# Downscaling Extremes—An Intercomparison of Multiple Statistical Methods for Present Climate

G. BÜRGER, T. Q. MURDOCK, A. T. WERNER, AND S. R. SOBIE

*Pacific Climate Impacts Consortium, University of Victoria, Victoria, British Columbia, Canada*

A. J. CANNON

*Environment Canada, Vancouver, British Columbia, Canada*

(Manuscript received 27 July 2011, in final form 14 December 2011)

ABSTRACT

Five statistical downscaling methods [automated regression-based statistical downscaling (ASD), bias correction spatial disaggregation (BCSD), quantile regression neural networks (QRNN), TreeGen (TG), and expanded downscaling (XDS)] are compared with respect to representing climatic extremes. The tests are conducted at six stations from the coastal, mountainous, and taiga region of British Columbia, Canada, whose climatic extremes are measured using the 27 Climate Indices of Extremes (ClimDEX; http://www.climdex.org/climdex/index.action) indices. All methods are calibrated from data prior to 1991, and tested against the two decades from 1991 to 2010. A three-step testing procedure is used to establish a given method as reliable for any given index. The first step analyzes the sensitivity of a method to actual index anomalies by correlating observed and NCEP-downscaled annual index values; then, whether the *distribution* of an index corresponds to observations is tested. Finally, this latter test is applied to a downscaled climate simulation. This gives a total of 486 single and 162 combined tests. The 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 combined tests. With respect to regions, there is some tendency of better performance at the coastal and mountaintop stations. With respect to methods, XDS performed best, on average, with 19% (48%) of passed combined (single) tests, followed by BCSD and QRNN with 10% (45%) and 10% (31%), respectively, ASD with 6% (23%), and TG with 4% (21%) of passed tests. Limitations of the testing approach and possible consequences for the downscaling of extremes in these regions are discussed.

## 1. Introduction

Interest in global warming is increasingly shifting from assessments of average behavior to understanding and analyzing the effects on extremes. Because of the very nature of extreme events being rare, corresponding statistical assessments are loaded with uncertainty. This can partly be overcome, at the expense of spatial and temporal detail, by emphasizing either global or continental scales and ensemble results of multiple climate models (Kharin and Zwiers 2000; Tebaldi et al. 2006; Kharin et al. 2007; Min et al. 2011). To regain the small-scale information and bridge the gap between the coarse scales of simulated climate and the local scales where climatic extremes usually materialize, a whole new discipline has evolved that is commonly referred to as ''downscaling.'' By employing physical or statistical methodology downscaling ''distills'' as much small-scale information out of global climate models (GCMs) as possible [see Wilby et al. (2004) for a comprehensive overview].

By focusing on local extremes, therefore, one is confronted with both the shrinkage of sample size and the weakening of theoretical linkage to increasing greenhouse gases, as compared to, for example, global climate models. Nevertheless, a whole body of studies has emerged that tackle the impact of global warming on local extremes using some form of downscaling (Schubert and Henderson-Sellers 1997; Olsson et al. 2001; Harpham and Wilby 2005; Dibike and Coulibaly 2006; Fowler et al. 2007; Vrac and Naveau 2007; Busuioc et al. 2008; Benestad 2010;

*Corresponding author address:* G. Bürger, P.O. Box 3060 Stn CSC, University House 1, Pacific Climate Impacts Consortium (PCIC), University of Victoria, Victoria BC V8W 3R4, Canada.
E-mail: gbuerger@uvic.ca

DOI: 10.1175/JCLI-D-11-00408.1

© 2012 American Meteorological Society

Mannshardt-Shamseldin et al. 2010); the European Statistical and Regional Dynamical Downscaling of Extremes for European Seasons (STARDEX; http://www.cru.uea.ac.uk/projects/stardex) was solely devoted to this topic. It is well known that downscaling comes in two ''flavors''—the dynamical, using regional climate models (RCMs), and the statistical, using empirical statistical techniques. It is also well known that both have their advantages and disadvantages, and often they are not even easy to compare because of their different target applications.

Because we are focusing on local extremes this study solely deals with the statistical approach (like STARDEX). Compared to its dynamical sister, statistical downscaling is more heterogeneous and often consists of a patchwork of different methods, recipes, and adjustments that would be unacceptable in a dynamical context. Bias correction techniques, as one example, are an integral part of many empirical methods (including those in this study), whereas they would be considered illegitimate as part of dynamical downscaling. The major problem here is that they usually represent fairly heterogeneous pre- and postprocessing procedures external to a core statistical model, which operate on long-term parameters of the simulated GCM climate. While the core method itself can be estimated and verified using standard statistical methodology based on short-term weather observations, verification of the external procedures requires an entire array of independent climates, reflected only in observational series spanning multiple decades. Methodologically, such adjustments are related to the flux correction schemes of earlier coupled GCMs (Sausen et al. 1988); because each are calibrated against observed datasets the coupling of atmosphere and ocean models to one another often resulted in a long-term drift that posed considerable problems for the interpretation of climate scenarios.

Consequently, with little or no foundation in physical principles all of these empirical tools need thorough verification, especially when it comes to extremes. In the words of Kundzewicz and Stakhiv (2010, p. 1087), bias correction techniques ''merely represent an ad hoc curve-fitting exercise of convenience, rather than a result of impeccable physically-based theory.'' Compared to standard statistical methods, which possess a known uncertainty, there is always an element of unknown uncertainty (cf. Department of Defense 2002) with regard to any of these nonstandard methods. On the other hand, most of these techniques operate just as a correction, so their effect is second order, unless the correction is a major one.

Given the multitude of techniques and results, downscaling intercomparison projects provide guidance for choosing the best method for a purpose in question. Ideally, there should be just one big intercomparison including all possible methods, with well-defined rules and performance measures, preferably stratified according to region. The climate Coupled Model Intercomparison Project (CMIP; Meehl et al. 2000) could certainly serve as a role model here. A very comprehensive intercomparison was undertaken by Schmidli et al. (2007), who analyze and compare the dynamically and statistically downscaled climate simulations at the RCM scale for the European Alps, using a number of statistics for daily precipitation. One of their main results is that all of the statistical methods underestimate the present interannual variability. For three watersheds in the western United States Hay and Clark (2003) compare the daily runoff simulated from downscaled atmospheric fields; a regression-based method produced the best results both with respect to the annual cycle as well as daily variations. Using a subset of the STARDEX indices, Haylock et al. (2006) compared the performance of several more advanced methods to downscale heavy precipitation over England, and obtained the best results from neural nets [which were not considered by Hay and Clark (2003)].

