The DOI for this manuscript is doi: 10.1175/JCLI-D-12-00249.1

Bürger, G., S. Sobie, A. Cannon, A. Werner, and T. Murdock, 2012: Downscaling extremes - an intercomparison of multiple methods for future climate. J. Climate. doi:10.1175/JCLI-D-12-00249.1, in press.

Downscaling extremes - an intercomparison of multiple methods for future climate

G. Bürger (*gbuerger@uni-potsdam.de*), Universität Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Deutschland; Pacific Climate Impacts Consortium (PCIC), University House 1, PO Box 3060 Stn CSC, University of Victoria, Victoria, British Columbia, Canada V8W 3R4

T. Q. Murdock, PCIC

A. T. Werner, PCIC

S. R. Sobie, PCIC

A. J. Cannon, PCIC

Abstract

This study follows up on a previous downscaling intercomparison for present climate. Using a larger set of 8 methods we downscale atmospheric fields representing present (1981-2000) and future (2046-2065) conditions, as simulated by 6 global climate models following 3 emission scenarios. At 20 locations in British Columbia we study local extremes as measured by the same set of 27 indices, ClimDEX, as in the precursor study. Present and future simulations give 2×3×6×8×20×27=155520 index climatologies whose analysis in terms of mean change and variation is the purpose of this study. The mean change generally reinforces what is to be expected in a warmer climate: that extreme cold events become less frequent and extreme warm events become more frequent, and that there are signs of more frequent precipitation extremes. There is considerable variation, however, about this tendency, caused by the influence of scenario, climate model, downscaling method and location. This is analyzed using standard statistical techniques such as ANOVA and multidimensional scaling, along with an assessment of the influence of each modeling component on the overall variation of the simulated change. We find that downscaling generally has the strongest influence, followed by climate model; location and scenario have only a minor influence. The influence of downscaling could be traced back in part to various issues related to the methods, such as the quality of simulated variability or the dependence on predictors. Using only methods validated in the precursor study considerably reduced the influence of downscaling, underpinning the general need for method verification.

§1 Introduction

This study presents a follow-up to a downscaling intercomparison study conducted for present climate [*Bürger et al.*, 2012, henceforth dip1]. While dip1 exclusively dealt with the statistics of present climate extremes and the verification of a number of downscaling methods, here we study and compare the same methods, plus several others, with respect to their simulation from future emission scenarios. For a similar set of regions in British Columbia, Canada, see Figure 1, essentially the same model chain is employed, with several different global climate models being driven by a set of emission scenarios that are subsequently downscaled by multiple methods to various stations to simulate daily weather. Likewise, as in dip1 we use the set of 27 core indices that by now forms the international standard to monitor climatic extremes, and that is recommended by the Expert Team on Climate Change Detection and Indices (cf. *http://cccma.seos.uvic.ca/ETCCDI*); these 'Climate inDices of EXtremes' (ClimDEX) are estimated from long-term statistics of daily temperature and precipitation series.

To recapitulate the main setup and findings of dip1, we had used a threefold testing procedure for each of five downscaling methods (ASD, BCSD, QRNN, TG, XDS, acronyms explained below) and each index: we checked the performance to reproduce ClimDEX statistics for present climate using analyzed (test 1) and simulated (test 2) atmospheres; finally we checked each method's ability to respond to observed climate anomalies, such as those expected from future change (test 3). To summarize the main findings: all temperature related indices pass about twice as many tests as the precipitation indices, and temporally more complex indices that involve consecutive days pass none of the tests; with

respect to regions, there is some tendency towards better performance at the coastal and mountain-top stations; with respect to methods, XDS performed best followed by (in descending order) BCSD, QRNN, ASD, and TG.

A major challenge for the current study was to connect those findings for present climate to the results for the future scenarios. For example, unlike for the present where the quality or adequacy for extremes of a method or simulation is immediately apparent from comparison to observations, no direct equivalent exists for the unobserved future, where all one has is an array of projections for any particular location and scenario in question. There is the possibility of a 'surrogate' future climate, however, if high-resolution simulations are available that can play the role of nature. This is increasingly the case for global climate models (GCMs) driving regional climate models (RCMs) of higher resolution (~50 km and finer). Although this resolution will not reflect proper local extremes it offers an interesting path for testing statistical downscaling methods [*Vrac et al.,* 2007].

Without any a priori knowledge about the projections, a first guess for the future mean climate is given by the average across all projections. Deviations from this mean are composed of at least four factors:

SCN: The particular choice of emission scenario.

GCM: The global climate model that was driven by **SCN**.

DSC: The downscaling method.

LOC: The particular location of interest.

It is probably helpful to memorize these bold-capital acronyms of the four factors, because they play a central role in this study and appear frequently in formulas and figures; we will switch between acronym and its colloquial meaning as it fits.

All four factors contribute in a specific way to the simulated change for any ClimDEX index, confounding the signal in a complicated way. One way to disentangle confounded signals of this type is by using *analysis of variance* (ANOVA). ANOVA comes in two flavors, one that is purely descriptive by giving influence estimates for the various factors, and a second one that is inferential by also providing significance estimates for each factor influence. The inferential flavor can be used to establish significant factor contributions in a noisy environment, such as climate predictability [*Zwiers*, 1996], but it relies on a number of conditions (normality of residuals, homogeneity of variance, etc.) that are not easily met in experimental setups like ours. But this is irrelevant insofar as our main question is not *whether* a factor has a (significant) influence but *how large* it is. Therefore, we employ ANOVA in the simple descriptive way similar to, e. g., [*Li et al.*, 2011]. In addition to that study we include **LOC** as an independent factor, which enables us to estimate the influence of location on the climate signal relative to the other sources. To our knowledge, no other study has previously subjected all four main sources of uncertainty, **SCN**, **GCM**, **DSC**, and **LOC** to a fully factorial ANOVA approach (§2.1). The ANOVA is accompanied with two related techniques; the first, influence of components (§2.2), puts a stronger focus on the single components, e.g. a specific **GCM**; the second, multidimensional scaling (§2.3), emphasizes the downscaling factor and renders a more geometric picture of the group of methods. We had also considered including natural variability as an extra factor but decided

against it, mainly in order to keep the study confined. Using *20 y* averages should largely moderate natural variations, with residual variations showing up as **GCM** variability.

By considering all factors with equal weight we are deliberately disregarding, for now, the results of dip1 or, for that matter, any other a priori knowledge that might affect the different factors (such as **GCM** performance for present climate, see also [*Giorgi and Mearns*, 2002]). We will, however, return to this important point later and discuss how dip1 fits into the overall picture. It is best to think of dip1 and this dip2 as providing independent evidence pro or contra a specific model setting.

Of the numerous studies devoted to assessing climate scenario uncertainty, global and local, [*Déqué et al.*, 2007] is probably the study closest to our approach. It came out of the European project PRUDENCE [*Christensen et al.*, 2007] whose main goal was to obtain high-resolution climate scenarios and corresponding uncertainty in order to improve climatic impact assessments of extremes. [*Déqué et al.*, 2007] analyze similar sources of uncertainty (they do not consider location and use dynamical instead of statistical downscaling) for a simulated shift in mean seasonal temperature and precipitation for Europe, and find the **GCM** to be the major source. [*Sain et al.*, 2011] analyze corresponding uncertainties within the North American Regional Climate Change Assessment Program (NARCCAP). A main characteristic of NARCCAP is the (almost) fully factorial GCM-RCM simulation matrix, which allows disentangling the various GCM and RCM influences on the final result. Towards this goal, [*Sain et al.*, 2011] employ a 2-dimensional (‘functional’) ANOVA design to obtain maps of the main factors. [*Schmidli et al.*, 2007] conduct a detailed intercomparison of various statistical and dynamical downscaling techniques (see also dip1). As men-

tioned, [*Li et al.*, 2011] use a similar ANOVA approach for the analysis of regional climate models and statistical emulators thereof.

In this study, we test a broad range of temperature and precipitation related extremes as measured by the set of 27 core indices, ClimDEX. The ClimDEX indices (cf. *http://www.climdex.org*), listed in *Table 1*, do not generally reflect the most extreme events conceivable, but instead represent 'the more extreme aspects of climate' which are a) known to be relevant to a broad range of impact fields [*Peterson*, 2005] and b) still manageable statistically so that they can be reliably estimated from current data for present and future. With both aspects in mind ClimDEX has been adopted as a standard for extremes by the World Climate Research Programme (*http://www.clivar.org/organization/extremes*) and will be used accordingly in the fifth assessment report of the Intergovernmental Panel on Climate Change (IPCC) [*Zhang et al.*, 2011].

For this study the downscaling methods from dip1 are augmented by four further methods that are widely used in the impact community. Likewise, instead of the three dip1 climate zones with six stations, the methods are now applied to a total of twenty stations covering eight different regions, ranging from coastal to mountainous to sub-arctic climate.

§2 Data and Methods

Each of the 27 ClimDEX indices of *Table 1* are calculated from daily values of precipitation, P, and minimum and maximum temperature, T_n and T_x, respectively; daily mean temperature will be denoted by T. Observed values of ClimDEX were calculated for 20 stations from the Environment Canada station network, as listed in Table 2. The selection of

stations was guided by two criteria: a) representative coverage of the study area and b) data completeness for present climate (at least 90% coverage for each variable between 1981 and 2000). This represents a considerable extension of the dip1 set of 6 stations.

The downscaling methods of dip1 that are tested here are

BCSD: Bias Correction Spatial Disaggregation [*Salathe Jr et al.*, 2007; *Wood et al.*, 2002]

QRNN: Quantile Regression Neural Networks [*Cannon*, 2011]

TG: TreeGen [*Stahl et al.*, 2008]

XDS: Expanded Downscaling [*Bürger*, 1996].

Details of the methods are described in dip1. Another frequently used method is the Statistical DownScaling Model [SDSM, *Wilby et al.*, 2002]). Its automated version, Automated regression-based Statistical Downscaling (ASD), was part of dip1 and would have fit very nicely into the scope of this follow-up study. But because it exists only in closed-source form (MatLab 'p-code') it could not be adapted to handle large numbers of simulations; hence neither SDSM (which is open source but not automated) nor ASD are covered here, unfortunately. Additionally to dip1, we have included four other methods:

BioSim: A stochastic weather generator [*Régnière and St-Amant*, 2007]

CDFt: A method using transfers of cumulative distribution functions [*Michelangeli et al.*, 2009]

DQM: detrended quantile mapping, a BCSD version without spatial and temporal disaggregation.

LARS-WG: A stochastic weather generator [*Semenov and Barrow*, 1997].

The main technical ingredients of all methods are summarized in Table 3. We should note that the sheer amount of simulations occasionally required quite inventive computing techniques, such as the automated pushing of buttons for two of the methods.

All downscaling methods were calibrated using NCEP I reanalysis fields [*Kalnay et al.*, 1996]. Projections of future climate were obtained from six GCMs, listed in Table 4, which belong to the multi-model dataset of the third Coupled Model Intercomparison Project (CMIP) conducted by the World Climate Research Program [*Meehl et al.*, 2007] and were selected based on the availability of predictor fields, the main limiting factor being daily upper level fields. The corresponding simulations for present climate are based on estimates of the relevant forcing agents for the 20th century (20C3M), and those for future climate on the well known scenarios B1, A1B, and A2 from the Special Report on Emission Scenarios (SRES) of the IPCC [*Nakicenovic and Swart*, 2000]; each of these simulations was downscaled using six methods, details of which are summarized in Table 3.

We analyze changes of annual ClimDEX values between the periods 1981 to 2000 for present and 2046 to 2065 (as set out by CMIP3) for future climate. We first calculate index-specific anomalies based on the operation Δ or Δ% as specified in *Table 1*, column 5, as follows: Relative to the index mean value for the present, C_0, we calculate annual index anomalies relative to C_0 either as a simple difference, denoted Δ, or, for many of the precipitation related quantities, as a relative change Δ/C_0, denoted Δ%. This way the anomalies

become somewhat more standardized. We are under no illusion, however, that this has a large effect on some of the highly non-Gaussian indices such as tropical nights or ice days (TR, ID); but fortunately these are exceptions, as will be confirmed below. But the standard deviation even of the transformed indices still depends on their original physical units, which hinders, among other things, a comparison across indices. By calculating the t-value of a corresponding test for the differences of means we obtain a climate change signal in each index that has (roughly) zero mean and unit variance; to allow for changing variances we use the t-statistic as described by [*Welch*, 1947].

For each combination of **SCN**, **GCM**, **DSC**, and **LOC** this defines a mapping

$$(\mathbf{SCN}, \mathbf{GCM}, \mathbf{DSC}, \mathbf{LOC}) \mapsto \quad \Delta CDX_t \tag{1}$$

The mapping (1) gives a total of 3 (emission scenarios) ×6 (GCMs) ×8 (downscaling methods) ×20 (locations) = 2880 estimates for each index change for the future climate of British Columbia. For each index, therefore, we can consider its mean change, that is, how it looks in the 2050s in BC in general, and how the 4 factors create variation around this mean change. While the entire study can be viewed as a sensitivity experiment with four independent agents, it should be noted that uncertainty originating from **LOC** on the one side and **SCN**, **GCM** and **DSC** on the other are inherently different: **LOC** effects are expected and physically consistent whereas any effect of the other factors on the outcome represents an unwanted uncertainty, as it entails a deviation from the truth.

