**20CR**	**ERA40**	**ERAInt**	***Sub***	**NCEP1**	**20CR**	**ERA40**	**ERAInt**	***Sub***	**Total**
VIC Forcings	**BCCA**	3	2	5	5	15	5	5	5	5	20	35
	DBCCA	3	2	5	5	15	5	5	5	5	20	35
	BCCI	3	4	5	5	17	5	5	5	5	20	37
	CI	2	4	5	5	16	5	5	5	5	20	36
	BCSD	2	3	3	4	12	5	5	5	5	20	32
	BCSDX	2	2	3	4	11	5	5	5	5	20	31
	BCCAQ	4	3	5	5	17	5	5	5	5	20	37
	Subtotal	19	20	31	33	**103**	35	35	35	35	**140**	
ANUSPLIN	**BCCA**	5	4	5	5	19	5	5	5	1	16	35
	DBCCA	5	5	5	5	20	5	5	5	5	20	40
	BCCI	3	5	5	5	18	5	5	5	5	20	38
	CI	1	5	5	5	16	5	5	5	5	20	36
	BCSD	1	2	4	5	12	5	5	5	5	20	32
	BCSDX	1	2	4	5	12	5	5	5	5	20	32
	BCCAQ	3	5	5	5	18	5	5	5	5	20	38
	Subtotal	19	28	33	35	**115**	35	35	35	31	**136**	

Table 12. Number of basins where downscaled results were significantly correlated and distributions were not significantly different than those from the VIC Forcings gridded observations derived winter 7-day low flow for downscaling method / reanalysis combinations for 1991-2005 (1991-2001 ERA40).

		Pearson's correlation					**KS Test**					
		NCEP1	**20CR**	**ERA40**	**ERAInt**	***Sub***	**NCEP1**	**20CR**	**ERA40**	**ERAInt**	***Sub***	**Total**
VIC Forcings	**BCCA**	5	5	5	5	20	5	5	5	5	20	40
	DBCCA	5	5	5	5	20	5	5	5	5	20	40
	BCCI	5	5	5	5	20	5	5	5	5	20	40
	CI	5	5	4	5	19	4	2	2	0	8	27
	BCSD	0	0	2	0	2	5	5	5	5	20	22
	BCSDX	0	0	2	0	2	5	5	5	5	20	22
	BCCAQ	5	5	5	5	20	5	5	5	5	20	40
	Subtotal	25	25	28	25	**103**	34	32	32	30	**128**	
ANUSPLIN	**BCCA**	5	5	4	5	19	1	3	4	3	11	30
	DBCCA	5	5	4	5	19	2	3	5	5	15	34
	BCCI	5	4	5	5	19	2	3	5	5	15	34
	CI	5	5	3	4	17	0	0	0	3	3	20
	BCSD	0	0	2	2	4	4	5	5	5	19	23
	BCSDX	0	1	2	0	3	5	5	5	5	20	23
	BCCAQ	5	4	5	5	19	1	3	5	5	14	33
	Subtotal	25	24	25	26	**100**	15	22	29	31	**97**	