In this study, we test a broad range of temperature- and precipitation-related extremes as measured by the set of 27 core indices, the Climate Indices of Extremes (ClimDEX). The ClimDEX indices (http://www.climdex.org), which are listed in Table 1, do not generally reflect the most extreme events conceivable, but instead represent ''the more extreme aspects of climate,'' which are (i) known to be relevant to a broad range of impact fields (Peterson 2005), and (ii) still statistically manageable so that they can be reliably estimated from current data for the 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 for the next Intergovernmental Panel on Climate Change (IPCC) report (Zhang et al. 2011). We apply five advanced downscaling methods to the highly varying climate zones of British Columbia, Canada, which range from coastal to mountainous to subarctic climate. The downscaling methods are as follows: automated regression-based statistical downscaling (ASD; based on Wilby et al. 2002), bias correction spatial disaggregation (BCSD; Wood et al. 2002; Salathe et al. 2007), quantile regression neural networks (QRNN; Cannon 2011), TreeGen (TG; Stahl et al. 2008), and expanded downscaling (XDS; Bürger 1996); details of the methods are described in section 2. The testing is threefold:

1) measure the sensitivity to actual (annual) climate anomalies;
2) test for the representation of present climate from reanalyses; and

TABLE 1. ClimDEX indices.

| ID | Indicator name | Definitions | Units |
|---|---|---|---|
| CDD | Consecutive dry days | Maximum number of consecutive days with RR < 1 mm | days |
| CSDI | Cold spell duration | Days with at least six consecutive days when TN < 10th percentile | days |
| CWD | Consecutive wet days | Maximum number of consecutive days with RR ≥ 1 mm | days |
| DTR | Diurnal T range | Monthly mean difference between TX and TN | °C |
| FD0 | Frost days | Annual count when TN (daily minimum) < 0°C | days |
| GSL | Growing season length | Days between first and last span of at least six warm enough days | days |
| ID0 | Ice days | Annual count when TX (daily maximum) < 0°C | days |
| PRCPTOT | Annual total wet-day precipitation | Annual total PRCP in wet days (RR ≥ 1 mm) | mm |
| R10 | Number of heavy precipitation days | Annual count of days when PRCP >= 10 mm | days |
| R20 | Number of very heavy precipitation days | Annual count of days when PRCP ≥ 20 mm | days |
| R95p | Very wet days | Annual total PRCP when RR > 95th percentile | mm |
| R99p | Extremely wet days | Annual total PRCP when RR > 99th percentile | mm |
| R25 | Number of days above 25 mm | Days when PRCP > 25 mm | Days |
| RX1day | Max 1-day precipitation | Monthly maximum 1-day precipitation | mm |
| Rx5day | Max 5-day precipitation amount | Monthly maximum consecutive 5-day precipitation | mm |
| SDII | Simple daily intensity index | Annual total precipitation divided by the number of wet days (PRCP ≥ 1 mm) | mm day$^{-1}$ |
| SU25 | Summer days | Annual count when TX (daily maximum) > 25°C | days |
| TN10p | Cool nights | Percentage of days when TN < 10th percentile | % |
| TN90p | Warm nights | Percentage of days when TN > 90th percentile | % |
| TNn | Min TN | Monthly minimum value of daily minimum temp | °C |
| TNx | Max TN | Monthly maximum value of daily minimum temp | °C |
| TR | Tropical nights | Annual count when TN (daily minimum) > 20°C | days |
| TX10p | Cool days | Percentage of days when TX < 10th percentile | % |
| TX90p | Warm days | Percentage of days when TX > 90th percentile | % |
| TXn | Min TX | Monthly minimum value of daily maximum temp | °C |
| TXx | Max TX | Monthly maximum value of daily maximum temp | °C |
| WSDI | Warm spell duration | Days with at least six consecutive days when TX > 90th percentile | days |

3) test for the representation of present climate from GCMs.

Unlike the European comparison studies above (Haylock et al. 2006; Schmidli et al. 2007), we provide well-defined criteria for the passing of each test, resulting in a unique set of successful method–index pairs. To our knowledge no previous study has conducted such a rigorous testing procedure. With some adjustments (mainly of the defining thresholds) the same procedure can be applied to RCM and even GCM fields to obtain entire maps of ClimDEX indices (Sillmann and Roeckner 2007). An intercomparison across statistical and dynamical techniques requires a transfer of scales, though, and this is a task that is largely nontrivial and beyond the scope of our study.

The tests will be done using two fully independent decades of daily data, details of which are outlined in section 3. For each of three focus regions of British Columbia (section 4) we analyze the results separately (section 5) and discuss the implications for the applications of future scenarios and the direction of further work and improvements (section 6). Note that test 1 (sensitivity to actual anomalies) is not suitable (and will fail) for methods that represent present climate in a purely stochastic form, such as weather generators. Accordingly, such approaches will not be considered here.

The focus of this study lies on verification relative to the present climate. A subsequent study is planned that considers the implications for future climate.

## 2. Statistical downscaling

As outlined in the introduction, statistical downscaling is a very heterogeneous enterprise drawing upon a large variety of sources, ranging from sophisticated statistical methods, such as stepwise regression or neural nets, to rather practical numerical recipes, such as the delta method (see below) or variance adjustment. Ideally, one should try to isolate the single tools and test their usefulness independently. Because of coupling effects, however, the combined performance, which is what counts in applications, is quite unpredictable from the single performance. For an intercomparison study such as this one, it is more feasible, therefore, to assess the combined effect of the various tools as they appear in a full downscaling application. Moreover, and most importantly, we wanted to test the methods as they are being used, to better rate the performance of past applications.

Common to most empirical downscaling methods is a training or calibration phase of a transfer function, where the adjustable function parameters are estimated from the observed atmosphere (reanalysis) and local station data in a period of overlapping data. That transfer function could then be applied to the same or similar atmospheric data to obtain corresponding downscaled local data and then, for example, compared to observations (verification). It could in principle be applied to GCM-simulated fields as well. However, systematic GCM biases (relative to reanalyses) interfere, which occasionally even exceed a projected climate change signal, with potentially dramatic effects on subsequent impact applications, such as a hydrologic model. Therefore, bias correction is an integral part of most downscaling methods (Dehn and Buma 1999; Easterling 1999; von Storch 1999; Chen 2002; Wood et al. 2002; Kysely 2002; Hay and Clark 2003; Huth 2004; Fowler et al. 2007; Maurer and Hidalgo 2008; Maraun et al. 2010).

When we set up our testing experiments, our goal was to use each method as it was used before in published studies by other research groups, or as it would be used by a novice user in publicly available, user-friendly methods. In the process, a few smaller errors were found and corrected before being applied here.

Of the five methods tested in this study at least four are fairly advanced and not easily available to the average user, so our choice is in part dictated by opportunity. However, they span a wide range of different approaches, which meets the objectives of this work.

### *a. Methods*

All of our methods, listed below, simulate variations about the long-term climatological mean. One such variation is the seasonal cycle, and it is important whether that cycle is treated internally, as imposed by seasonally varying predictors, or externally, as imposed deterministically through fitted harmonics. For ASD and BCSD the former (internal) applies and for QRNN, TG, and XDS the latter (external).

#### 1) Automated regression-based statistical downscaling

Driven by the need to make easy-to-use downscaling tools available to a larger community, ASD (Hessami et al. 2008) was developed as an automated version of the statistical downscaling model (SDSM) of Wilby et al. (1999, 2002). These sources describe SDSM as a ''hybrid of the stochastic weather generator and transfer function methods.'' The method first links large-scale circulation patterns and atmospheric moisture variables with local-scale weather parameters (precipitation occurrence and amount, minimum and maximum temperature), using an (auto)regression approach, and then adjusts simulated local variables to account for loss of variance and residual bias (the stochastic weather generator component).

The automatization is designed to replace SDSM's subjective method of predictor selection, requiring significant input on the part of the user, with a more objective approach. For example, it uses a backward stepwise multiple linear regression (e.g., Seber and Lee 2003), starting from a fairly large suite of standard potential predictors (cf. Hessami et al. 2008). Loss of variance, a characteristic of any regression method, is accounted for by adding a white noise process, the variance of which, along with some bias adjustments, can be set either manually or automatically. Our version of ASD, along with most others, does not apply any extra measures to deal with predictor collinearities; this was implemented only in recent versions of ASD. Its hybrid nature renders each ASD simulation partly stochastic so that extra care is required for the validation (see section 3).