First we study the overall mean change for each index. We then analyze the variation about this mean and the influence of the factors, using three different techniques all of

which are related but focus on a different aspect: analysis of variance (§2.1), influence of components (§2.2), and multidimensional scaling (§2.3).

§2.1 Analysis of variance

In the analysis of variance (ANOVA) approach one tries to explain the overall variation of some quantity x_i from contributions of a finite number of *factors*, each assuming a finite number l_f of *levels*. This approach is particularly simple and appealing in *balanced* experiments where the number of observations ('responses') is constant across all factors and levels [e.g. *Toutenburg*, 2009]. If this 'cell count' is n, the overall sum of squared variations V can be decomposed into independent (orthogonal) contributions of the single factors, as

$$V = \sum_f \sum_{l=1}^{l_f} n(\bar{x}_l^f - \bar{x})^2 + \varepsilon^2 \qquad (2)$$

with $\bar{x}_l^f$,$\bar{x}$ denoting cell and total mean of the x_i and ε^2 the residual sum of squared errors; note that ε^2 describes the unresolved variance within each factor level. The contribution of each factor can thus be expressed as a ratio to the overall variation V, which is usually called *explained variance* (EV) and measured in percentage. Since we employ a fully factorial ANOVA the condition of equal cell count is satisfied and Eq. (2) can be applied. We shall conduct the ANOVA using the four single factors, **SCN**, **GCM**, **DSC**, **LOC**, and the 6 factor 'interactions' **SCN**×**GCM**, **SCN**×**DSC**, **SCN**×**LOC**, **GCM**×**DSC**, **GCM**×**LOC**, **DSC**×**LOC**. One could further decompose the residual variance into 3-factor and 4-factor interactions, but those are difficult to interpret so we do not use them; note that [*Li et al.*, 2011] introduce all interactions but actually never use them.

§2.2 Influence of components

By way of ANOVA, a given total variance of some index change ΔCDX_t is decomposed into the variations of factors. ANOVA does not provide information on how that variation changes when factor levels are added; that is, how the total variation is influenced by each single factor level. For example, from ANOVA we know that downscaling has a strong influence *in general*, but how any particular method affects the variation remains unknown. To fill this gap we perform an extra analysis on the influence of each single component of the model chain on the overall uncertainty. We do this in a differential way, by calculating for any index the relative change of variance introduced by adding a single component. Let σ^2 denote the variance of ΔCDX_t across all simulations (that is, the normalized V from Eq. (2)), and for any component C, such as C = CGCM3 or C = BioSim, let $\sigma^2_{\neg C}$ denote the variance of ΔCDX_t across all simulations *except* those where C is involved. Our measure is then defined as

$$\text{influence of C} \quad = \quad \frac{\sigma^2}{\sigma^2_{\neg C}} \,. \tag{3}$$

This measure takes positive values, with values <1 indicating a damping and values >1 an amplifying influence on the variance. Note that this always relates to the variations from the other components of that factor, that is for C = CGCM3, from all other GCMs, so that factors with many components (levels) such as **LOC** will show a smaller influence (on average). By the same reason, influence from components that belong to the same factor should roughly have unit mean (with deviations caused by the nonlinear variance operation). We calculate this ratio for all indices of ClimDEX and all 3+6+8+20=37 single model components from **SCN**, **GCM**, **DSC**, and **LOC**.

§2.3 Multidimensional scaling

Each single downscaling method is characterized by an array of 3×6×20=360 different numbers, consisting of all possible combinations of scenarios, GCMs, and locations. Reducing such wealth of criteria based on some ad-hoc argumentation for or against some of the components, e.g. to discard some GCMs, is never really free of cherry-picking, so we avoid this. Towards simplicity, there are several ways to accomplish this type of reduction, most of them based on some fairly general mathematical principles: How can a set of 8 (**DSC**) points in a 360-dimensional space be represented mathematically in a space of much lower dimension, without loosing too much information? - In climate research one immediately thinks of principal component analysis (PCA) with its numerous applications for atmospheric or oceanic fields. Out of a large sample of realizations PCA extracts the main directions (i.e. patterns) of variability – and there are usually only a few - and projects each single case onto these few axes to obtain a low-dimensional representation. Hence naturally PCA depends on a large sample size, so that in our case of 'only' 8 downscaling methods it is not applicable. A very general approach to the problem of dimension reduction is one that tackles the high-dimensional geometry solely through the concept of *mutual distance*. For any group of points living in a high-dimensional space the main question then is: How can the set of mutual (Euclidean) distances be *realized* by another group of points in some other, preferably low-dimensional space, and how accurate is that approximation? This is called Multidimensional scaling (MDS), and it provides a concise and usually quite illustrative way of describing high-dimensional *dissimilarities* in a simple plot of low dimensions (two or three is often enough).

Specifically, given the mutual dissimilarities of n entities in a matrix, $D=(d_{ij})$, the d_{ij} are approximated by the Euclidean distances of n 'real' points z_i in a low-dimensional space, the approximation being performed based on some measure of closeness. Using least-squares, one has to minimize the so-called stress function

$$S(z_1,\ldots,z_n)=\sum_{i\neq j}(d_{ij}-\|z_i-z_j\|)^2 \tag{4}$$

It is known that problems of this type are complicated by the presence of local minima, so that standard optimization recipes such as gradient descent methods often become trapped in these minima [*Groenen and Heiser*, 1996]. A method that appears to be apt to this special form (4) of the cost function is the so-called "Scaling by majorizing a convex function" (SMACOF) [*De Leeuw and Heiser*, 1977]. To minimize (4), SMACOF iteratively replaces the cost function by suitable smooth convex functions and applies standard techniques of convex analysis to optimize those (convex functions have only one global minimum). SMACOF has proved to be more resilient for the stress optimization (4), although local minima are hard to overcome in general. But note that even the global minimum is only unique up to a group of orthogonal (Procrustes) rotations. To further guard against local minima we have applied MDS multiple times, using 50 random initializations reflecting the general scale of distances, and selecting from the appropriately rotated solutions the one with minimum stress function.

We will employ MDS for the 8 downscaling methods. Any particular method is characterized by 2880/8=360 ΔCDX_t values per index; by grouping them into T and P indices this accumulates to 360×16=5760 and 360×11=3960 values, respectively. These are the dimen-

sions of the space in which we take the Euclidean distance of any two downscaling 'points' as a dissimilarity measure.

§3 Results

All ClimDEX definitions are based on the three variables P, T_x, and T_n. Several indices, moreover, are given relative to a base period that serves to represent the climate normals. For example, TN90p (warm nights) describes the ratio of days with T_n being warmer than the normal (present) upper 10% quantile (relative to calendar day). For a future scenario, therefore, the signal size depends both on the projected anomaly itself *and* on the base variability. The above ratio can easily be calculated if the effect of climate change is a uniform shift of the entire distribution, by simply solving the corresponding integral equations for the distribution function. For Gaussian quantities the ratio is then a simple function of the future shift relative to the present standard deviation. In the case of TN90p, for which normality is a reasonable, albeit heuristic first order approximation, the ratio grows from 10% for zero change to values near 80% for a shift of 2 standard deviations. With or without normality, the future mean change relative to the present variability provides a good heuristic for the change of extremes in P, T_x, and T_n. Present variability is calculated as the standard deviation of each series after removing the seasonal cycle (as anomalies per calendar date). We show this heuristic for all simulations, grouped by downscaling, in Figure 2. A few things require attention: First, especially for T_x and T_n the LARS-WG markers appear shifted to the left, as compared to the other methods, which indicates a loss of simulated variability. Also note that, for the low variability range LARS-WG simulates fairly large mean changes, which inevitably has an impact on the projected extremes, as we shall see.

At the high variability range we see quite different scales for the methods. For example, maximum *P* variability is much larger for QRNN, TG and XDS, approaching and for QRNN exceeding 12 mm/d, and for T_x and T_n, maximum variability is especially low for LARS-WG (less than 6 and 5 deg, resp.). Again, this lack of variability is bound to produce large extremes. Note also that near the scale of *8 mm/d*, LARS-WG produces an obvious outlier with very large mean signals for one particular station (which happens to be the mountain station 117CA90 at 1875 m altitude).

For each location, observed and simulated variability are compared in Figure 3. It confirms that LARS-WG underestimates present temperature variability, showing the largest deviations with root mean square (rms) values of about 0.7. Interestingly, the *P* variability agrees best for this method, which may point to problems in deriving correct temperature values from the wet and dry spells. The Figure also shows an overestimation of T_x and T_n variability for TG and an overestimation of *P* variability for QRNN. Note also the T_n outliers for BioSim.

An interesting feature of Figure 2 is the increase of the projected T_x and T_n change with variability, which is evident at least for QRNN and XDS. A closer inspection reveals that such a nonzero proportionality is indeed seen in all methods, with varying degree. As Figure 4 shows, all simulated temperature signals show this proportionality to the simulated (present) variability, with significantly (α=1%) nonzero slopes in almost all cases. A similar but weaker proportionality holds for *P*, except for QRNN and XDS, interestingly, for which the T proportionality was strongest. This is unlikely a common artifact of all methods, but instead indicates real (mathematical, physical) phenomena. First, the simulated mean change may to some extent be proportional to the overall scale of variability, which would

apply especially for the long-tailed *P* distribution. From a more physical reasoning, proximity of the ocean and corresponding larger thermal capacity leads to attenuated temperature variability, a tendency that is also seen for the 20 locations of this study (not shown); moreover, for the decadal timescales pertinent to radiative heating, different warming rates of land and sea are the result of a more effective evaporative heat-loss over the wet ocean surface, an argument that goes back to [*Manabe et al.*, 1991].

As suggested by Figure 2, XDS in particular tends to project negative *P* signals, especially so for the sites with large variability (analogous to the case for temperature). Apart from this, larger signals of decreasing *P* are only simulated by LARS-WG and TG. Because of its general importance we did an extra analysis for *P*, as shown in Figure 5. For any combination of **GCM** and **DSC** it displays the simulated mean SRESA1B changes, denoted $\Delta P($**GCM**,**DSC**$)$. For better resolution we show the results in two dimensions, by projecting each original value $\Delta P($**GCM**,**DSC**$)$ onto two slightly different axes: $x=\lambda\langle\Delta P($**GCM**,**DSC**$)\rangle+(1-\lambda)\bullet\langle\Delta P($**GCM**$,:)\rangle$ and $y=\lambda\langle\Delta P($**GCM**,**DSC**$)\rangle+(1-\lambda)\bullet\langle\Delta P(:,$**DSC**$)\rangle$, with '$\langle...\rangle$' denoting average and using a weight of $\lambda=0.5$. It shows that negative signals are produced mainly by GFDL2, the largest being $\Delta P(GFDL2,XDS)$. Most positive signals come from CGCM3, with moderate signals from QRNN, TG, and XDS and larger ones from the rest. The same **DSC** clustering is apparent from all other **GCM**.

Each index is now analyzed in terms of the t-value, ΔCDX_t, of the mean difference between the 20 simulated annual values of future (2046-2065) and present climate (1981-2000), leading to 2880 simulated changes for each index. It turned out that for all indices except CSDI and TR and a few cases of ID, most of the annual index anomalies are in fact Gaussian, according to a Kolmogorov-Smirnov test (not shown); in these cases ΔCDX_t is t-

distributed. Unlike the more standard case of a t-test with equal variance, the distribution parameters in this case, including the significance level for nonzero values (=climate change), depend on the sample variance. This dependence of the significance level turns out to be quite weak, however, with a mean value of 2.73 and a standard deviation of 0.03 for the α=0.01 level.

Despite having a unified scale for all index changes now - roughly unit sampling variance (= unchanging climate) - the projected change proves to be very different for the temperature and precipitation related indices, to which we refer simply as T and P indices, respectively. We will discuss both separately. The overall change for each index, using all of **SCN**, **GCM**, **DSC**, and **LOC**, is shown in the boxplot of Figure 6. We remind the reader that each box represents the inter-quartile range (IQR, between the 25% and 75% quantiles) of the sample; sample minimum and maximum are indicated by the whiskers, unless those extremes are beyond 1.5×IQR in which case a "•" is displayed to indicate an 'outlier' (an "x" for 3×IQR). This phrase should not be taken in a literal statistical sense because, if the results can be interpreted at all as coming from a random distribution, that distribution is very likely non-Gaussian. By an outlier we merely indicate simulations that may require special attention. Along with the boxplots we have indicated the level of significance of any single change being nonzero, by using the constant mean value of 2.73 (see above) across all indices. This is only to indicate the scale that single random simulations may attain and as such does not pertain to e. g. the significance of the overall mean.