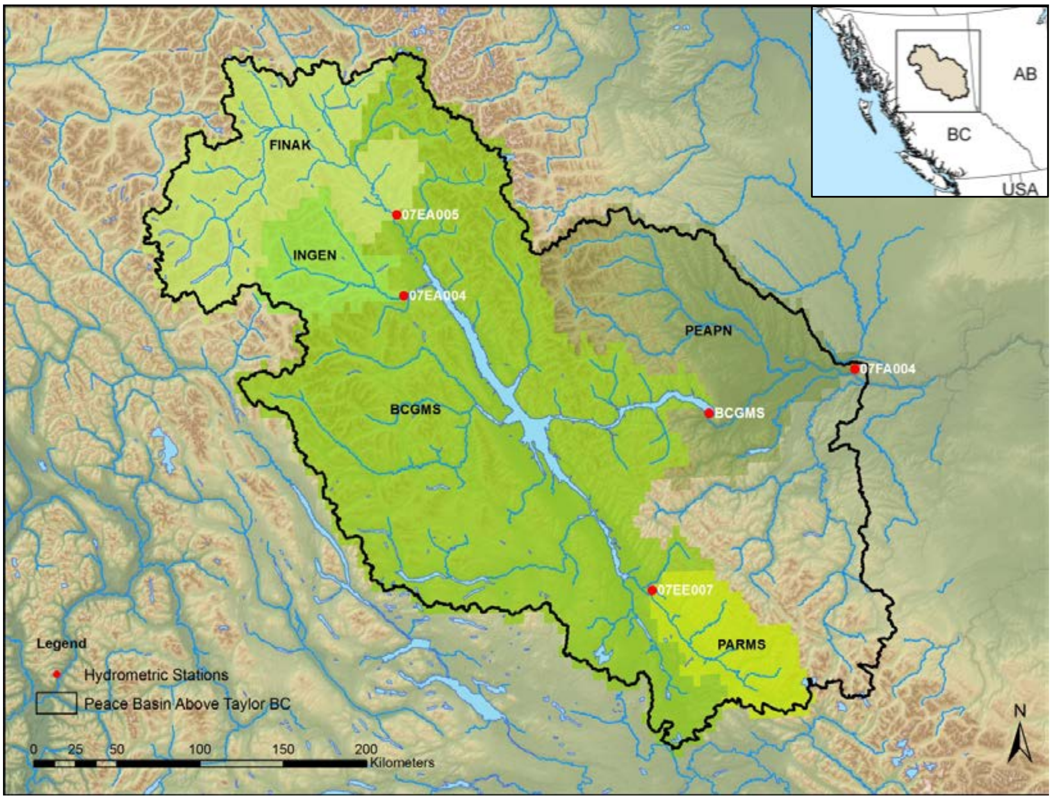


Figure 1. The Peace River basin (above Taylor BC) study area analyzed for ClimDEX indices (black boundary) and the five sub-basins investigated for hydrologic extremes including the Finlay River above Akie River (FINAK), Ingenika River above Swannell River (INGEN), Parsnip River above Misinchinka River (PARMS), the Peace River above Pine River (PEAPN), and the Peace River at Bennett Dam (BCGMS).

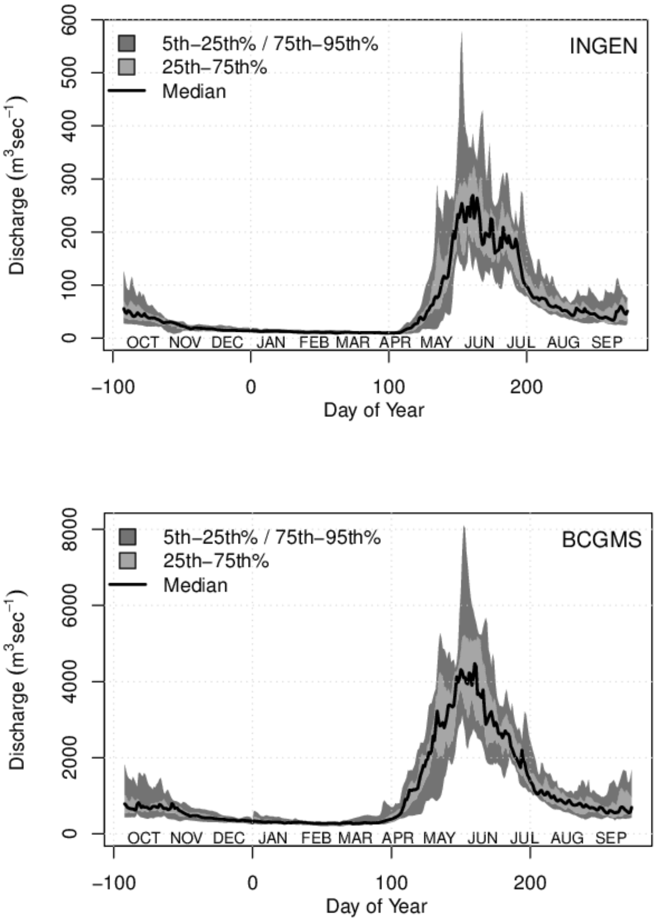
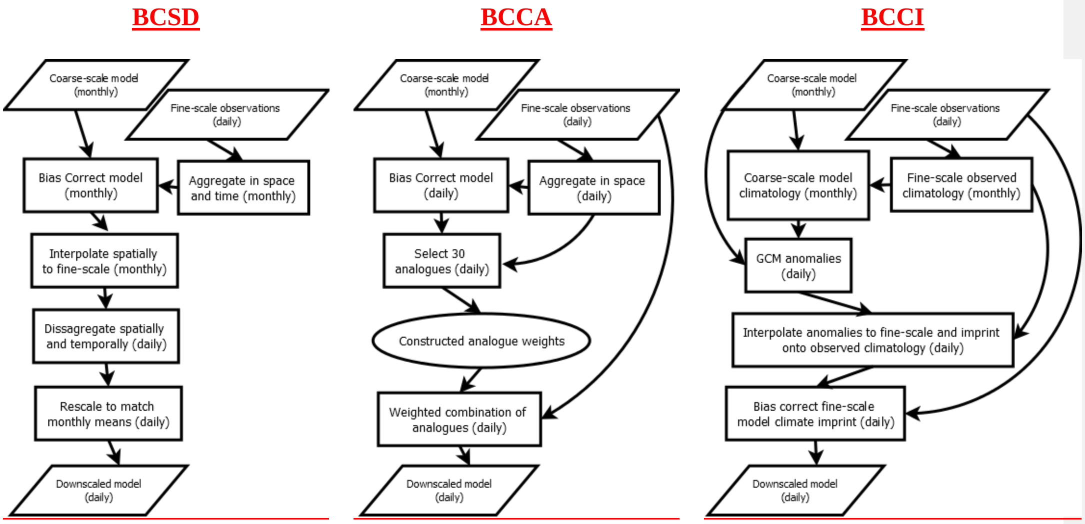