Additional studies that make use of ASD or SDSM include Wilby et al. (2003, 2006), Wilby (2005), Khan et al. (2006), and Gachon and Dibike (2007).

#### 2) Bias-corrected spatial disaggregation

BCSD (Wood et al. 2002) is geared toward providing gridded, high-resolution temperature and precipitation fields over relatively large domains, mainly for driving hydrologic models. GCM data are bias corrected, spatially disaggregated (to a finer grid or station data) and finally temporally disaggregated to a daily time step. To avoid the extra handling of wet and dry days the bias correction part of BCSD is applied to monthly GCM data only. The present and most widely used version of BCSD uses mean temperature as a driver; minimum and maximum temperature are derived indirectly using climatological temperature range. The main steps of BCSD are as follows:

1) Bias correction of monthly GCM fields, using aggregated observations. A quantile mapping (Panofsky and Brier 1958) is employed to adjust the monthly large-scale fields. The adjustment is done on the detrended data, which are retrended afterward to have the same climatic trends as the original large-scale

fields. To formalize quantile mapping, we denote the modeled series and modeled and observed cumulative distribution functions by $M$, $F_M$, and $F_O$, respectively. Let $\varphi$: [0 1] → [0 1] denote the identity map of the unit interval. The final mapping of the modeled series is done as follows:

$$\begin{tikzcd} M \arrow[r, Rightarrow, "f"] \arrow[d, "F_M"'] & M_{\mathrm{qm}} \\ {[0,1]} \arrow[r, "\vec{\varphi}"] & {[0,1]} \arrow[u, "F_O^{-1}"'] \end{tikzcd}, \tag{1}$$

which is formally $f = F_M\ \varphi F_O^{-1}$: $M \rightarrow M_{\mathrm{qm}}$. The mapping must be estimated using data from the calibration period. Data outside the range of the calibration percentiles are extrapolated using Weibull (Gumbel) fits to $F$ for precipitation minimum (maximum) and a Gaussian fit for temperature. Note that these fits strongly influence the characteristics of the (monthly) extremes.

2) Spatial and temporal disaggregation using a delta approach. The monthly large-scale values of precipitation ($P$) and temperature ($T$) are first disaggregated spatially, using high-resolution correction factors (for $P$) and summands (for $T$). These values are finally temporally disaggregated by picking a random historic month and adjusting its daily values (multiplicatively and additively) to reproduce the monthly value.

Additional studies that make use of BCSD include Christensen et al. (2004), VanRheenen et al. (2004), Maurer and Duffy (2005), Christensen and Lettenmaier (2007), Hayhoe et al. (2007), Maurer et al. (2007), and Schnorbus et al. (2011).

### 3) Quantile regression neural network

QRNN (Taylor 2000) estimates conditional values of an individual quantile using a multilayer perceptron neural network. If one develops QRNN methods for a range of quantiles $q = 0.1, \ldots, 0.9$, then the result is an estimate of the full predictand distribution, without extra assumptions about the parametric form of the distribution. In a downscaling context, this means that the shape of the distribution may change under future climate conditions as the large-scale GCM predictors change. Details are given in Cannon (2011). QRNN is methodologically similar to the multilayer perceptron neural network MLPR used by Haylock et al. (2006).

Unlike XDS (see below), QRNN predictors need not be normalized because of the nonlinearity of the underlying model. However, when making predictions for GCM scenarios, it is assumed that the statistical characteristics of the GCM predictors match those from the reanalysis over the calibration period. If this is not the case, then a bias correction step, as in BCSD, must be applied. Here, a simple linear rescaling is used to match climatological means and variances.

Because QRNN downscaling models are probabilistic, results in this study are based on 20 simulations from the estimated conditional distributions. Specifically, the conditional quantile function on a given day is obtained by linearly interpolating between predicted quantiles from the set of fitted QRNN models. Outside the range of the fitted quantile probabilities, exponential lower/upper tails are assumed following Quinonero-Candela et al. (2006). Simulated values are obtained by entering a uniform random variate into the conditional quantile function.

Because of its novelty QRNN has not been used apart from (Cannon 2011). Its source code (R) can be obtained online (http://cran.r-project.org/web/packages/qrnn).

### 4) TreeGen

TG is a hybrid downscaling technique that draws upon several approaches to statistical downscaling, including synoptic weather typing, regression modeling, analog resampling, and stochastic weather generation. TG is driven by daily atmospheric fields. The following steps are carried out independently for each station:

1) Predictor selection is based on common principal components (PCs) of National Centers for Environmental Prediction (NCEP) and the GCM fields (see section 2b).
2) Synoptic types are determined using a multivariate regression tree that recursively splits observed data into increasingly homogeneous groups on the basis of thresholds in the PC scores (Cannon et al. 2002). Values of the thresholds are optimized so that the associated surface temperature and precipitation observations are placed into groups (or weather map types) that minimize within-group sums of squares error. Following (Cannon et al. 2002), 25 map types are identified for the study domain.
3) Once thresholds have been identified using the historical record predictor PCs are then entered into the regression tree, defining a unique map type for each simulation day.
4) For each simulation day an observation from the predicted weather type is picked stochastically, with a probability that is inversely proportional to the Euclidean distance between its predecessor and the previous simulated day.

5) Because simulated low-frequency variability in surface climate conditions is exclusively due to changes in the frequency and timing of the synoptic map types, a low-frequency correction is applied to each weather type separately, as follows: (i) interannual variability at the GCM grid point is superimposed onto the nonparametric weather generator outputs; and (ii) trend corrections are applied additively for temperature and multiplicatively for precipitation to match the trends from the nearest GCM grid point.

The stochastic component of TG requires (as for ASD) extra consideration for the test setup described in section 3.

Studies that make use of TG include Stahl et al. (2008) and Allen et al. (2010).

### 5) Expanded downscaling

XDS is born out of the idea of simulating local events that are as close to and consistent with the prevailing atmospheric circulation, but at the same time generate realistic local covariability (of variables and stations). Let predictors and predictands be denoted by $\mathbf{x}(t)$ and $\mathbf{y}(t)$, respectively. Whereas in the classical linear regression approach one seeks a matrix $\mathbf{Q}$ that minimizes the error $\mathbf{xQ} - \mathbf{y}$, this unconditional error criterion is relaxed in favor of searching only through those $\mathbf{Q}$ that preserve local covariability. This leads to the following constraint optimization problem,

$$\mathrm{XDS} = \arg\min_{Q}(\|\mathbf{xQ} - \mathbf{y}\|), \quad \text{subj. to} \quad \mathbf{Q}'\mathbf{x}'\mathbf{xQ} = \mathbf{y}'\mathbf{y}, \tag{2}$$

which simply means that the matrix $\mathbf{Q}$ that minimizes the error $\mathbf{xQ} - \mathbf{y}$ is sought among those that preserve local covariance $(\mathbf{xQ})'\mathbf{xQ} = \mathbf{Q}'\mathbf{x}'\mathbf{xQ} = \mathbf{y}'\mathbf{y}$. Equation (2) has as a unique solution, XDS, which we call the expanded downscaling or XDS model. As an optimization problem the solution is only approximate and in previous applications its estimation required large computing resources. Only recently it was found (Bürger et al. 2009) that Eq. (2) can be recast as an orthogonal Procrustes problem from statistical shape analysis (Dryden and Mardia 1998). Denoting the Cholesky factors of $\mathbf{x}'\mathbf{x}$ and $\mathbf{y}'\mathbf{y}$ as $\mathbf{G}_\mathbf{x}$ and $\mathbf{G}_\mathbf{y}$, respectively, the solution has the form

$$\begin{aligned} \mathbf{U}\Lambda\mathbf{V} &= \mathbf{G}_\mathbf{y}\mathbf{y}'\mathbf{x}\mathbf{G}_\mathbf{x}^{-1} \qquad \text{(SVD)} \\ \mathrm{XDS} &= \mathbf{G}_\mathbf{x}^{-1}\mathbf{V}\mathbf{U}'\mathbf{G}_\mathbf{y}. \end{aligned} \tag{3}$$