The T indices show the behavior expected in a warmer climate, that is, signals that are significantly decreasing for FD, ID, TN10p, TX10p, and increasing for most of the others, and this frequently applies to the entire IQR of the index. Most of the changes of the P indi-

ces are insignificant or ambiguous, with both increasing and decreasing tendency. The 1% significance level shown in the Figure is based on a normally distributed quantity, a condition that is not necessarily met here.

Figure 7 shows the results for the individual downscaling methods. For the T indices, the signals shown in Figure 6 are generally reproduced by the single methods, with varying amplitudes. As can be seen from the different axis scale and significance strip, BioSim and LARS-WG have exceptionally strong signals, and the other methods share similarly moderate signals. The ΔCDX_t values are outside the significance band for most indices, similar in direction between methods, but differing in magnitude, indicating unique responses for all simulations; only CSDI and DTR are ambiguous across all methods. The opposite is true for the P indices, which also reaffirms the results for the mean result (Figure 6), with little significance for the main body (IQR) of the simulations. Note that QRNN and TG, but especially XDS produce a number of significantly negative signals (CWD, PRCPTOT, R10mm, R20mm, R95p).

Because temperature signals are relatively large compared to the 'noise', but also because of their greater normality, the ΔCDX_t values are generally larger for the T indices (see discussion above). Moreover, probably due to their sensitivity to the simulated present variability 'outliers' are found more often for these indices (see Figure 2). This is almost certainly the case for several of the extreme outliers such as CSDI, ID, TN10p, TN90p, TXx, WSDI, which all originate from LARS-WG. This method generally produces less interannual variability than the others, which leads to very strong change signals for these indices. While the outliers for T are generally in the direction of the overall signal of the IQR, P out-

liers are seen in both directions, reinforcing their overall uncertainty. The strongest positive P outliers are produced by CDFt and DQM and the strongest negative ones by QRNN and especially by XDS.

For the most extreme positive and negative ΔCDX_t values we show the corresponding simulations in Figure 8, separately for *T* and *P* indices. For *T*, we see two striking examples of very strong positive and negative index changes as projected by LARS-WG. One is warm nights (TN90p, SRESA2, CGCM3, 1021480), whose frequency increases from 10% (by definition) to about 80%, and the other is frost days (FD, SRESA1B, MIRO3, 1125700), whose number decreases from over 100 to less than 40. The case for TN90p (SRESA2, CGCM3, 1021480), which is a quantile-based index, happens to correspond to a minimum of simulated present T_n variability, cf. Figure 2, so that, following the normality argument outlined above, even a moderate shift in the mean can lead to very large increases in the index. The FD case is less obvious but also threshold based, so it may come from the same reduced variability. For *P*, the strongest positive outlier is also from LARS-WG, which projects daily intensity (SDII, SRESA1B, CGCM3, 1090660) to increase from *6.5 mm/d* to *8.0 mm/d*; this will further be discussed in §4. The negative precipitation projections from GFDL2 downscaling, and here especially by XDS, were already mentioned around Figure 5.

§3.1 Analysis of variance

The results of the 4-way ANOVA are shown in Figure 9. It is obvious that **DSC** has the strongest influence, especially for the T indices; **DSC** often explains more than 50% of the variation while for the P indices that value varies around 10%. For *P*, the main source of

variation comes from the **GCM** (along with **GCM•LOC**) which explains 20-30% for PRCPTOT and R10mm. Generally, the contribution of **LOC** alone is marginal, with the exception of ID with about 20% and particularly TR with more than 40% EV. But there is a sizable influence of the combined factors **DSC•LOC**, with EV values of at least 10% across all indices and more for the T indices. There is some influence of **SCN** on the T indices, especially SU, TNx, and TXx with EV values of about 10%; there is hardly any influence on the P indices. Figure 9 also reveals that variations in the T indices are overall better explained by this ANOVA (4-way with simple interactions), where several of them exceed levels of 80% of EV. P indices vary about 50% EV.

§3.2 Influence of components

The results for the influence of the single components is presented in Figure 10. Please note that a damping (values < 1) or amplifying (values > 1) influence says nothing about the direction of the signal (such as warming or cooling), only about the increasing or decreasing uncertainty. First, the lesser relative influence of the 20 locations is obvious, almost all showing values near 1 (remember that influence depends on the number of factor components, see §2.2). There are a few exceptions, though, for ID (1090660) and TR (1125700, 1123992). We have seen the ID instance also in two outlier cases of Figure 8, which may point to a data problem for the location 1090660 (although some ad-hoc checking did not reveal anything obvious). A more likely explanation is, however, unreliable statistics due to poor sampling for some stations (too few cases for some climates), which would explain the TR instance as well. Obviously a very large influence on the variation comes from LARS-WG, by strongly affecting all T indices, and partly from BioSim with a

large impact on minimum and maximum temperatures (TNn, TNx, TXn, TXx), but also CDD. Of the downscaling methods, XDS has an amplifying influence on the P index uncertainties. This is likely related to the fact that XDS produces negative P signals in several cases, see Figure 7. The strongest source of GCM uncertainty comes from CGCM3, affecting most of the Tn indices, followed by GFDL2 with an impact on some P values. CNRM3 has some affect on TR, interestingly. Note the marked and widespread damping of all T index variations from most of the downscaling components, to counter the amplification from BioSim and LARS-WG. With respect to scenarios, SRESB1 has a damping and especially SRESA1B an amplifying influence. Note that this must not be mistaken as a cooling or drying influence; it points, instead, to enhanced uncertainty from the overall larger signal amplitudes in the case of SRESA1B.

§3.3 Multidimensional scaling

Figure 11 shows the results of the MDS conducted on the T and P indices. With regard to the T indices, LARS-WG and BioSim are obvious outliers. The other methods are much closer, with a central cluster containing DQM, CDFt and BCSD and another cluster formed by QRNN, TG and XDS; these secondary clusters are of course somewhat arbitrary. This constellation reflects the findings of Figure 7 where BioSim and LARS-WG showed by far the strongest signals. The results for the P-indices show a larger spread. DQM and CDFt are fairly close again and occupy the center, being surrounded uniformly by the other methods.

Please remember that the MDS axes do not represent physical dimensions (such as a climate change signal) but a mathematically optimal two-dimensional embedding of the

original 5760 (for T) and 3960 (for P) dimensional space. The quality of this embedding is shown in Figure 12. For each pair of downscaling "points" it displays their Euclidean distance in the original vs. that in the reduced space. A clustering of distances is noticeable for the T indices, which is likely due to the 'outlying' role of BioSim and LARS-WG. Relative to this expansion of distances the T indices appear to be better approximated.

§3.4 Selecting downscaling methods

As evident from Figure 9, the 8 downscaling methods have the strongest influence on the overall uncertainty of the scenarios. It seems natural, therefore, to try to reduce this influence by selecting only some of the methods. But to make that selection, additional independent evidence is needed. The most obvious criterion is the performance for present climate, which was done in dip1 with a not fully overlapping set of methods. Here we take the three best-performing methods of dip1, namely XDS, QRNN, and BCSD, and repeat the ANOVA of Figure 9 with these. For comparison, we take another three methods, namely CDFt, DQM, and TG which occupy the center of MDS (cf. Figure 11). The results are shown in Figure 13. Evidently, in both cases the influence of downscaling is sharply reduced compared to Figure 9, which at least for T is most likely due to the removal of BioSim and LARS-WG from the **DSC** set, and now **GCM** forms the main source of uncertainty. This holds for both T and P indices. For the verified methods, **DSC** and especially the coupled factor **DSC•LOC** still shows some influence, varying at about 20% in total; for the three methods from the MDS center the influence of **DSC** has practically vanished, leaving only **GCM** and **GCM•LOC** as the main source of variation. But given that the selection is

based on MDS similarity this is of course to be expected, since ANOVA and MDS provide rather similar and partly redundant information.

§4 Discussion

We first turn our attention to the overall simulated mean for each index. The results were generally very different for the T and P indices. For the former, the main message is that the simulations tend to agree that extreme warm events are getting considerably more frequent and the opposite holding for cold events; this is of course no surprise and hardly requires any downscaling to determine [e.g. *Kharin and Zwiers*, 2000]. For the latter, the bulk of the simulated changes are insignificant. Several significant outliers, nevertheless, simulate increasing or decreasing heavy precipitation, as discussed further below.

It was the purpose of this study to analyze and understand the variations about this mean signal, with a particular focus on the downscaling methods. And in fact, downscaling turned out to be the major factor influencing the simulated change. Along with **LOC** as a combined factor, **DSC** explains more than 60% of the variation in the signal for many of the T indices. The strong dependence on **DSC** as a single factor shows that different methods have a uniform affect across the entire region, the effect of **LOC** being only secondary. The P indices, on the other hand, show a stronger susceptibility to the **GCM** (~10%). The influence of **SCN** is negligible in both cases, which is somewhat expected due to comparatively little difference in emissions for the 2046 to 2065 time horizon, especially for the P indices. From the influence and MDS analyses it became apparent that the large spread especially for the *T* indices was caused by somewhat ’outlying’ index projections of BioSim and LARS-WG.

This outlying behavior of BioSim and LARS-WG could be traced back to a bias in simulated present variability, which has a strong influence especially on all quantile-based indices. In both cases variability is generated purely stochastically from a parametric weather generator. LARS-WG simulates the length of dry and wet spells as a Poisson process and derives all other variables from that process; a misfit of the corresponding exponential parameters therefore has large consequences; moreover, interannual variability is not (explicitly) simulated at all. The LARS-WG cases of Figure 8 may all be related to this, including precipitation intensity (SDII) which depends on the number of wet days. BioSim, on the other hand, generates stochastic weather sequences from an array of long-term monthly statistics, including interannual, but lacks about 15% of the variance [*Régnière and St-Amant*, 2007].

The P indices, on the other hand, did not show comparable systematic outliers. The extreme low end was a 40% reduction of PRCPTOT projected by XDS using GFDL2 (SRESA2, 1090660). This must be seen in the context that GFDL2 is the driest of the GCMs, with QRNN, TG, and XDS being the driest of the GFDL2-driven methods. These are the methods that make use of upper level predictor fields, which in the case of GFDL2 contain fairly large undefined grid cells over the Rocky Mountains. Accordingly, these potentially crucial gridpoints cannot be used for the NCEP calibration and may render the methods sub-optimal in some cases. But as long as there is no objective criterion our judgment may not even be relevant or, in fact, the downscaling may be even more consistent with the projected GFDL2 fields; at least the corresponding calibration statistics do not indicate otherwise. In any case, this points to the general sensitivity of QRNN, TG, and XDS to the usually large set of potential predictors and, consequently, their exposure to over- or underfitting.

The overestimation of P variability by QRNN and of T_x and T_n variability by TG was likely caused by an imperfect choice of independent predictors. Note, however, that this fact does not per se invalidate the corresponding (negative) P projections. The remaining methods (BCSD, CDFt, DQM) did not show any major issues. Being based on quantile mapping they are generally closest to the driving GCM, and their calibration consists of "merely" finding the best mapping of the respective distributions.

Many of the facets of the different downscaling methods, it thus appears, derive from their particular methodological setup. Moreover, they roughly correspond to the clusters of the MDS analysis, at least for the T indices. We summarize therefore the main features of these three groups in Table 5.

All methods, especially QRNN and XDS, exhibit a proportionality of simulated present variability and future signal for T_x and T_n, which, as we have seen, is equivalent to larger warming rates inland. All methods, with the interesting exception of QRNN and XDS, also show a proportionality for P, which is likely a simple scaling effect from the long-tailed distribution (but why not for QRNN and XDS?), and perhaps also some influence of location since the increase was mainly at the coast. Note, however, that the proportionality of P is weak after all.

The described projection spread changes drastically if the analysis is confined to downscaling methods that were established as reliable through independent verification. Using only XDS, QRNN, and BCSD as the most reliable methods of dip1, the influence of downscaling was strongly diminished, leaving GCMs as the main source of uncertainty. Unfortunately, the remaining methods (BioSim, CDFt, DQM, LARS-WG) were not part of

dip1 and are thus untested, and testing them here as in dip1 was beyond the scope of the current study.

This enforces the need for independent verification of all components. In fact, only *after* all parts of the model chain have undergone thorough validation is it justified to view the corresponding set of projections as an *ensemble* in a statistical sense, with no criterion left to constrain the projections any further and rendering them as truly indistinguishable. With such an ensemble of projections a fully *inferential* ANOVA becomes possible. Using the likely setting of **SCN**, **GCM**, and, **DSC** as random and **LOC** as a fixed factor, one could establish the influences of these factors in a strict statistical sense. But note that while dip1 provided a quite thorough verification against present climate that all non-tested methods should undergo, some questions remain whether that will sufficiently constrain the future projections. For example, some elements of the methods such as the detrending/retrending component of BCSD or TG, or any sort of bias correction, all of which crucially affect future simulations, can only be tested on a longer time horizon of several decades which is not (yet) available.

Given the strong seasonality that is evident in the study region, a useful extension of the present analysis would be to include the four seasons (or the 12 months), both in the testing for present climate and as an additional ANOVA factor for future projections. It should be noted, however, that the current intercomparison setup, whose GCMs were all taken from the CMIP3 suite, would be better suited to the new and extended suite of CMIP5 models.