Figure 2. Annual daily hydrograph 1985 to 1995 for (top) Ingenika and (bottom) BCGMS hydrometric sites.



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 (to avoid trends and drizzle introduced by BCCA).

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).

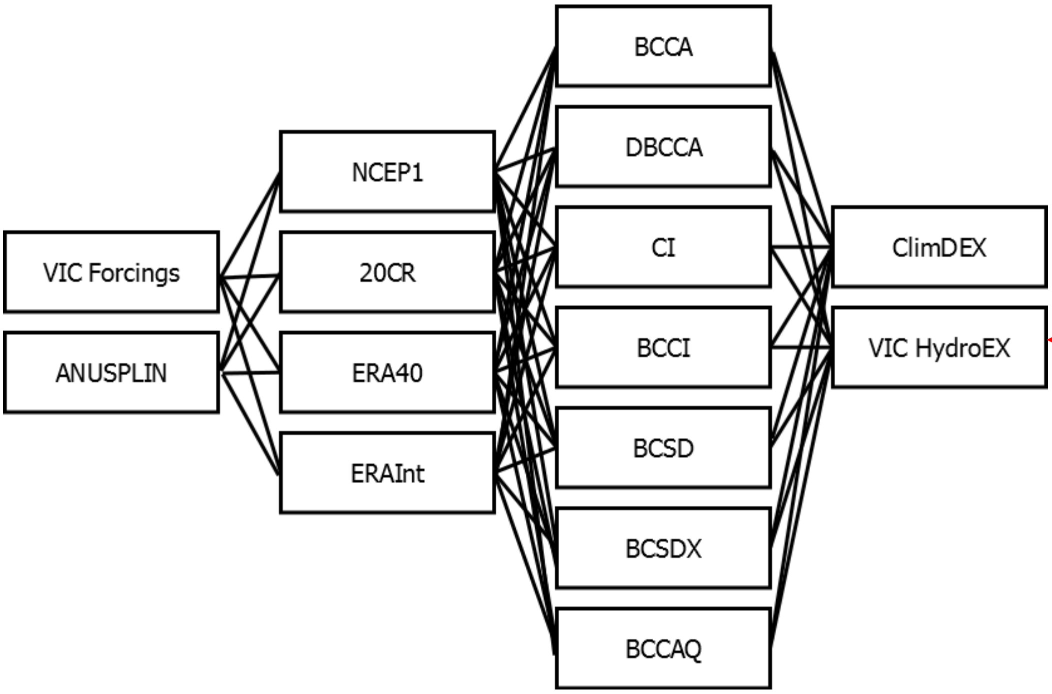


Figure 3~~.~~b. Workflow diagram for assessment of downscaling techniques in replicating ClimDEX and hydrologic extremes.