Because the main building blocks of Eq. (3) are derived from (cross-)covariances and only normal (Gaussian) distributions are fully characterized by the first and second moments, for the estimation of XDS the variables in question ($\mathbf{x}$ and $\mathbf{y}$) should be transformed into the normalized domain. This can be achieved for any random variable $X$ using the probability integral transform (probit), as follows: the image of $X$ under the cumulative distribution function is uniformly distributed, and applying to this image the inverse normal distribution (Gaussian quantile) function results in a normally distributed variable. The statistics of extremes are therefore mainly defined through the probit parameters. The preservation of local covariance, including realistic variability and interstation correlations, renders XDS particularly useful for hydrologic applications where it has been applied most frequently (Dehn et al. 2000; Müller-Wohlfeil et al. 2000; Menzel and Bürger, 2002; Bronstert 2004; Menzel et al. 2006; Bürger et al. 2009; see also (Wilby et al. 2004). Experimental sources (Octave/Matlab) can be obtained online (http://xds.googlecode.com).

## *b. Predictors*

Each method usually comes with its own set of atmospheric predictors. For example, some methods use upper-level atmospheric information and others only surface data. The matrix of (potential) predictors is shown in Table 2; note that QRNN, TG, and XDS predictors are synoptic fields.

ASD: We selected from each level an optimum grid point (based on correlation) and defined the average of its neighbors as a potential predictor. The stepwise regression identified an average of six predictors per variable.

BCSD: We take for each station the nearest grid point of the corresponding variable, that is, one predictor per predictand.

QRNN: The domain between (35°N, 215°E) and (65°N, 255°E) is used. Predictors were screened separately for each station and predictand variable using the regression tree methodology described by Cannon (2008). Seasonal anomalies of gridpoint predictors were entered as inputs to a regression tree targeted on the station predictand. Only those variables that formed splits in the tree were used as predictors. Sine and cosine of the day of year were added as additional predictors to account for seasonal variations in the predictor–predictand relationships.

TG: It uses the dominant common PCs (95%) of the QRNN predictors, based on pooling NCEP and the historic GCM covariances. This leads to about 10–20 predictors per predictand.

TABLE 2. The matrix of (potential) predictors. Boldface indicates predictor fields.

| | ASD | BCSD | QRNN | TG | XDS |
|---|---|---|---|---|---|
| 500 hPa | $T, q, u, v, d, \zeta$ | | $z$ | $z$ | |
| 700 hPa | | | | | $\boldsymbol{T, q, u, v}$ |
| 850 hPa | $T, q, u, v, d, \zeta$ | | **RH**, $\boldsymbol{T, q, u, v, z}$ | **RH**, $\boldsymbol{T, q, u, v, z}$ | $\boldsymbol{T, q, u, v}$ |
| Surface | $T, q, u, v, d, \zeta$ | $T, P$ | $\boldsymbol{T, P}$, **SLP** | $\boldsymbol{T, P}$, **SLP** | $\boldsymbol{P}$ |

XDS: We used a rectangle of about 14° longitude and 10° latitude about the center of each region in a 2.5° resolution and projected each field onto the dominant PCs, retaining 99% of the variance. The optimum number of retained PCs was estimated using a split sample method, by training XDS on 1961–75 data and validating on 1976–90. This leads to about 50 predictors for each region.

The testing would certainly have benefited from the use of a common predictor set. However, the employment of specific forcing fields is so interwoven with each method that we have taken them as part of the method itself. For example, a crucial component of the ASD is the stepwise regression, starting from a rather large number of potential predictors. If that component is removed then the method is no longer automatic and would no longer deserve the "A" in ASD. Likewise, much of the simplicity of BCSD derives from its use of a sparse predictor set (temperature and precipitation).

## 3. Test setup

The testing consists of comparing observed and downscaled indices of ClimDEX from the independent period from 1991 to 2010. It is made up of three single tests, all of which are crucial to derive future information from present climate in a reliable way. We require that a method

(test 1) adequately responds to actual anomalies in the reanalyses,

(test 2) reproduces the present distribution of an index from reanalyses,

(test 3) reproduces the present distribution of an index from the GCM.

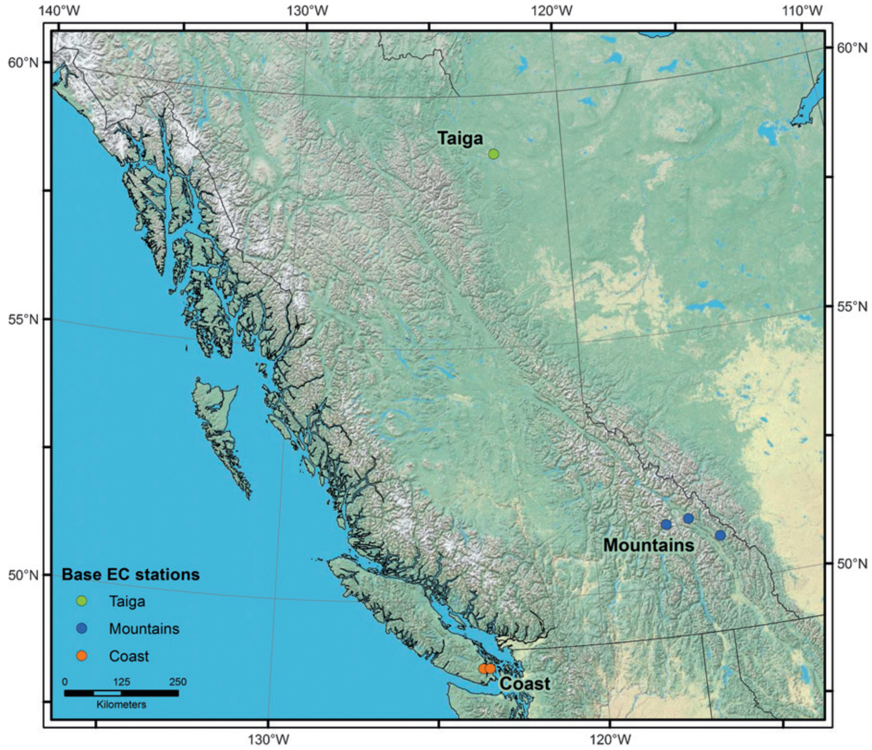

FIG. 1. The study area with stations.