An interesting future path of research is offered by the concept of surrogate future climate mentioned in the introduction, where methods are tested against the simulated high-resolution fields of an RCM. This approach is not limited to RCMs, of course, as high-resolution surrogate climates can be provided by *any* downscaling technique, including statistical. Following that path, the *testing* of a single method is turned into the mutual *consistency* of any two methods, including the consistency of a method with itself, and with end results that may yield consistency clusters resembling those of the MDS plots here. We are presently working towards putting this under a sound methodological framework.

Acknowledgment

Dave Spittlehouse initiated this study and dip1 through a proposal to the Future Forests and Ecosystems Scientific Council of BC, leading to financial support from the BC Ministry of Forests and Range; additional funding was provided from the BC ministry of Environment and from the University of Victoria. We are thankful to Francis Zwiers for helpful comments and to Hailey Eckstrand for preparing the figures.

Reference

Bürger, G. (1996), Expanded downscaling for generating local weather scenarios, *Climate Research*, *7*(2), 111–128, doi:10.3354/cr007111.

Bürger, G., T. Q. Murdock, A. T. Werner, S. R. Sobie, and A. J. Cannon (2012), Downscaling extremes - an intercomparison of multiple statistical methods for present climate, *Journal of Climate*, 120119124955001, doi:10.1175/JCLI-D-11-00408.1.

Cannon, A. J. (2011), Quantile regression neural networks: Implementation in R and application to precipitation downscaling, *Computers & geosciences*, *37*(9), 1277–1284.

Christensen, J. H., T. R. Carter, M. Rummukainen, and G. Amanatidis (2007), Evaluating the performance and utility of regional climate models: the PRUDENCE project, *Climatic Change*, *81*, 1–6, doi:10.1007/s10584-006-9211-6.

Déqué, M., D. Rowell, D. Lüthi, F. Giorgi, J. Christensen, B. Rockel, D. Jacob, E. Kjellström, M. De Castro, and B. van den Hurk (2007), An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections, *Climatic Change*, *81*, 53–70.

Giorgi, F., and L. O. Mearns (2002), Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the “Reliability Ensemble Averaging” (REA) Method, *Journal of Climate*, *15*, 1141–1158, doi:10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2.

Groenen, P. J. F., and W. J. Heiser (1996), The tunneling method for global optimization in multidimensional scaling, *Psychometrika*, *61*(3), 529–550.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, and J. Woollen (1996), The NCEP/NCAR 40-year reanalysis project, *Bulletin of the American Meteorological Society*, *77*(3), 437–471.

Kharin, V. V., and F. W. Zwiers (2000), Changes in the Extremes in an Ensemble of Transient Climate Simulations with a Coupled Atmosphere–Ocean GCM, *J. Climate*, *13*(21), 3760–3788, doi:10.1175/1520-0442(2000)013<3760:CITEIA>2.0.CO;2.

De Leeuw, J., and W. J. Heiser (1977), Convergence of correction matrix algorithms for multidimensional scaling, *Geometric representations of relational data*, 735–752.

Li, G., X. Zhang, F. Zwiers, and Q. H. Wen (2011), Quantification of uncertainty in high resolution temperature scenarios for North America, *Journal of Climate*, 111221122205005, doi:10.1175/JCLI-D-11-00217.1.

Manabe, S., R. Stouffer, M. Spelman, and K. Bryan (1991), Transient responses of a coupled ocean-atmosphere model to gradual changes of atmospheric CO2. Part I: Annual mean response, *J. climate*, *4*(8), 785–818.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor (2007), The WCRP CMIP3 multimodel dataset, *Bull. Am. Meteorol. Soc*, *88*, 1383–1394.

Michelangeli, P. A., M. Vrac, and H. Loukos (2009), Probabilistic downscaling approaches: Application to wind cumulative distribution functions, *Geophysical Research Letters*, *36*(11), L11708.

Nakicenovic, N., and R. Swart (2000), *Special Report on Emissions Scenarios*, Cambridge University Press.

Peterson, T. C. (2005), Climate change indices, *WMO bulletin*, *54*(2), 83–86.

Régnière, J., and R. St-Amant (2007), Stochastic simulation of daily air temperature and precipitation from monthly normals in North America north of Mexico, *International Journal of Biometeorology*, *51*(5), 415–430.

Sain, S. R., D. Nychka, and L. Mearns (2011), Functional ANOVA and regional climate experiments: a statistical analysis of dynamic downscaling, *Environmetrics*, *22*(6), 700–711, doi:10.1002/env.1068.

Salathe Jr, E. P., P. W. Mote, and M. W. Wiley (2007), Review of scenario selection and downscaling methods for the assessment of climate change impacts on hydrology in the United States pacific northwest, *International Journal of Climatology*, *27*(12), 1611–1621, doi:10.1002/joc.1540.

Schmidli, J., C. M. Goodess, C. Frei, M. R. Haylock, Y. Hundecha, J. Ribalaygua, and T. Schmith (2007), Statistical and dynamical downscaling of precipitation: An evaluation and comparison of scenarios for the European Alps, *J. Geophys. Res.*, *112*, 20 PP., doi:200710.1029/2005JD007026.

Semenov, M. A., and E. M. Barrow (1997), Use of a stochastic weather generator in the development of climate change scenarios, *Climatic Change*, *35*(4), 397–414.

Stahl, K., R. D. Moore, J. M. Shea, D. Hutchinson, and A. J. Cannon (2008), Coupled modelling of glacier and streamflow response to future climate scenarios, *Water Resources Research*, *44*(2), W02422.

Toutenburg, H. (2009), *Statistical analysis of designed experiments*, Springer Verlag.

Vrac, M., M. L. Stein, K. Hayhoe, and X.-Z. Liang (2007), A general method for validating statistical downscaling methods under future climate change, *Geophys. Res. Lett.*, *34*, 5 PP., doi:200710.1029/2007GL030295.

Welch, B. L. (1947), The generalization ofstudent's' problem when several different population variances are involved, *Biometrika*, *34*(1/2), 28–35.

Wilby, R. L., C. W. Dawson, and E. M. Barrow (2002), SDSM—a decision support tool for the assessment of regional climate change impacts, *Environmental modelling and software*, *17*(2), 145–157.

Wood, A., E. Maurer, A. Kumar, and D. P. Lettenmaier (2002), Long-range experimental hydrologic forecasting for the eastern United States, *Journal of Geophysical Research (Atmospheres)*, *107*, 4429.

Zhang, X., L. Alexander, G. C. Hegerl, P. Jones, A. K. Tank, T. C. Peterson, B. Trewin, and F. W. Zwiers (2011), Indices for monitoring changes in extremes based on daily temperature and precipitation data, *Wiley Interdisciplinary Reviews: Climate Change*, *2*(6), 851–870, doi:10.1002/wcc.147.

Zwiers, F. (1996), Interannual variability and predictability in an ensemble of AMIP climate simulations conducted with the CCC GCM2, *Climate Dynamics*, *12*(12), 825–847.

660

Tables

Table 1. ClimDEX indices.

index	Indicator name	Definitions	UNITS	change
CDD	Consecutive dry days	Maximum number of consecutive days with RR<1mm	Days	Δ
CSDI	Cold spell duration	Days with at least 6 consecutive days when T_n<Q_{10}	Days	Δ
CWD	Consecutive wet days	Maximum number of consecutive days with RR>=1mm	Days	Δ
DTR	Diurnal T range	Monthly mean difference between T_x and T_n	ºC	Δ
FD0	Frost days	Annual count when T_n < 0ºC	Days	Δ
GSL	Growing season Length	Days between first and last span of at least 6 warm enough days	Days	Δ
ID0	Ice days	Annual count when T_x < 0ºC	Days	Δ
PRCPTOT	Annual total wet-day precipitation	Annual total PRCP in wet days (RR>=1mm)	mm	Δ%
R10	Number of heavy precipitation days	Annual count of days when PRCP>=10mm	Days	Δ
R20	Number of very heavy precipitation days	Annual count of days when PRCP>=20mm	Days	Δ
R95p	Very wet days	Annual total PRCP when RR>95th percentile	mm	Δ%
R99p	Extremely wet days	Annual total PRCP when RR>99th percentile	mm	Δ%
R25	Number of days above 25 mm	Days when PRCP>25mm	Days	Δ
RX1day	Max 1-day precipitation	Monthly maximum 1-day precipitation	mm	Δ%
Rx5day	Max 5-day precipitation amount	Monthly maximum consecutive 5-day precipitation	mm	Δ%
SDII	Simple daily intensity index	Annual total precipitation divided by the number of wet days (PRCP>=1mm)	mm/day	Δ%
SU25	Summer days	Annual count when T_x > 25ºC	Days	*q*
TN10p	Cool nights	Percentage of days when T_n<10th percentile	%	Δ
TN90p	Warm nights	Percentage of days when T_n>90th percentile	%	Δ

TNn	Min Tmin	Monthly minimum value of daily minimum temp	ºC	Δ
TNx	Max Tmin	Monthly maximum value of daily minimum temp	ºC	Δ
TR20	Tropical nights	Annual count when T_n > 20ºC	Days	Δ
TX10p	Cool days	Percentage of days when T_x < 10th percentile	%	Δ
TX90p	Warm days	Percentage of days when T_x > 90th percentile	%	Δ
TXn	Min Tmax	Monthly minimum value of daily maximum temp	ºC	Δ
TXx	Max Tmax	Monthly maximum value of daily maximum temp	ºC	Δ
WSDI	Warm spell duration	Days with at least 6 consecutive days when $T_x > Q_{90}$	Days	Δ

665

Table 2. The 8 regions with the corresponding stations. The naming of regions is taken in part from dip1.

region	id	lon [deg]	lat [deg]	alt [m]
Campbell River	1021261	-125.27	49.95	106
	1021480	-125.43	50.33	23
	1046390	-124.55	49.88	52
Mountains	1154400	-116.05	50.88	1170
	1173210	-116.98	51.30	785
	117CA90	-117.70	51.23	1875
Taiga	1192940	-122.60	58.83	1201
Okanagan	1123992	-119.48	49.87	350
	1125700	-119.42	50.03	501
	1123970	-119.38	49.95	430
Fernie	1152850	-115.07	49.48	1001

	1158692	-115.47	49.47	762
Prince George	1090660	-121.52	53.07	1283
	1096450	-122.68	53.88	691
	1096630	-122.52	53.03	545
Vancouver	1101158	-122.92	49.28	366
	1103332	-122.57	49.27	147
	1108447	-123.18	49.20	4
Coast	1017230	-123.63	48.65	138
	1018620	-123.43	48.65	19

Table 3. The downscaling methods used. Large and small scales are abbreviated as LS and SS, respectively.

name	reference	characteristic
BCSD	*Wood et al., 2002*	- LS quantile mapping - spatial disaggregation - temporal disaggregation
BioSim	*Regniere and Bolstad, 1994*	- weather generator conditioned on 11 monthly statistics - *P* occurrence modeled from monthly total P and T range - disaggregation of P intensity - T_x and T_n anomalies as 2-dim AR(2) process
CDFt	*Michelangeli et al., 2009*	- common scales for LS and SS - cdf transformation (LS) present → future - cdf transformation LS → SS (present)
DQM		- detrending - quantile matching LS → SS (present) - retrending (simplified BCSD)
LARS-WG	*Semenov and Barrow, 1997*	- Poisson process of wet and dry spells - T_n and T_x stochastic Gaussian modeled from wet and dry spells, *P* intensity from semi-empirical distribution
QRNN	*Taylor, 2000*	- nonlinear quantile regression
TG	*Stahl et al., 2008*	- weather typing
XDS	*Bürger, 1996*	- probit normalization - covariance preserving regression - probit rescaling

Table 4. The set of GCMs used.

GCM	institution	resolution
CGCM3 T63	CCCma (CA)	T63
CNRM CM3	CNRM (FR)	T63
CSIRO MK 3_5	CSIRO (AU)	T63
GFDL CM 2_1	GFDL (US)	2.5°x2°
MIROC 3_2 MEDRES	JAMSTEC (JP)	T42
MPI_ECHAM5	MPI (DE)	T63

680 *Table 5. Characteristics and issues for the three main downscaling groups.*

downscaling group	main method	characteristic / issues
BioSim, LARS-WG	stochastic weather generator	(interannual) variability underestimated, hence many of the quantile-based indices likely overestimated
BCSD, CDFt, DQM	quantile mapping	everything inherited from nearest GCM grid point (T and P); minimum amount of calibration; BCSD with extra spatial and temporal disaggregation, has some issues from mimicing T_x and T_n from daily average and climatological range of T
QRNN, TG, XDS	predictor fields	sensitive to predictor selection (over- and under-fitting); possible issues for XDS with missing values and/or low resolution

Figure Captions

Figure 1. The study area with the typical regions ('Coast', 'Mountains', and 'Taiga' taken from dip1).