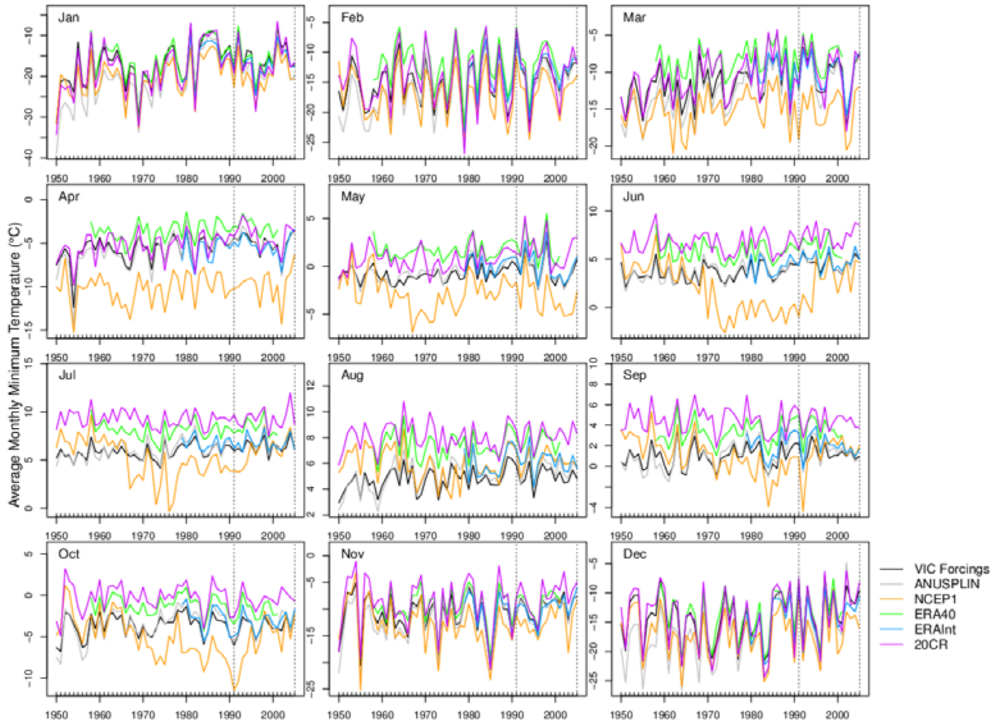


Figure 4. Monthly average minimum temperature by gridded observations (VIC Forcings and ANUSPLIN) and reanalysis (NCEP1, ERA40, ERAInt, 20CR) over the Peace River basin.

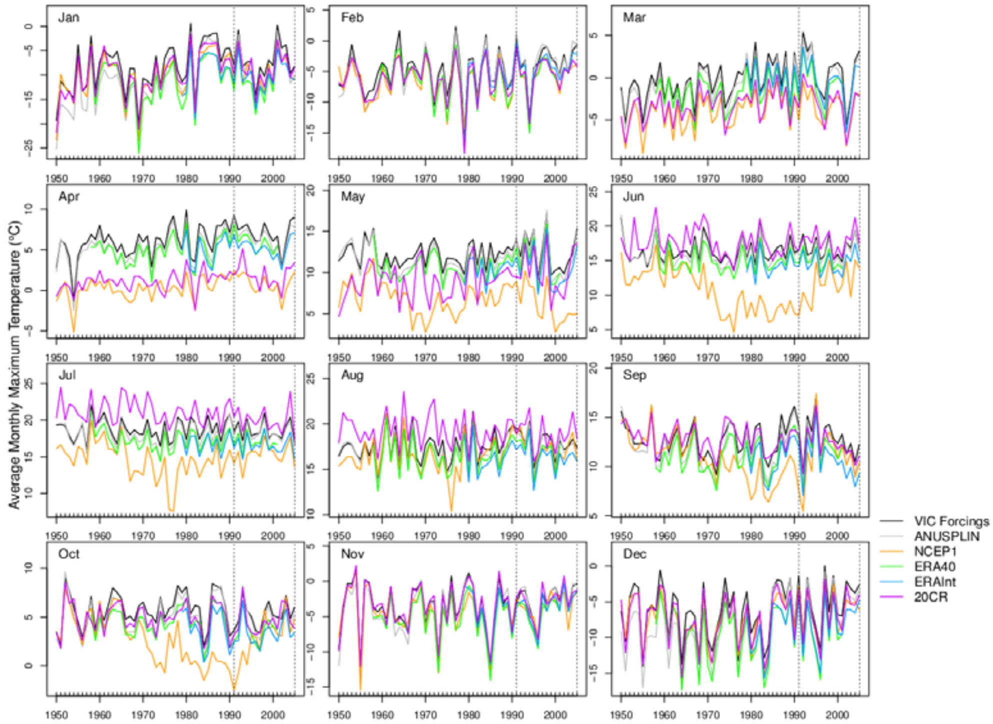


Figure 5. Monthly average maximum temperature by gridded observations (VIC Forcings and ANUSPLIN) and reanalysis (NCEP1, ERA40, ERAInt, 20CR) over the Peace River basin.