TABLE 3. Regions and stations used for the comparison, along with characteristic temperature ($\overline{T}$) and precipitation ($\overline{P}$). Parenthesized symbols are used in summary tables.

| Region | Station | ID | Lon (°) | Lat (°) | Alt (m) | $\overline{T}$ (°C) | $\overline{P}$ (mm day$^{-1}$) |
|---|---|---|---|---|---|---|---|
| Mountains | | | | | | | |
| | Kootenay ($M_1$) | 1154400 | −116.05 | 50.88 | 1170 | 2.3 | 511.2 |
| | Golden ($M_2$) | 1173210 | −116.98 | 51.30 | 785 | 4.7 | 475.2 |
| | Mt. Fidelity ($M_3$) | 117CA90 | −117.70 | 51.23 | 1875 | 0.2 | 2101.6 |
| Taiga | | | | | | | |
| | Fort Nelson ($T_1$) | 1192940 | −122.60 | 58.83 | 382 | −0.7 | 451.7 |
| Coast | | | | | | | |
| | Shawnigan ($C_1$) | 1017230 | −123.63 | 48.65 | 138 | 9.6 | 1247.6 |
| | Victoria ($C_2$) | 1018620 | −123.43 | 48.65 | 19 | 9.7 | 883.3 |

Accordingly, for any given pair of method and index all tests should be passed in order to be deemed reliable. The passing of test 3 likely depends on the GCM in question, so one may want to assign some average measure to a method across many GCMs. For the present study only one GCM was available that satisfied all of the necessary data requirements to apply all of the methods (see below).

Test 1 checks whether a method is sensitive to climatic anomalies. For example, if one year shows an exceptionally high value of annual total wet-day precipitation (PRCPTOT; see Table 1) one would attribute this to some anomalous large-scale circulation pattern. If this incident is considered to be more likely in a changing climate, then its occurrence or nonoccurrence ought to be reflected by the downscaling. The test consists of checking whether a simulated index (i.e., one derived from simulations) has significantly nonzero correlations to the observed index, in which case the method/index *passes the test*. We base this test on annual values as a compromise between what is considered to be a relevant time scale of climate change and the size of

TABLE 4. Summary table for test 1; see Table 3 for symbols.

| Index | ASD | BCSD | QRNN | TG | XDS |
|---|---|---|---|---|---|
| CDD | | | | | |
| CSDI | $M_3$ | | $C_1M_1M_2M_3$ | $C_1C_2M_1M_2$ | $M_1M_2M_3$ |
| CWD | | | | | $C_1$ |
| DTR | | | $C_1$ | | $C_1M_1$ |
| FD | $C_1C_2M_2M_3T_1$ | $C_1C_2M_3T_1$ | $C_1C_2M_3T_1$ | $M_1M_3T_1$ | $C_1C_2M_2M_3T_1$ |
| GSL | $M_3$ | | $C_2M_1M_3T_1$ | | $M_1M_2M_3T_1$ |
| ID | $C_1C_2M_1M_2M_3$ | $M_1M_2$ | $C_1C_2M_1M_2M_3T_1$ | $C_1M_1M_2M_3T_1$ | $C_1M_1M_2M_3T_1$ |
| PRCPTOT | $C_2$ | $C_1C_2$ | $C_1C_2$ | $C_1C_2$ | $C_1C_2M_3T_1$ |
| R10 | | $C_1C_2$ | $C_2$ | $C_2$ | $C_2M_3$ |
| R20 | | $C_1$ | | | $C_1M_3T_1$ |
| R95p | | | | | $C_2M_3T_1$ |
| R99p | | | | | |
| R25 | | $C_1$ | | | $C_2M_3$ |
| RX1day | | | | | $C_1$ |
| RX5day | | $M_2$ | | | |
| SDII | | $C_1C_2M_2$ | $C_2$ | | $M_3T_1$ |
| SU | $C_1C_2M_1M_2M_3$ | $C_1C_2M_1M_2M_3T_1$ | $M_1M_2M_3T_1$ | $C_1M_1M_2$ | $C_1C_2M_1M_2M_3T_1$ |
| TN10p | $C_1C_2M_1M_2M_3T_1$ | $M_1M_2M_3T_1$ | $C_1C_2M_1M_2M_3T_1$ | $M_1M_2M_3T_1$ | $C_1C_2M_1M_2M_3T_1$ |
| TN90p | $C_1C_2M_2M_3$ | $C_1M_1M_2M_3$ | $C_1C_2M_1M_2M_3T_1$ | $M_2M_3$ | $C_1M_2M_3$ |
| TNn | $C_1$ | | $C_1M_1M_2M_3T_1$ | $C_1T_1$ | $C_1M_1M_3T_1$ |
| TNx | $C_1M_2$ | | | | $C_1$ |
| TR | | | | | |
| TX10p | $C_1C_2M_1M_2M_3T_1$ | $M_1M_2T_1$ | $C_1C_2M_1M_2M_3T_1$ | $M_1M_3T_1$ | $C_1C_2M_1M_2M_3$ |
| TX90p | $C_1C_2M_1M_2M_3$ | $M_1M_2M_3T_1$ | $C_1C_2M_1M_2M_3T_1$ | $M_1M_3$ | $C_1C_2M_1M_2M_3$ |
| TXn | $C_1$ | | $C_1C_2M_2M_3T_1$ | $M_2M_3$ | $C_1C_2M_2M_3T_1$ |
| TXx | $M_3$ | $M_3$ | $C_1M_3$ | | $C_1M_2M_3$ |
| WSDI | | $M_1$ | | | |
| Total No. | 44 | 39 | 63 | 33 | 75 |

TABLE 5. Summary table for test 2.

| Index | ASD | BCSD | QRNN | TG | XDS |
|---|---|---|---|---|---|
| CDD |  | $C_1C_2M_2$ |  |  | $T_1$ |
| CSDI |  |  |  |  |  |
| CWD |  | $C_2T_1$ |  |  | $M_1$ |
| DTR | $C_1$ |  |  | $M_3$ | $C_1C_2$ |
| FD |  | $C_2M_3T_1$ | $C_2M_3T_1$ | $T_1$ | $C_1C_2M_3T_1$ |
| GSL | $C_1$ | $C_2M_2M_3T_1$ | $C_2M_2M_3T_1$ |  | $C_1C_2M_2M_3T_1$ |
| ID | $M_1$ | $C_1M_2M_3T_1$ | $C_1M_2T_1$ | $M_2T_1$ | $M_1M_2M_3T_1$ |
| PRCPTOT | $T_1$ | $M_1T_1$ | $C_1$ |  | $C_1C_2M_1M_3T_1$ |
| R10 | $C_2T_1$ | $M_1M_3T_1$ | $C_1$ | $M_3$ | $C_1C_2T_1$ |
| R20 |  | $M_2$ | $C_1$ |  | $C_1M_2M_3$ |
| R95p |  | $C_1C_2M_1M_2T_1$ | $C_1C_2$ | $C_1C_2T_1$ | $C_1C_2M_2M_3$ |
| R99p |  | $C_1M_2M_3$ |  |  | $C_1C_2M_3$ |
| R25 |  | $T_1$ | $C_1$ | $C_1$ | $C_2$ |
| RX1day |  | $C_1M_1M_2T_1$ |  |  | $C_1C_2M_2$ |
| RX5day |  | $M_2T_1$ |  |  | $C_1C_2M_2$ |
| SDII | $T_1$ | $M_1M_2$ | $T_1$ | $C_2$ | $C_2$ |
| SU | $C_1$ | $C_1C_2M_2$ | $C_1$ | $C_1M_2$ | $C_1M_2$ |
| TN10p | $C_2M_2M_3T_1$ | $C_1C_2M_1M_2T_1$ | $C_1C_2M_2T_1$ | $C_1C_2M_1M_2T_1$ | $C_1C_2M_1M_2M_3T_1$ |
| TN90p | $C_2T_1$ | $C_1C_2M_2M_3T_1$ | $C_1C_2M_2M_3T_1$ | $C_1M_3T_1$ | $C_1C_2M_2M_3T_1$ |
| TNn | $T_1$ | $C_1C_2M_1M_2T_1$ | $C_1C_2T_1$ | $C_2$ | $C_1M_1M_2M_3T_1$ |
| TNx | $M_1$ | $C_1C_2M_1M_2$ | $C_1$ |  | $M_1$ |
| TR | $C_1C_2M_1M_3T_1$ | $C_1C_2M_1M_3T_1$ | $C_1C_2M_1M_3T_1$ | $C_1C_2M_1M_3T_1$ | $C_1C_2M_1M_3$ |
| TX10p | $M_2M_3T_1$ | $C_1C_2M_2M_3T_1$ | $C_2M_2M_3T_1$ | $C_1M_3T_1$ | $C_1C_2M_1M_2M_3T_1$ |
| TX90p | $C_1C_2M_2M_3$ | $C_1C_2M_2M_3T_1$ | $C_1C_2M_2M_3T_1$ | $C_1C_2M_2M_3T_1$ | $C_1C_2M_2M_3T_1$ |
| TXn | $C_1C_2$ | $C_1C_2M_1M_2M_3T_1$ | $C_1C_2M_3T_1$ |  | $C_1C_2M_2M_3T_1$ |
| TXx | $C_1$ | $C_1M_2T_1$ | $C_1C_2$ |  | $C_1C_2$ |
| WSDI |  |  |  |  |  |
| Total No. | 31 | 85 | 51 | 34 | 84 |