Figure 2. Mean future climate change vs. present variability, as simulated by the 8 downscaling methods. Per panel this gives 3 scenario points (y-axis) for each of 6×20=120 present-day simulations (x-axis). Variability is given as standard deviation after removing the seasonal cycle.

Figure 3. For each downscaling (rows) we show for the core variables P, T_x, and T_n (columns) the 6 **GCM** *simulations of present variability vs. the 20 observations from* **LOC***, with corresponding root mean square values. The line represents identity.*

Figure 4. Proportionality (fitted linear) of simulated present variability and future climate signal. Solid lines indicate a significantly positive slope. The axes scale is the same as in Figure 2.

Figure 5. Mean projected change of P from **DSC** *vs.* **GCM***, based on the A1B scenario. For better visibility, the single results for each (***GCM***,* **DSC***) pair, which would lie on the diagonal, are projected onto two slightly different axes, reflecting the* **GCM** *part (x-axis) and the* **DSC** *part (y-axis, see text).*

Figure 6. Mean change of T (top) and P indices (bottom). A box represents the IQR of the sample (see text); the range is indicated by the whiskers unless it is beyond the 1.5×IQR in which case a "•" is displayed to indicate an 'outlier' (an "x" for 3×IQR). The horizontal levels (dashed) of t = ±2.75 indicate the significance for any single simulation, if the index is Gaussian.

Figure 7. Range of simulated change in ClimDEX from the seven downscaling methods. Left: temperature related ClimDEX; right: precipitation related ClimDEX. The significance levels (dashed) are as in Figure 6.

Figure 8. Most extreme climate change results, with respect to temperature (left) and precipitation indices (right), and with increasing (top) or decreasing (bottom) tendency. Titles indicate emission scenario, driving GCM (Table 4), downscaling method (Table 3) and station (Table 2).

Figure 9. Contribution to variations in projected ClimDEX values, based on a 4-way ANOVA.

Figure 10. Influence of individual components on ClimDEX (cf. Eq. (3)). Blue colors indicate damping, red colors amplifying influence of spread.

Figure 11. Multidimensional scaling of the 8 downscaling methods for the T and P indices. The axes represent the optimum two-dimensional embedding of the original space of 5760 and 3960 dimensions, respectively (see text).

Figure 12. Original vs. MDS reduced Euclidean distance of the 8 **DSC** *"points", for T (left) and P indices (right). The line indicates identity.*

Figure 13. Same as Figure 9, but using only the verified downscaling methods BCSD, QRNN, and XDS (upper panel) or only CDFt, DQM, and TG which occupy the center of the MDS in Figure 11 (lower panel).

Figures

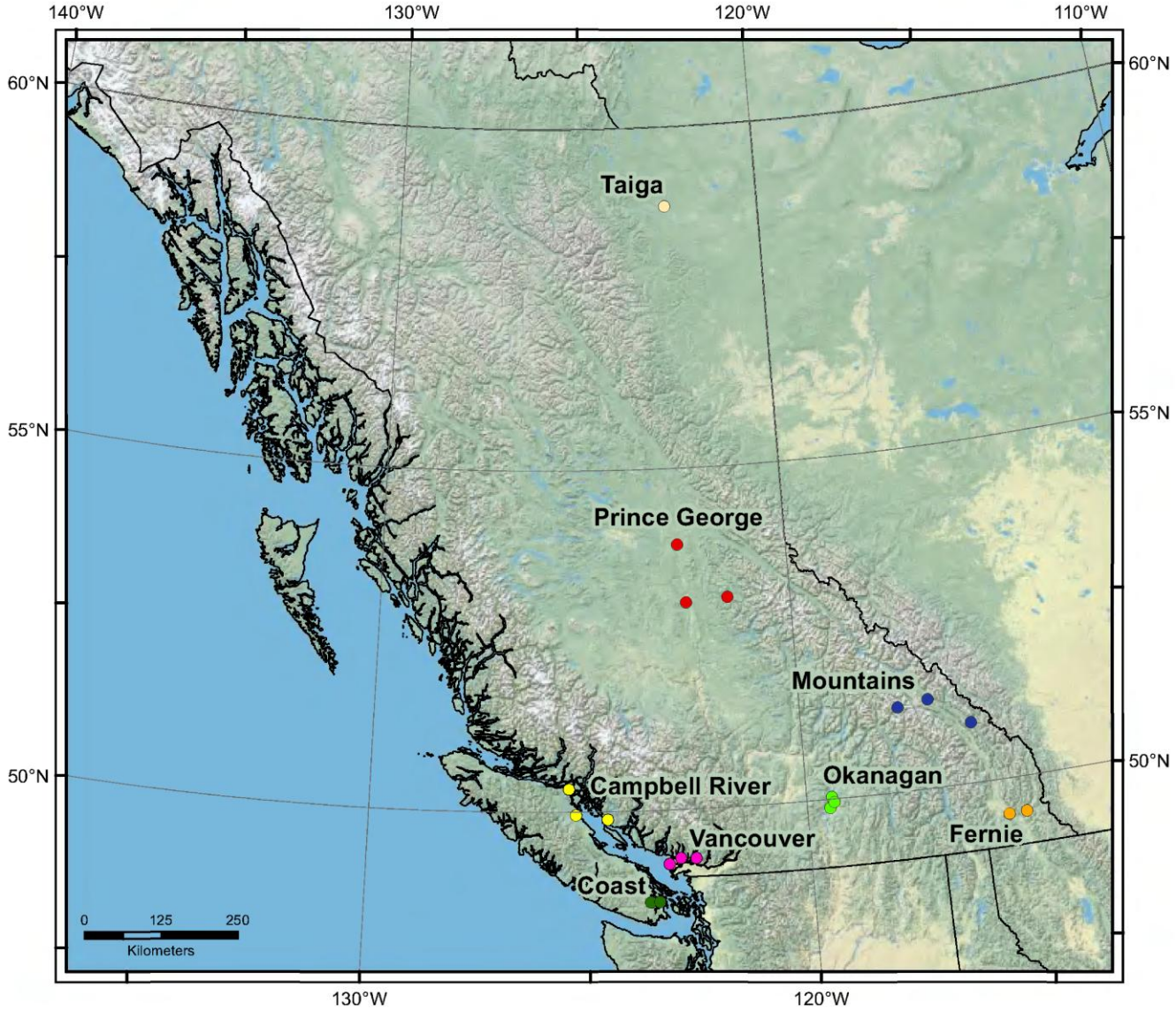

Figure 1. The study area with the typical regions (’Coast’, ’Mountains’, and ’Taiga’ taken from dip1).

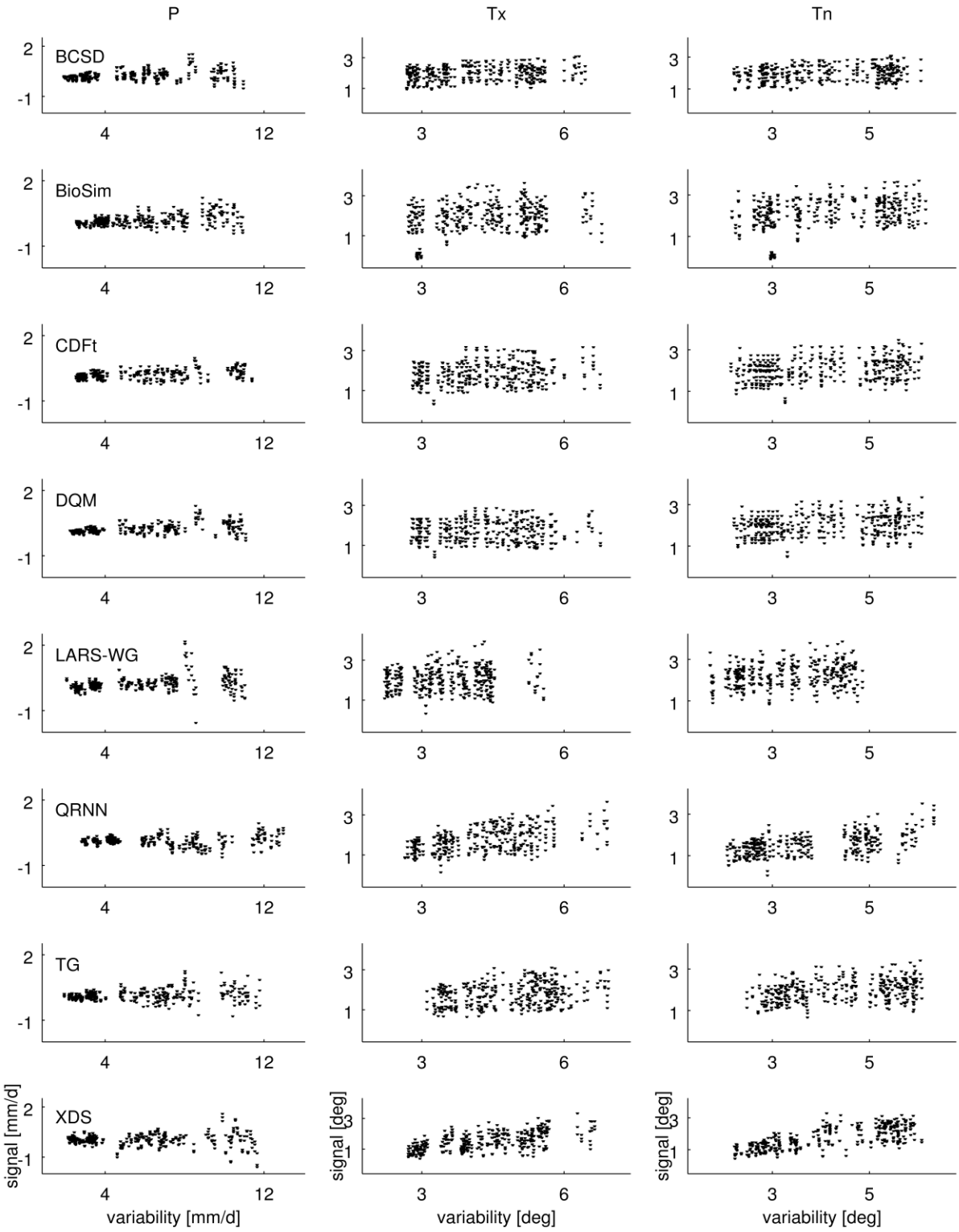

Figure 2. Mean future climate change vs. present variability, as simulated by the 8 downscaling methods. Per panel this gives 3 scenario points (y-axis) for each of 6×20=120 present-day simulations (x-axis). Variability is given as standard deviation after removing the seasonal cycle.

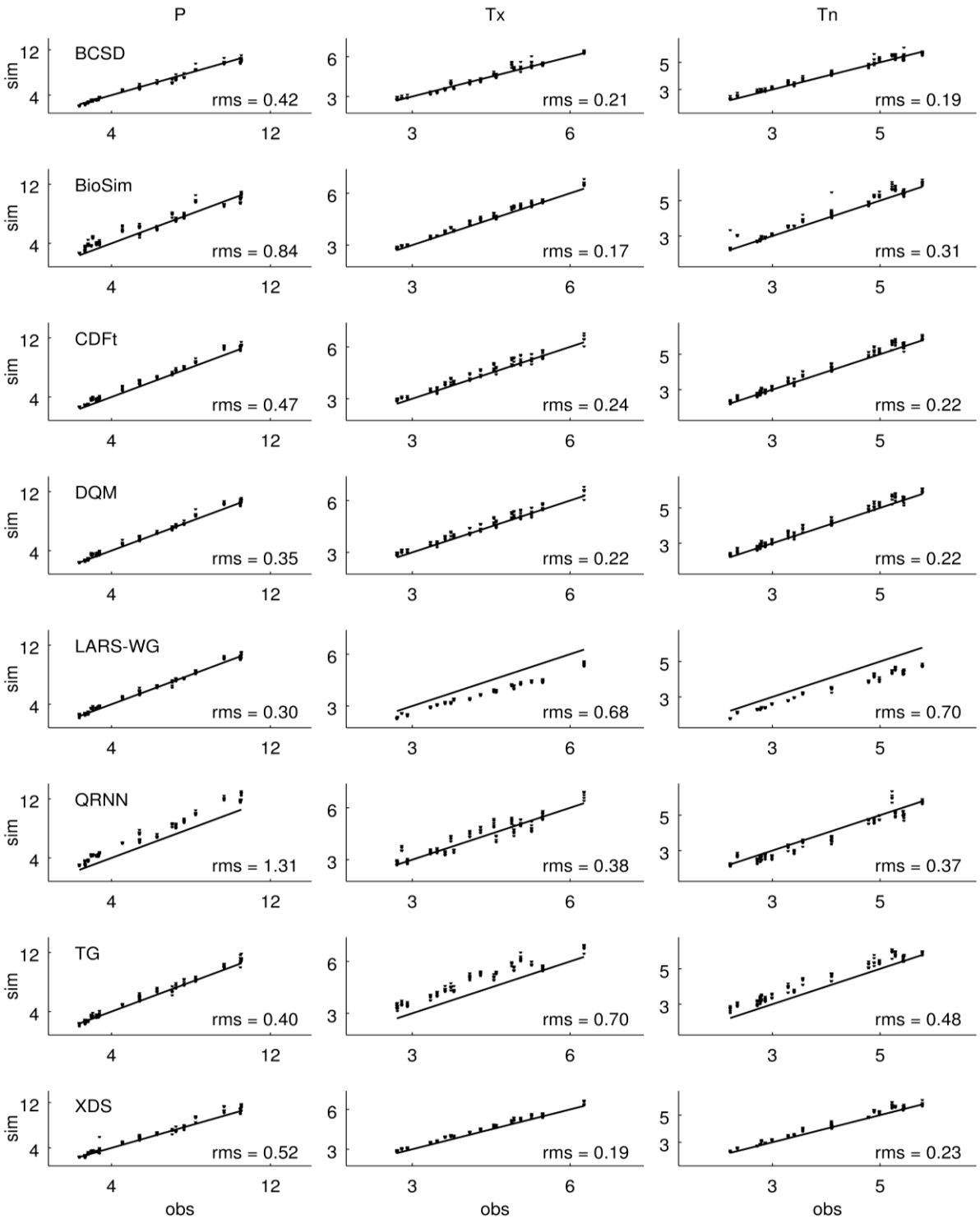

Figure 3. For each downscaling (rows) we show for the core variables P, T_x, and T_n (columns) the 6 **GCM** *simulations of present variability vs. the 20 observations from* **LOC***, with corresponding root mean square values. The line represents identity.*