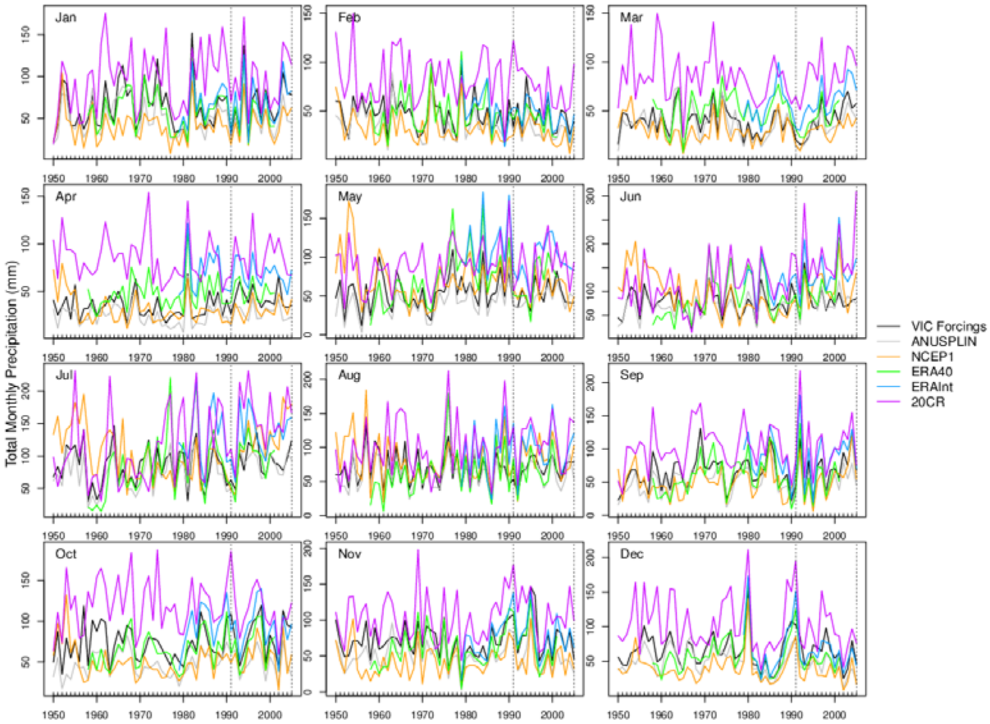


Figure 6. Monthly total precipitation by gridded observations (VIC Forcings and ANUSPLIN) and reanalysis (NCEP1, ERA40, ERAInt, 20CR) over the Peace River basin.

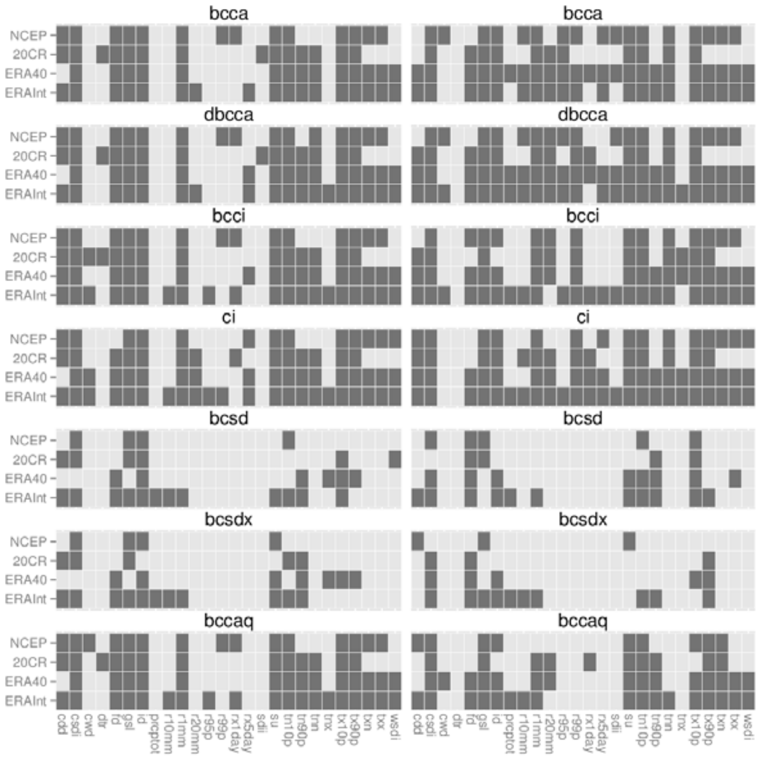


Figure 7. Field significant correlations based on the Walker field significance test over the Peace River basin between ClimDEX indices for downscaled reanalysis versus target gridded observation, VIC Forcings (left) and ANUSPLIN (right), by downscaling method for 1991-2005 (1991-2001 ERA40). **Dark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