independent samples to test. Serial correlation is taken into account (Ebisuzaki 1997).

To compare the simulated *distribution* for an index, either from NCEP (test 2) or from a GCM (test 3), with the observed one we use the quantile–quantile (qq) plot and its corresponding sampling uncertainty. The mapping of any given quantile probability $q$, $0 \leq q \leq 1$, to the corresponding physical variable $x$ is given by $F^{-1}(x)$, where $F(x)$ is the empirical cumulative distribution function of $x$; the confidence limits in the qq plot are therefore given by the confidence band about $F$, which is derived from the Kolmogorov–Smirnov (KS) test. Using a significance level of $\alpha$, as specified below, this uncertainty is indicated in all of the following qq plots as a gray $(1 - \alpha)$ confidence band about the diagonal. According to the KS test, a simulated distribution is regarded as significantly different from observations if only one simulated quantile lies outside the confidence band; this is simply a consequence of the KS statistics, which is based on the maximum difference ($\| \|_\infty$ norm) between two cumulative distribution functions (CDFs). A method/index pair passes the test if the downscaled distribution is not significantly different from that observed. We have used a general significance level of $\alpha = 0.01$, leading to fairly wide confidence bands. The distribution test is applied to downscaled reanalyses as well as downscaled simulations of present climate driven by observed radiative forcings, that is, greenhouse gases, aerosols, solar, and volcanic ash [twentieth-century GCM simulations (20C3M; cf. http://www.ipcc-data.org/ar4/scenario-20C3M.html)].

The testing is complicated by the stochastic components of ASD, QRNN, and TG, where any two realizations can differ considerably. We need to test, therefore, whether the underlying simulation population differs significantly from observations when sampled accordingly. In the case of the distributions (test 2 and 3) this is straightforward: a method/index pair passes if the rate of rejected realizations is at most $\alpha$. For the sensitivity testing (test 1) the situation is more involved. While the sampling uncertainty for the distribution of a time series is estimated quite universally using the KS theory, no corresponding concept exists for the time series itself. We have used the following criterion for test 1: a method/index pair passes if 50% of the realizations pass the single test. This would be equivalent to requiring the median of all realized correlations to be significant, except if the corresponding significance level depends on the time series itself (which it does). For all stochastic methods

TABLE 6. Summary table for test 3.

| Index | ASD | BCSD | QRNN | TG | XDS |
|---|---|---|---|---|---|
| CDD | | $C_1C_2M_1M_2T_1$ | | | $C_1C_2M_1M_2M_3$ |
| CSDI | | | | | |
| CWD | | $C_2M_2$ | | | $M_1M_3T_1$ |
| DTR | $C_2$ | | | $C_1$ | $C_2$ |
| FD | | $T_1$ | $T_1$ | | $T_1$ |
| GSL | $C_1M_3$ | $C_1M_2T_1$ | | | $C_2M_3$ |
| ID | | $M_2T_1$ | | $T_1$ | $T_1$ |
| PRCPTOT | | $C_1C_2M_1T_1$ | $M_3$ | $C_2$ | $C_1C_2T_1$ |
| R10 | | $C_1C_2T_1$ | | | $C_1C_2M_3$ |
| R20 | | $C_1C_2M_2T_1$ | $C_2$ | | $C_1C_2$ |
| R95p | $C_2$ | $C_1C_2M_1M_2T_1$ | $C_1C_2$ | $C_2$ | $C_1C_2M_2T_1$ |
| R99p | | $C_1M_3$ | | | $C_2M_3$ |
| R25 | | $C_1C_2T_1$ | | $C_2$ | $C_1C_2$ |
| RX1day | | $C_1M_2T_1$ | | | $C_1C_2M_2T_1$ |
| RX5day | $C_2M_2$ | $C_1M_1M_2$ | | $C_1$ | $C_1C_2T_1$ |
| SDII | | $C_1M_2T_1$ | | | $C_1C_2$ |
| SU | $C_1C_2T_1$ | $C_1C_2M_2T_1$ | $T_1$ | $C_1C_2T_1$ | $C_1C_2M_2$ |
| TN10p | $C_2M_2T_1$ | $C_1C_2M_1M_2T_1$ | $C_2M_2T_1$ | $C_2M_2T_1$ | $C_1C_2M_1M_2T_1$ |
| TN90p | $C_1C_2M_2M_3T_1$ | $C_1C_2M_3T_1$ | $M_3T_1$ | $C_2M_3T_1$ | $C_1C_2M_2M_3T_1$ |
| TNn | $C_1M_3$ | $M_1$ | | | $C_1M_1M_2M_3$ |
| TNx | $C_1C_2M_2$ | $C_1C_2M_1M_2M_3$ | $C_1$ | | $C_1$ |
| TR | $C_1C_2M_1M_3T_1$ | $C_1C_2M_1M_3T_1$ | $C_1C_2M_1M_3T_1$ | $C_1C_2M_1M_3T_1$ | $C_1C_2M_1M_3T_1$ |
| TX10p | $M_1$ | $C_1C_2M_1M_2M_3$ | $M_1M_2T_1$ | $M_2$ | $C_2M_1M_2T_1$ |
| TX90p | $C_1C_2M_2M_3T_1$ | $C_1C_2M_2M_3T_1$ | $C_1C_2M_2M_3T_1$ | $C_1C_2M_2M_3T_1$ | $C_2M_3T_1$ |
| TXn | | $C_1C_2M_1$ | | | $M_1M_2M_3$ |
| TXx | $C_1$ | $C_1C_2M_2T_1$ | | $C_1$ | $M_2T_1$ |
| WSDI | | | | | |
| Total No. | 34 | 84 | 25 | 27 | 73 |

(ASD, QRNN, TG) we have used 20 realizations. Note that BCSD also contains a stochastic component in the daily disaggregation scheme; but because that is largely ''tamed'' by constraining it to prescribed monthly means, we treat BCSD as a deterministic method.