720

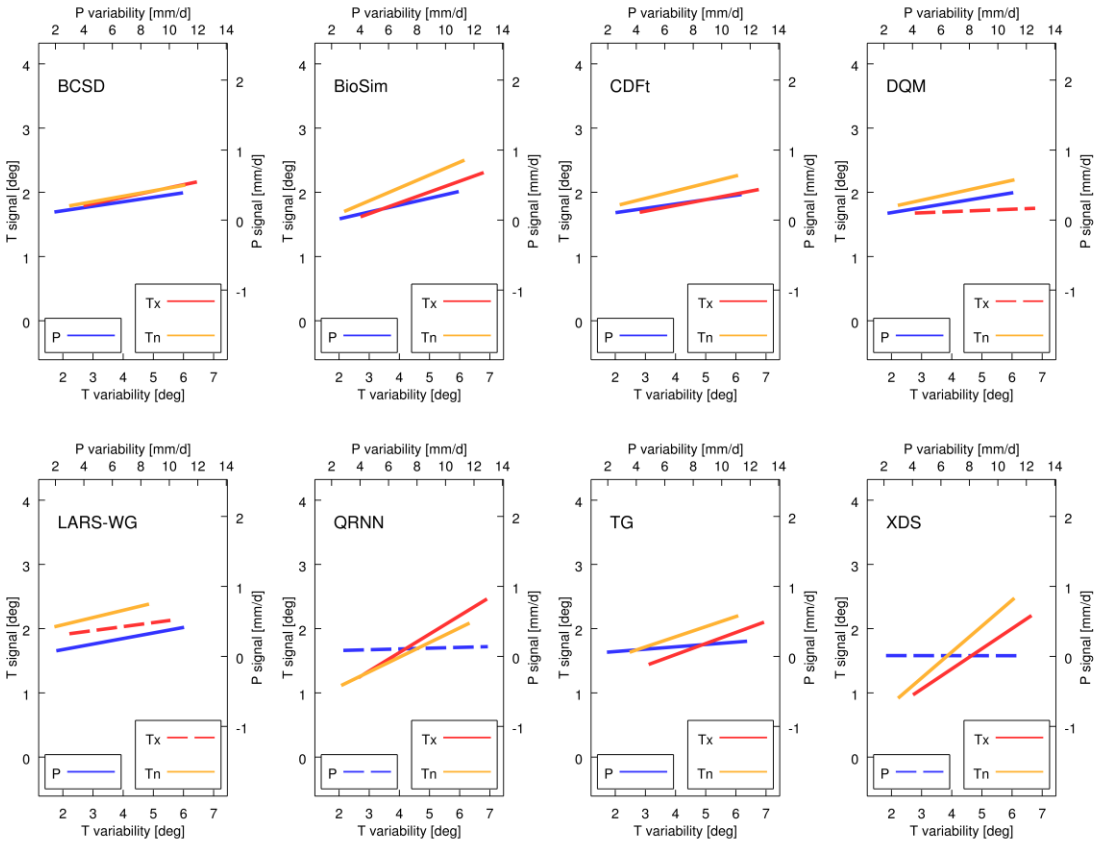

Figure 4. Proportionality (fitted linear) of simulated present variability and future climate signal. Solid lines indicate a significantly positive slope. The axes scale is the same as in Figure 2.

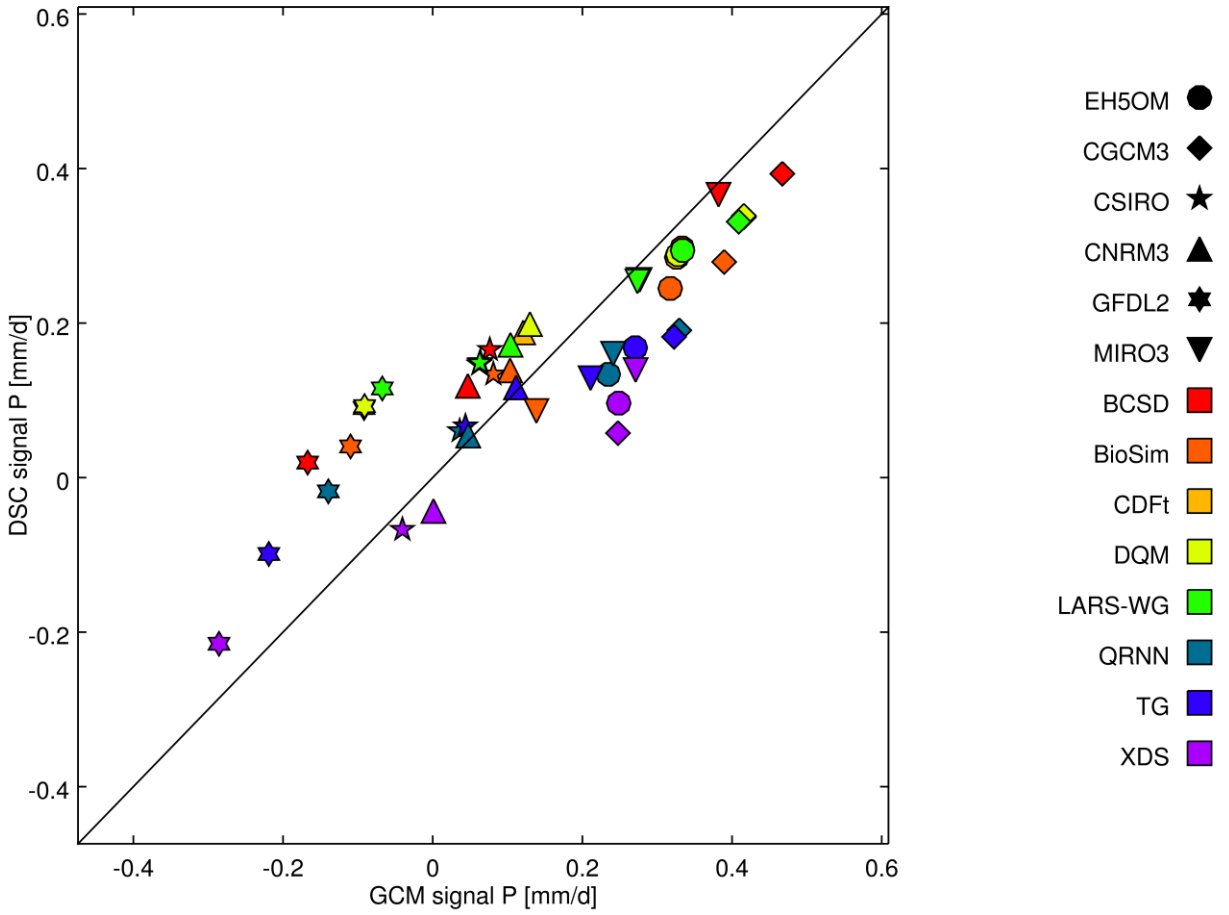

Figure 5. Mean projected change of P from **DSC** *vs.* **GCM**, *based on the A1B scenario. For better visibility, the single results for each (***GCM***, ***DSC***) pair, which would lie on the diagonal, are projected onto two slightly different axes, reflecting the* **GCM** *part (x-axis) and the* **DSC** *part (y-axis, see text).*

725

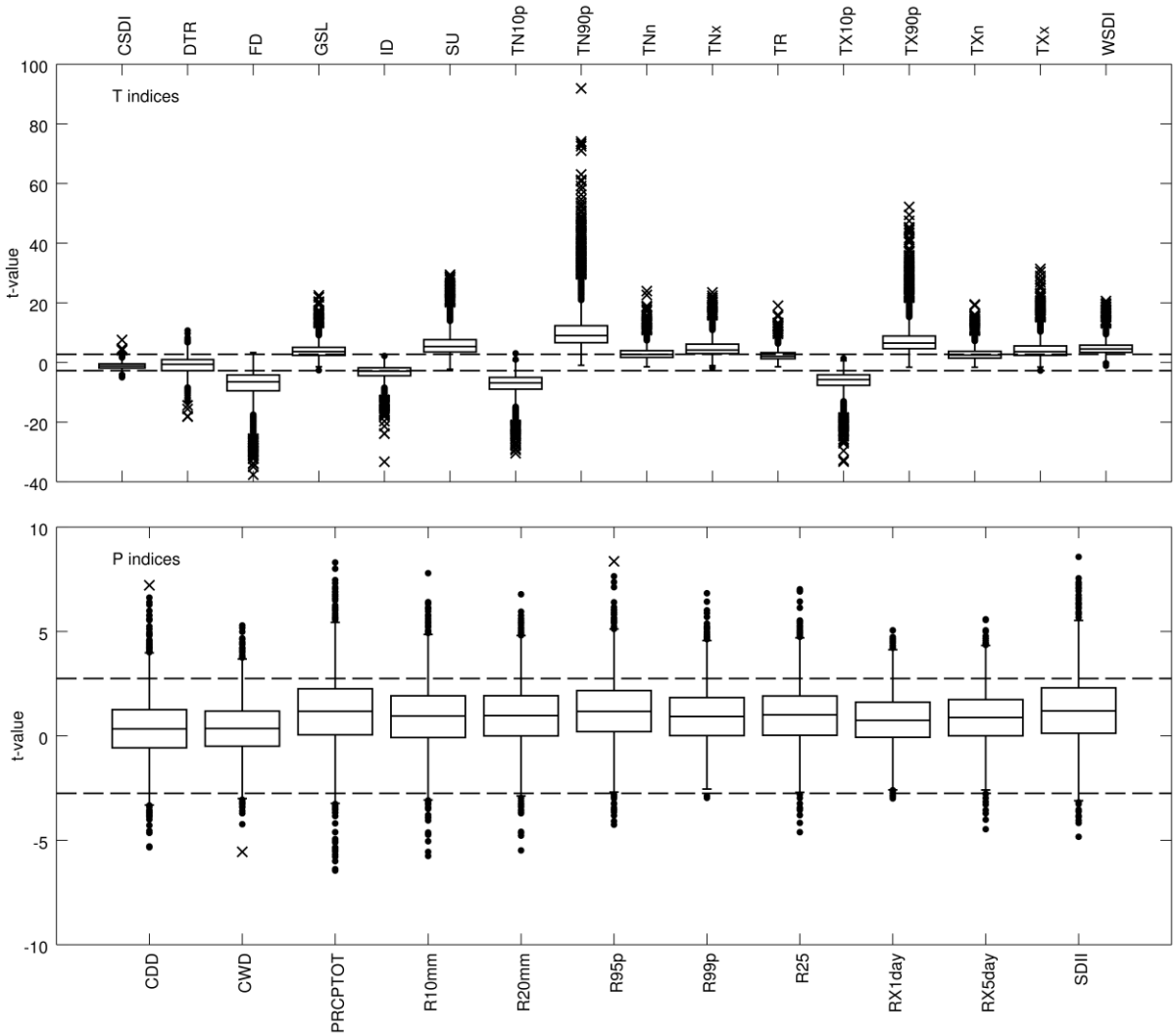

Figure 6. Mean change of T (top) and P indices (bottom). A box represents the IQR of the sample (see text); the range is indicated by the whiskers unless it is beyond the 1.5×IQR in which case a "•" is displayed to indicate an 'outlier' (an "x" for 3×IQR). The horizontal levels (dashed) of t = ±2.75 indicate the significance for any single simulation, if the index is Gaussian.