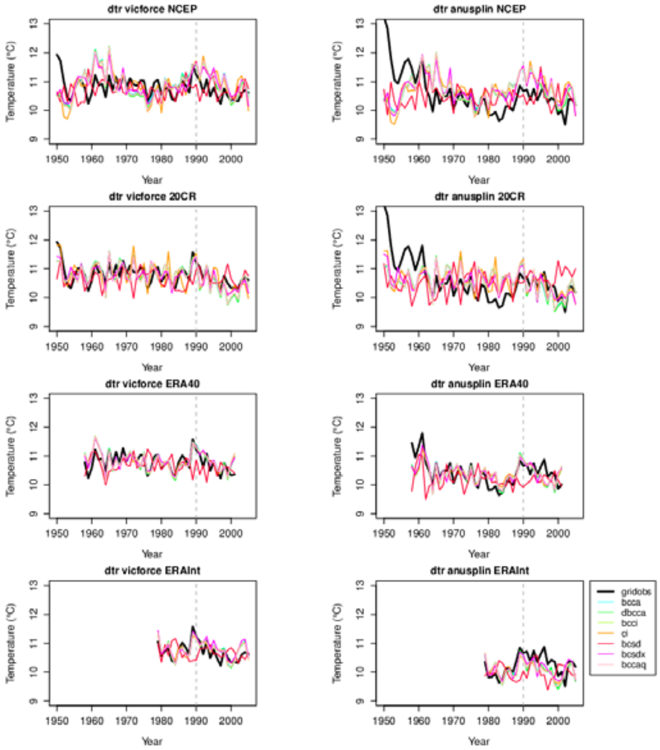


Figure 8. Time series of average DTR from VIC Forcings (left) and ANUSPLIN (right) for NCEP1 (top), 20CR (second), ERA40 (third) and ERAInt (bottom) downscaled using BCCA, DBCCA, BCCI, CI, BCSD, BCSDX and BCCAQ over the Peace River basin.

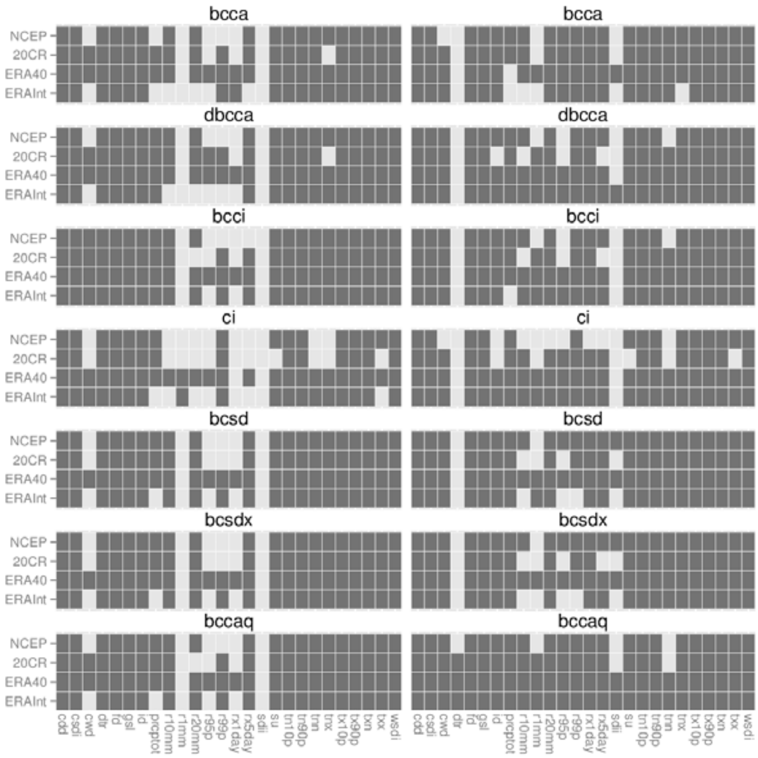


Figure 9. Field significant ~~differences~~similarities in distributions based on the Walker field significance test over the Peace River basin between ClimDEX indices for downscaled reanalysis versus target gridded observation, VIC Forcings (left) and ANUSPLIN (right), by downscaling method for 1991-2005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