Estimating a distribution, as in tests 2 and 3, follows standard statistical practice of data sampling and calculating summary measures, and thus comes with a natural uncertainty as embodied by the Kolmogorov–Smirnov test. Accordingly, the notion of a 99% confidence interval has a solid statistical foundation, and in our study it provides the criterion for the passing of tests 2 and 3. For a time series there is no equivalent. For example, observational records do not come naturally with a confidence interval, and no corresponding test exists for the similarity to that record. As a workaround, summary measures for the similarity of any *two* series are usually employed, such as correlation or coherence. The dependence on a second argument, however, greatly confounds the significance assessments of these measures. All one can do is estimate their level of significance ''from below,'' by calculating them from random time series (''noise''). The passing of test 1, accordingly, involves the rejection of the null hypothesis of noise. Note that the character of this noise (color or memory) is often quite controversial.

## 4. Focus regions and data

As evident from Fig. 1, the study area of British Columbia offers a varied landscape where several climate zones exist. To reflect a broad range of these zones in the testing, we selected the

- coastal zone (C) with two stations, where climate is maritime and mild, and the seasonality is relatively weak for temperature, but strong for precipitation;
- southern interior mountains (M) with three stations, representing an alpine climate with strong topographic gradients for temperature and precipitation; and
- taiga plains (T) with one station, which is characterized with long sub-Arctic winters and cloudy and unstable weather in summer.

For each of these regions we selected a maximum of three climate stations that

- record daily values of precipitation $P$ and minimum and maximum temperature $T_n$ (=TN) and $T_x$ (=TX), respectively,

TABLE 7. Summary table for test 1, test 2, and test 3 combined.

| Index | ASD | BCSD | QRNN | TG | XDS |
|---|---|---|---|---|---|
| CDD | | | | | |
| CSDI | | | | | |
| CWD | | | | | |
| DTR | | | | | |
| FD | | $T_1$ | $T_1$ | | $T_1$ |
| GSL | | | | | $M_3$ |
| ID | | $M_2$ | | $T_1$ | $T_1$ |
| PRCPTOT | | | | | $C_1C_2T_1$ |
| R10 | | | | | $C_2$ |
| R20 | | | | | $C_1$ |
| R95p | | | | | $C_2$ |
| R99p | | | | | |
| R25 | | | | | $C_2$ |
| RX1day | | | | | $C_1$ |
| RX5day | | $M_2$ | | | |
| SDII | | $M_2$ | | | |
| SU | $C_1$ | $C_1C_2M_2$ | | $C_1$ | $C_1M_2$ |
| TN10p | $C_2M_2T_1$ | $M_1M_2T_1$ | $C_2M_2T_1$ | $M_2T_1$ | $C_1C_2M_1M_2T_1$ |
| TN90p | $C_2$ | $C_1M_3$ | $M_3T_1$ | $M_3$ | $C_1M_2M_3$ |
| TNn | | | | | $C_1M_1M_3$ |
| TNx | | | | | |
| TR | | | | | |
| TX10p | | $M_2$ | $M_2T_1$ | | $C_2M_1M_2$ |
| TX90p | $C_1C_2M_2M_3$ | $M_2M_3T_1$ | $C_1C_2M_2M_3T_1$ | $M_3$ | $C_2M_3$ |
| TXn | | | | | $M_2M_3$ |
| TXx | | | | | |
| WSDI | | | | | |
| Total No. | 9 | 16 | 13 | 6 | 31 |

- have a fairly complete data coverage for 1971–2010, and
- represent the altitudinal profile of the region.

The selected regions and stations are tabulated in Table 3.

We reserved the two decades from 1991 to 2010 for validation; this leaves sufficient daily data (depending on station, but at least two decades) for the calibration. It should be noted, however, that a sample size of 20 annual values from the validation is bound to limit the statistical power of each of the tests. From among several possible choices of reanalysis products, such as NCEP–National Center for Atmospheric Research (NCAR) Global Reanalysis 1 (GR-1) and NCEP/Department of Energy Global Reanalysis 2 (GR-2), 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40), or ECMWF Interim Reanalysis (ERA-Interim; cf. http://reanalyses.org/atmosphere/overview-current-reanalyses), GR-1 is the only one with sufficient pre-1991 data to calibrate the methods. As GCM data we used the 20C3M simulation (for pre-2001 values) followed by the A2 scenario of the ECHAM5/Max Planck Institute Ocean Model (MPI-OM) (EH5OM) run 1, as described online (http://dx.doi.org/10.1594/WDCC/EH5-T63L31_OM-GR1.5L40_20C_1_6H and http://dx.doi.org/10.1594/WDCC/EH5-T63L31_OM-GR1.5L40_A2_1_6H). At the time of writing, no other GCM was available to us that had sufficient predictor data for all methods. The original resolution is 2.5° × 2.5° for GR-1 (1.875° × 1.875° for precipitation), and 1.875° × 1.9° for EH5OM. For the downscaling, all fields are regridded to the finest common resolution.

This study deals with present climate throughout, either observed/analyzed or simulated from greenhouse gas concentrations that are prescribed according to observed or projected (up to 2010) values (IPCC 2007).

## 5. Results

All ClimDEX indices are based on annual values (either originally or aggregated by us from monthly values). For the stochastic methods we always show the first realization in the time series plots. All single test results can be inspected from Table 4 (test 1), Table 5 (test 2), and Table 6 (test 3), and the combined results are found in Table 7. We present the full tables mainly