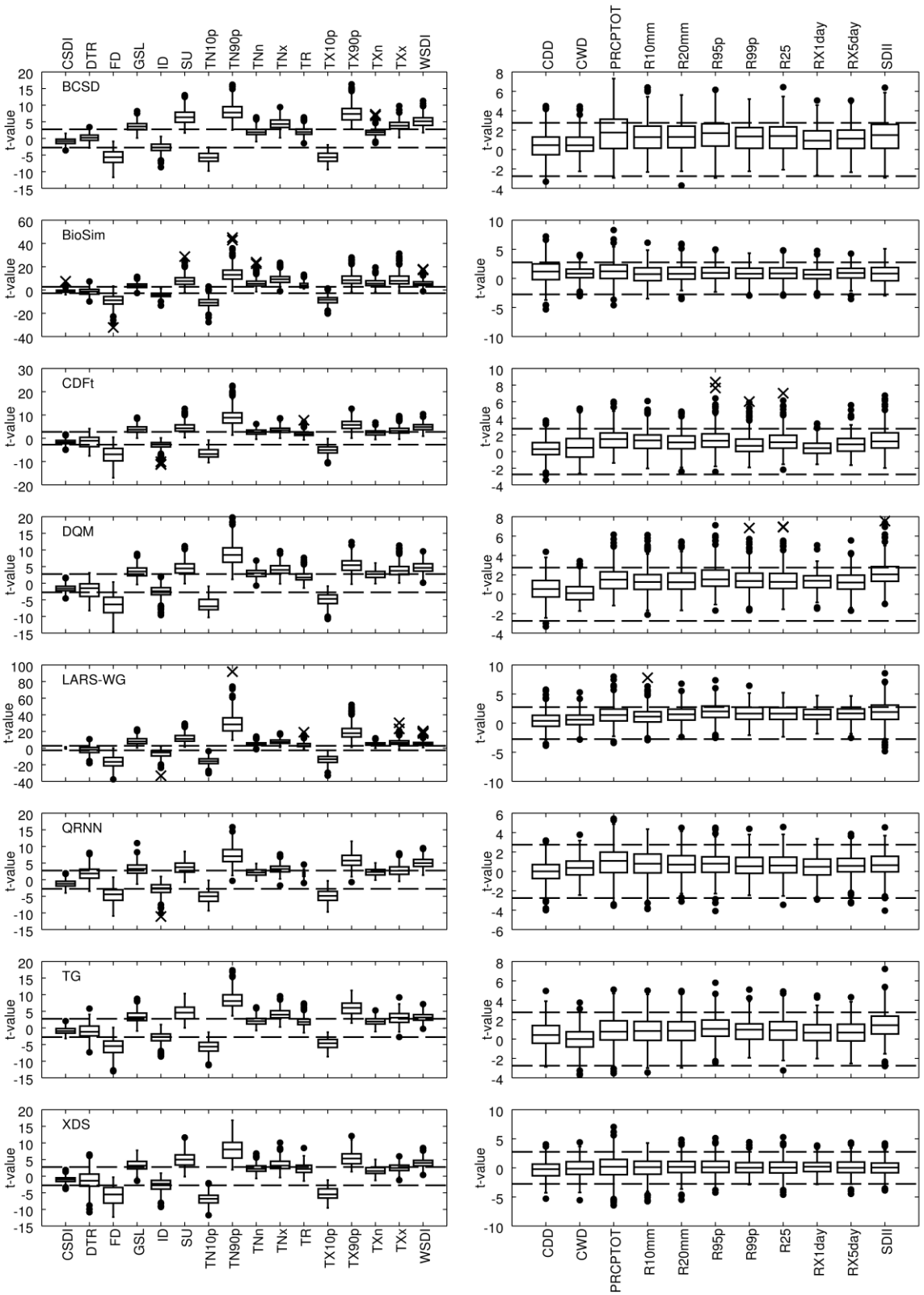

Figure 7. Range of simulated change in ClimDEX from the seven downscaling methods. Left: temperature related ClimDEX; right: precipitation related ClimDEX. The significance levels (dashed) are as in Figure 6.

730

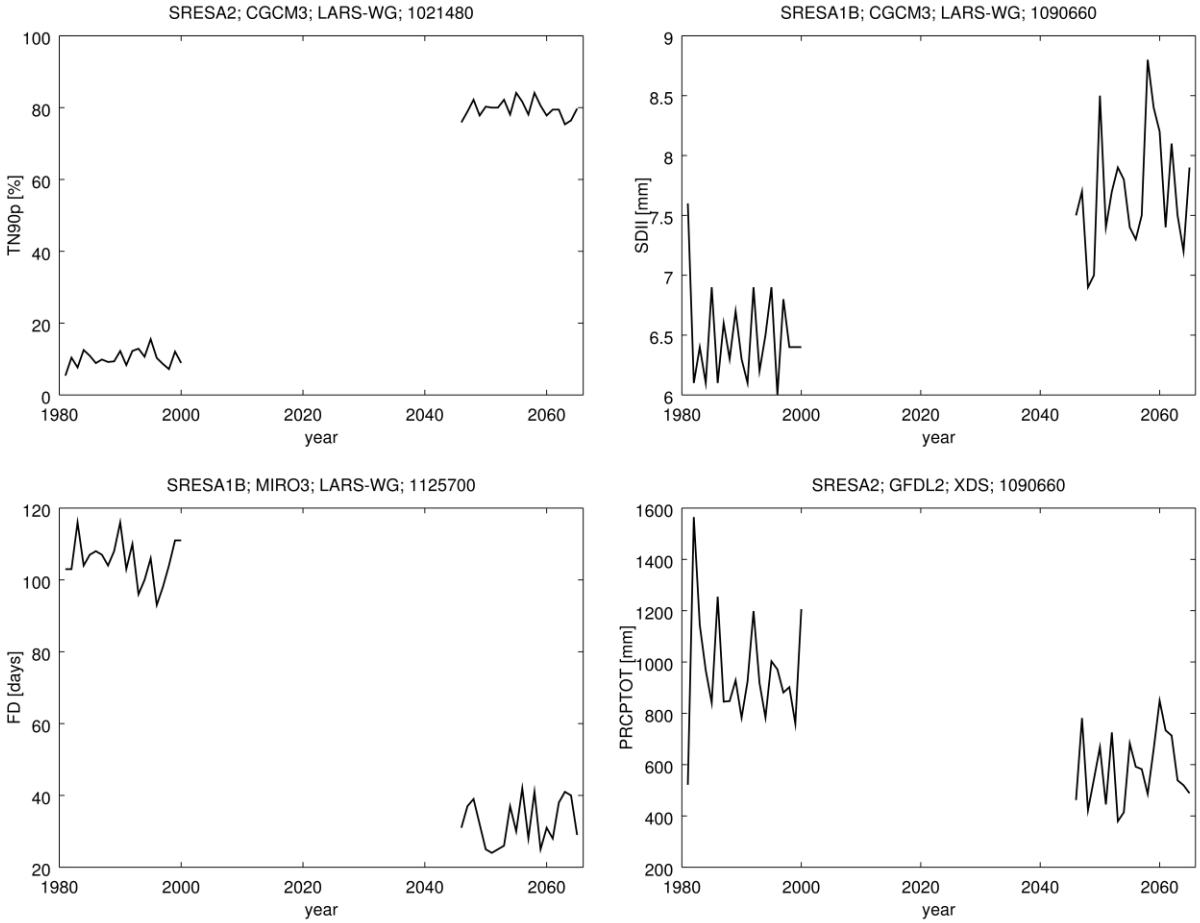

Figure 8. Most extreme climate change results, with respect to temperature (left) and precipitation indices (right), and with increasing (top) or decreasing (bottom) tendency. Titles indicate emission scenario, driving GCM (Table 4), downscaling method (Table 3) and station (Table 2).

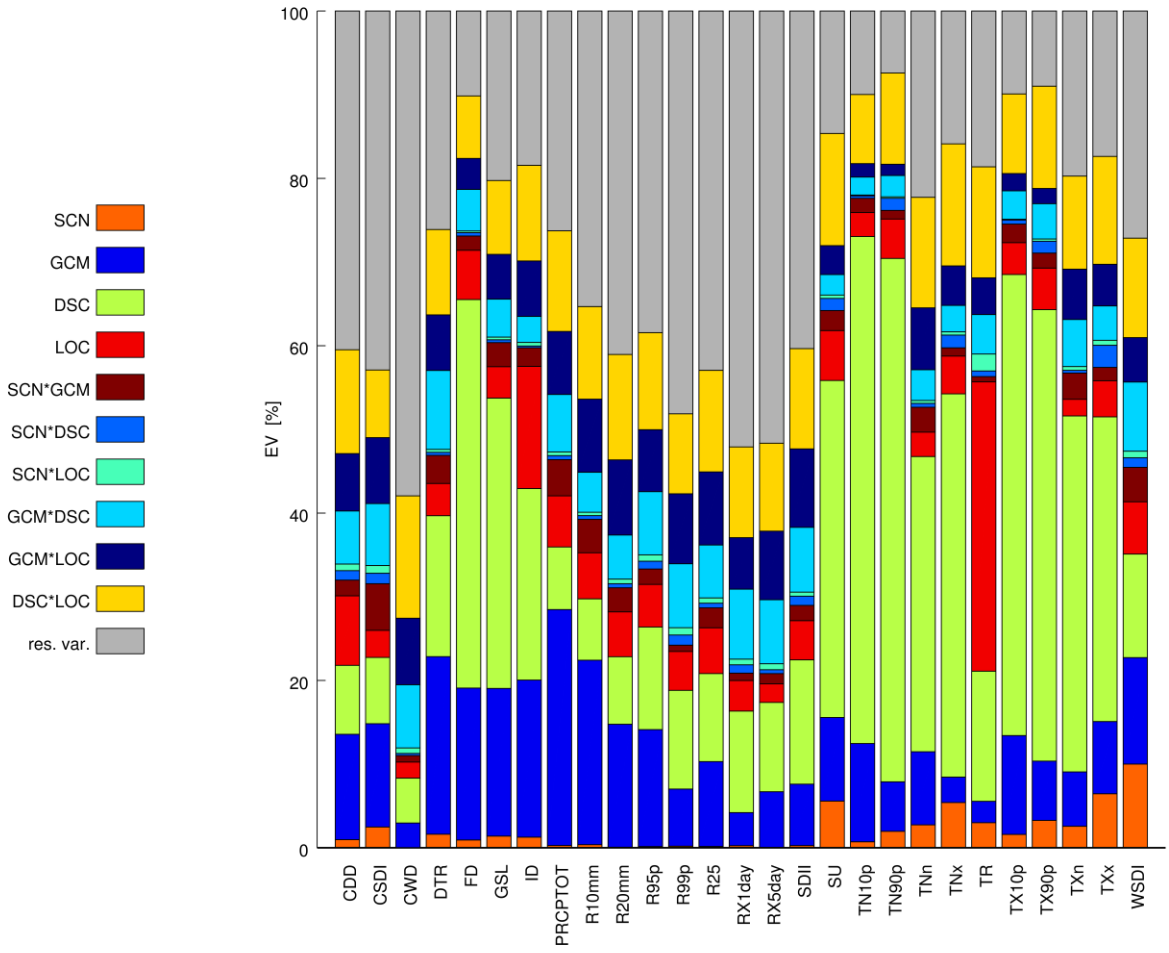

Figure 9. Contribution to variations in projected ClimDEX values, based on a 4-way ANOVA.