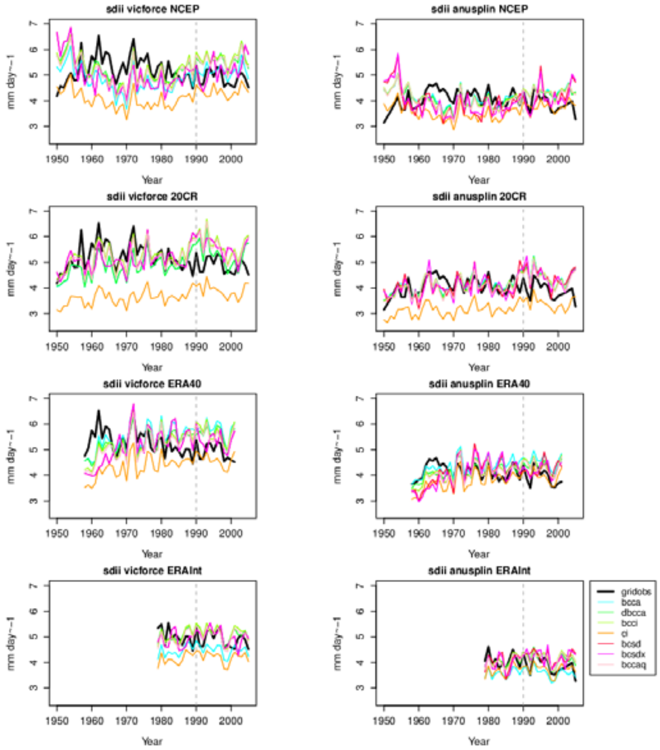
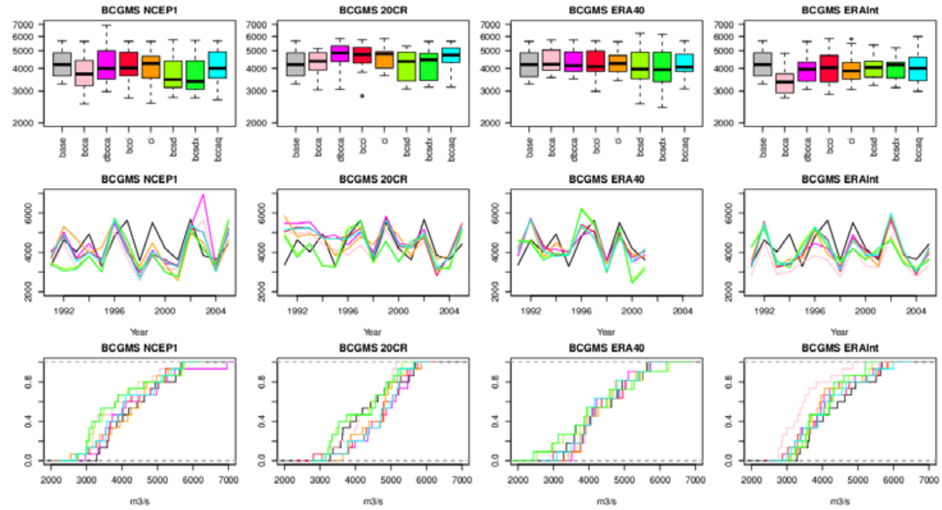


Figure 10. Time series of average SDII from VIC Forcings (left) and ANUSPLIN (right) for NCEP1 (top), 20CR (second), ERA40 (third) and ERAInt (bottom) downscaled using BCCA, DBCCA, BCCI, CI, BCSD, BCSDX and BCCAQ over the Peace River basin.



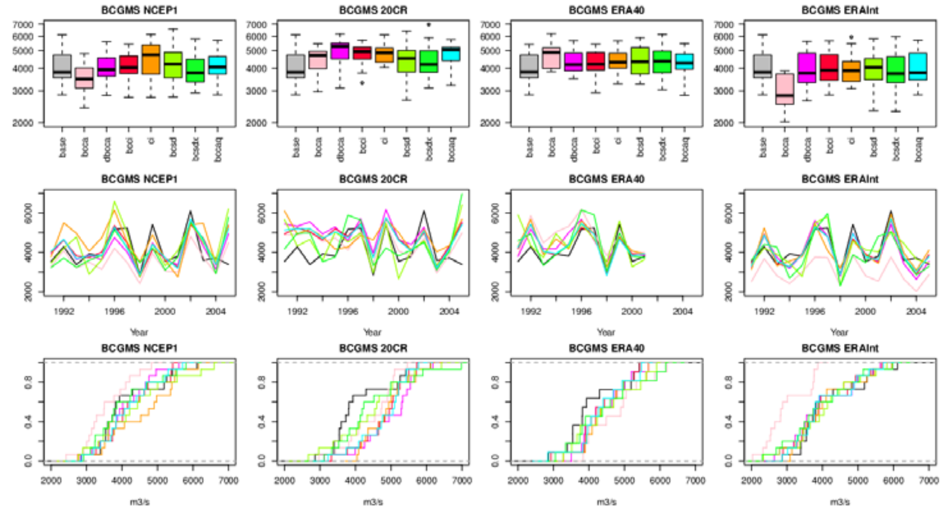


Figure 11. Boxplots, time series and distributions of 3-day peak flow in the spring months (May-July) for NCEP1, 20CR, ERA40 and ERAInt in the BCGMS basin based on VIC Forcings (top) ANUSPLIN (bottom). Legend same as Figure 9.

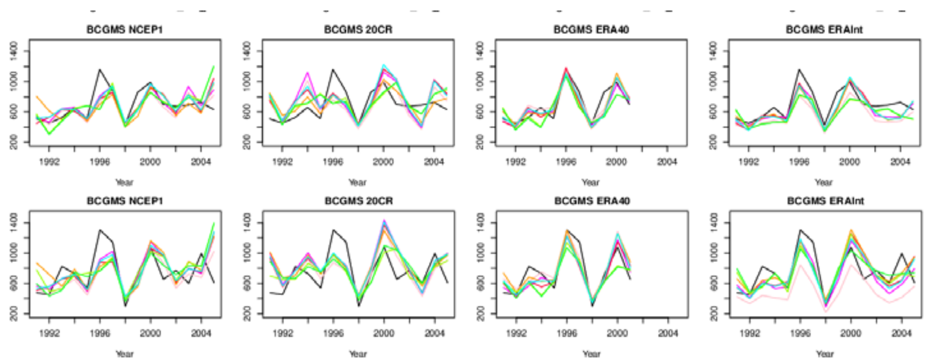


Figure 12. Time series of 7-day low flow in the summer months (July-September) for NCEP1, 20CR, ERA40 and ERAInt in the BCGMS basin based on VIC Forcings (top) ANUSPLIN (bottom). Legend same as Figure 9.

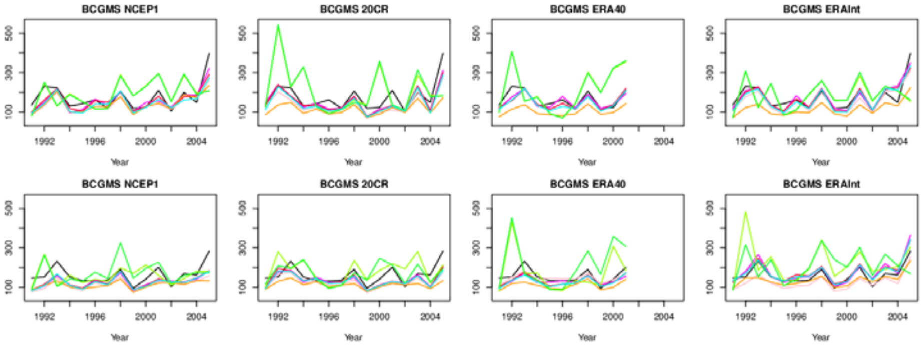


Figure 13. Time series of 7-day low flow in the winter months (November to April) for NCEP1, 20CR, ERA40 and ERAInt in the BCGMS basin based on VIC Forcings (top) ANUSPLIN (bottom). Legend same as Figure 9.

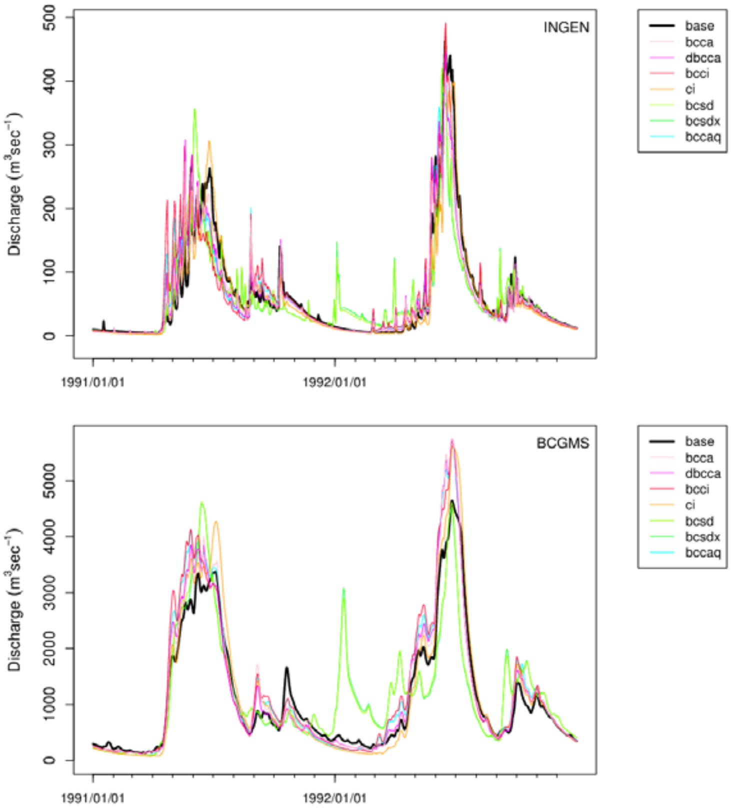


Figure 14. Time series of daily streamflow in the BCGMS basin as driven by ANUSPLIN (base) and ERA40 downscaled to ANUSPLIN with the BCCA, DBCCA, BCCI, CI, BCSD, BCSDX and BCCAQ methods over 1991 to 2005.