735

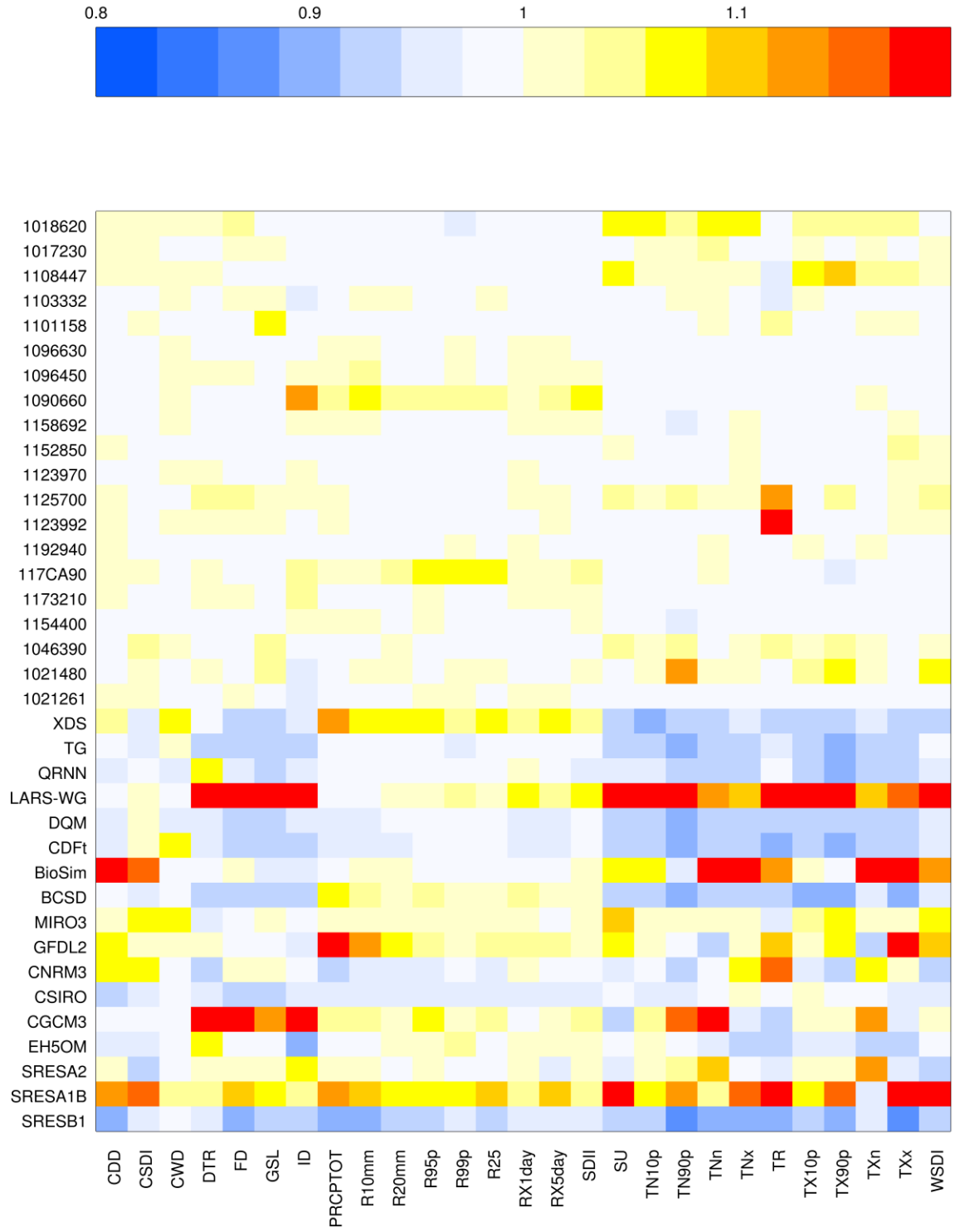

Figure 10. Influence of individual components on ClimDEX (cf. Eq. (3)). Blue colors indicate damping, red colors amplifying influence of spread.

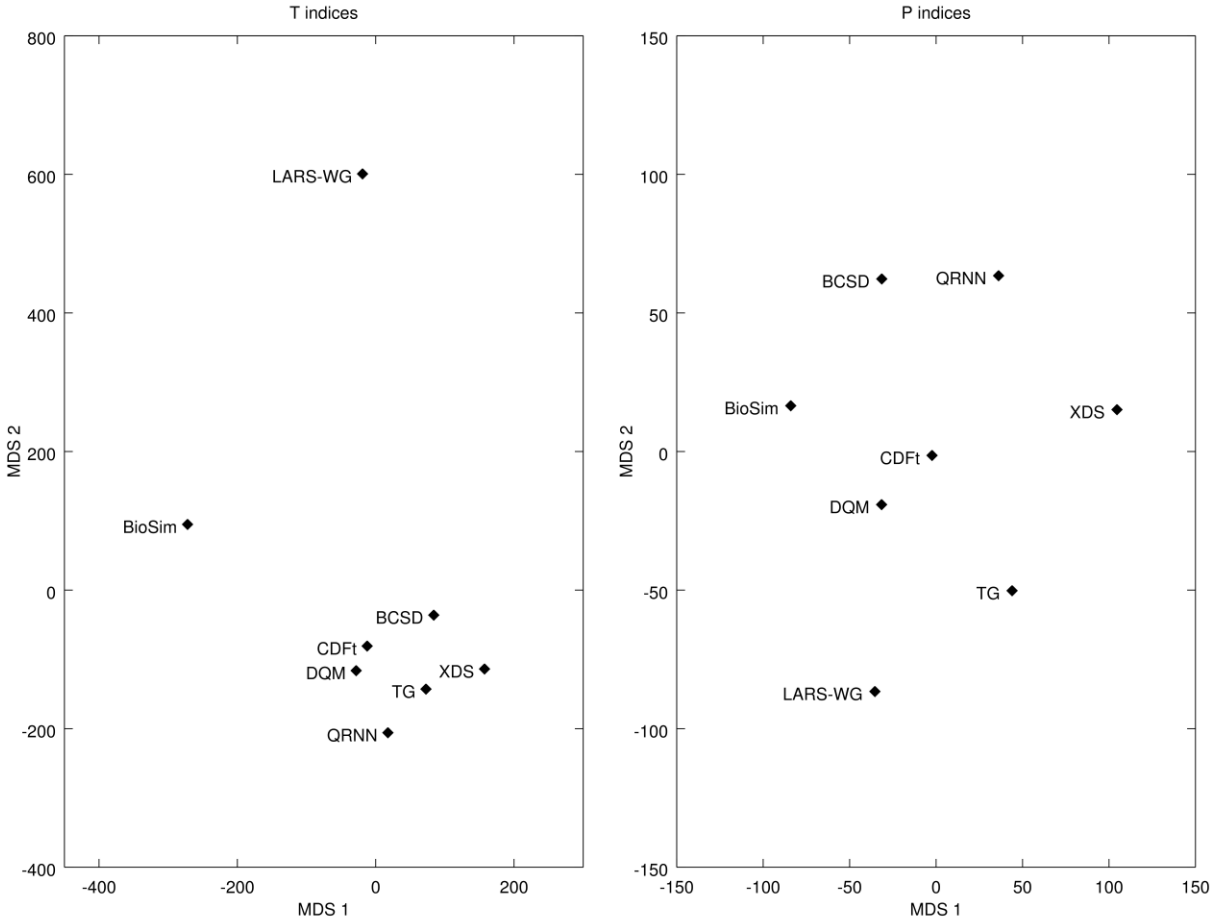

Figure 11. Multidimensional scaling of the 8 downscaling methods for the T and P indices. The axes represent the optimum two-dimensional embedding of the original space of 5760 and 3960 dimensions, respectively (see text).

740

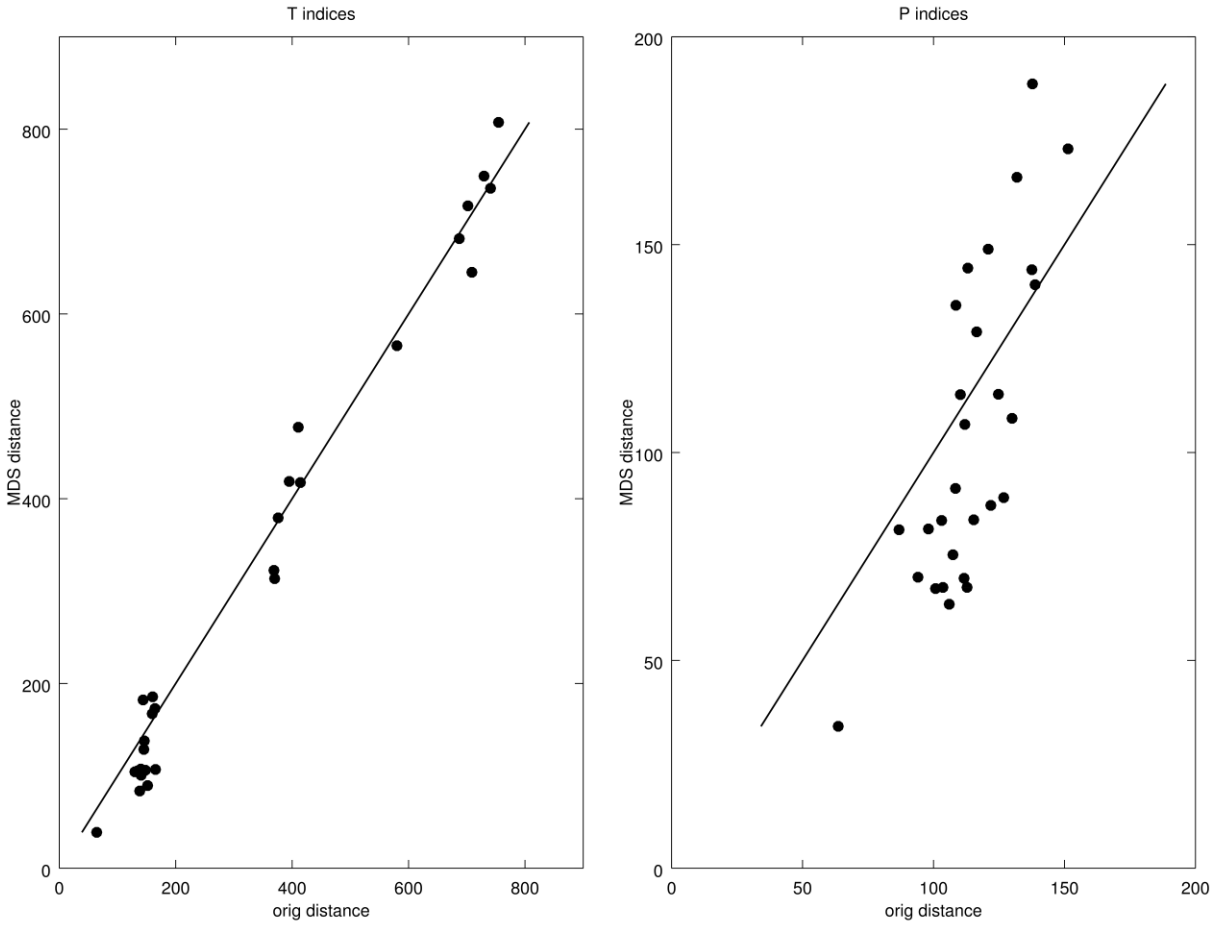

Figure 12. Original vs. MDS reduced Euclidean distance of the 8 **DSC** *"points", for T (left) and P indices (right). The line indicates identity.*

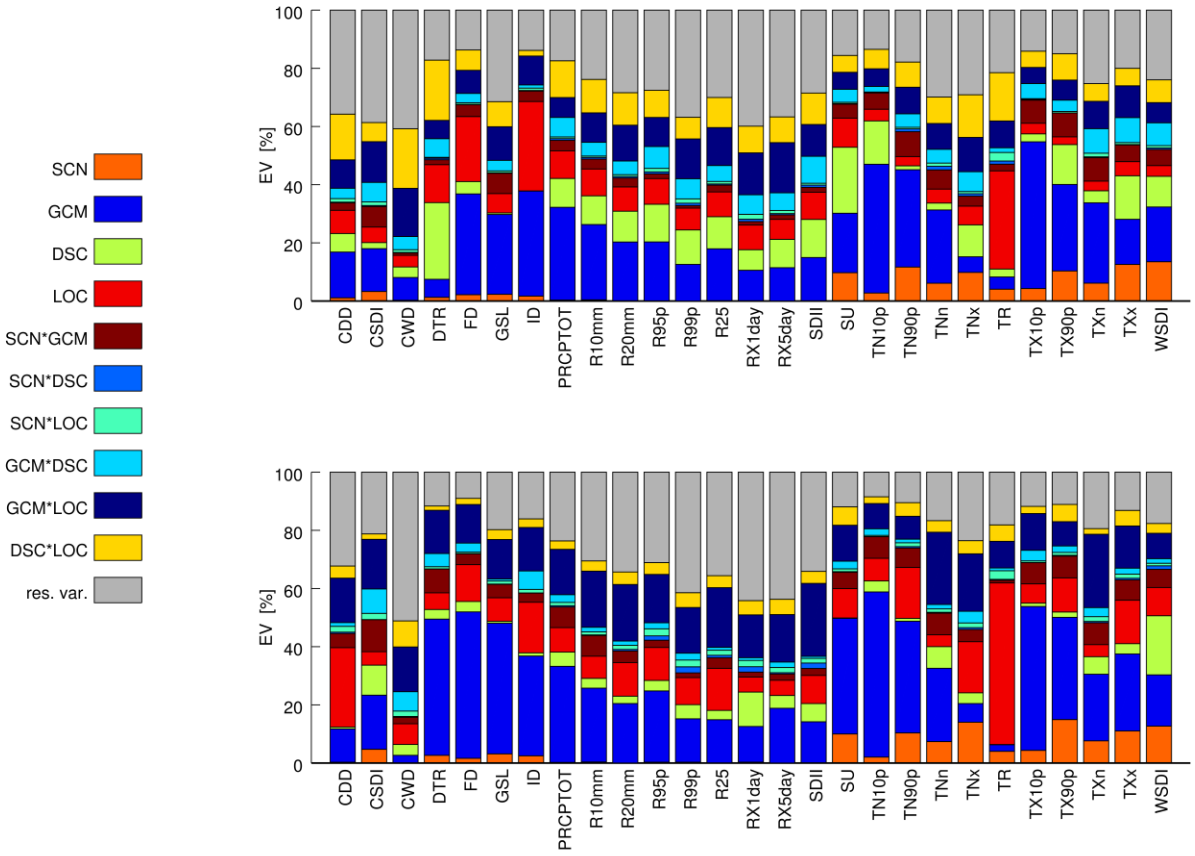

Figure 13. Same as Figure 9, but using only the verified downscaling methods BCSD, QRNN, and XDS (upper panel) or only CDFt, DQM, and TG which occupy the center of the MDS in Figure 11 (lower panel).

745