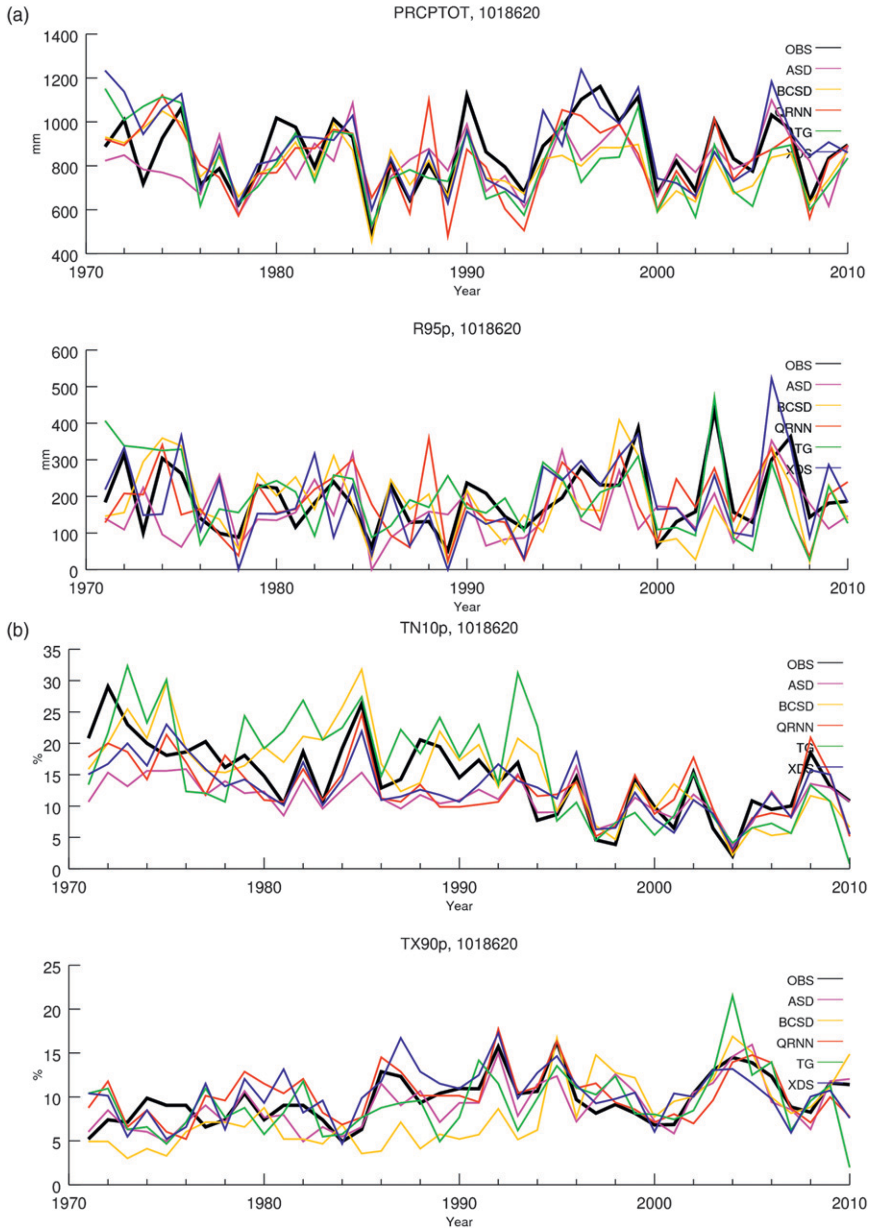

FIG. 2. (a) NCEP-downscaled vs observed annual values of PRCPTOT and R95p for the coastal region (Victoria, 1018620). For the stochastic methods (ASD, QRNN, and TG) we show one realization. (b) Same as (a), but for TN10p and TX90p.

for reasons of completeness, because their content is easily overwhelming and, apart from a few obvious features, difficult to assess. We attempt to interpret them along the three marginals of region (section 5a), ClimDEX (section 5b), and method (section 5c). Among the obvious general features is the better performance of most of the temperature-based methods as compared to precipitation. Moreover, indices derived from percentiles are better simulated as well. It is noteworthy that BCSD outperforms the other methods for the distribution tests, but shows poor performance for the sensitivity test 1. The combined test is passed by much less method–index pairs, with XDS showing the best results.

## *a. Regions*

### 1) COAST

The time series of four typical indices for the coastal region (C) are shown in Fig. 2. For PRCPTOT, interannual

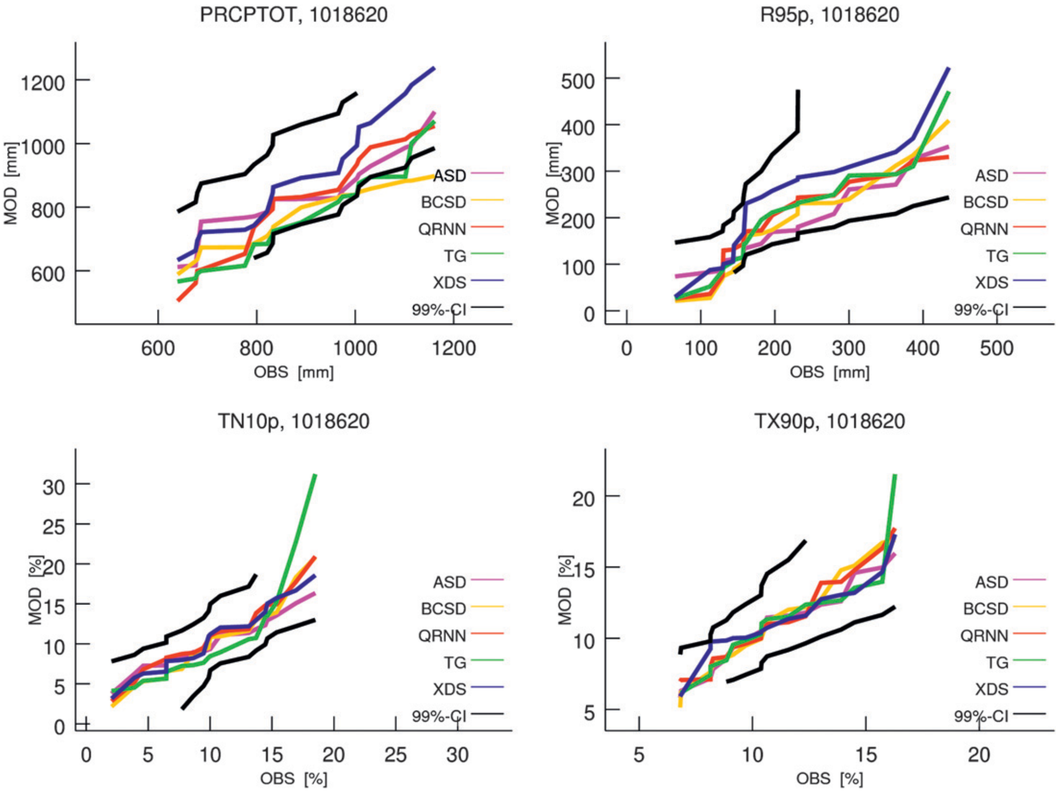

FIG. 3. The qq plot of NCEP-downscaled vs observed annual ClimDEX values for Victoria (1018620), based on data from 1991 to 2010.

variations are traced quite well in most methods, except perhaps for the persistent positive anomaly of the late 1990s whose full scale is only reflected by the XDS. For the verification period, TG appears to be least sensitive to climatic anomalies. A good example of this kind of climate sensitivity is the high-rainfall periods followed by sharp declines in the early 1980s and late 1990s. Peak precipitation years, as measured by very wet days (R95p), such as 1999 and 2003, are less well represented by the methods (1999 by XDS and 2003 by TG); note the strong overestimation of the year 2006 by XDS. For the temperature values, the apparent trend for both cool nights (TN10p, negative) and warm nights (TX90p, positive) is captured by all methods. TG and partly BCSD persistently overestimate TN10p, while XDS and QRNN are very similar and closer overall; ASD shows an underestimation in the early part. Single anomalous years are well represented in most of the methods, such as the "cold" anomalies in 1985, 1996, and 2008. For TX90p, the most obvious feature is the underestimation of BCSD in the early part of the series. The two warm years of 1992 and 1995 are well captured by all methods except BCSD (these years are relatively warm but not warm enough) and TG, while the extended warming observed throughout most of the 2000s is reflected in all methods.

The corresponding qq plot, based on the verification period, of four typical ClimDEX indices is depicted in Fig. 3. For the precipitation indices PRCPTOT and very wet days (R95p), only ASD, QRNN, and XDS are within the confidence band; note, however, that the first two are stochastic, and together with the other realizations might still fail the corresponding test 2 (and infact do so, cf. Fig. 6). Note the widening of the confidence band for the extreme quantiles, which reflects the larger uncertainty in estimating the likelihood of the corresponding events. For the two temperature indices TN10p and TX90p all methods are well within the confidence band and pass the test. The corresponding results for EH5OM/20C3M are shown in Fig. 4 and are overall similar. The only difference is an upward shift of the precipitation values across all methods; for ASD, this leads to a strong positive PRCPTOP bias and a failure of the corresponding test.

The results of test 1, that is, the annual correlation to observations from the verification period from 1991 to 2010, are shown in Fig. 5 for Victoria. There is a noticeable drop in performance for all precipitation derived values. Overall, XDS and QRNN perform best, with high scores for most temperature and many precipitation indices. Note that a corresponding analysis for

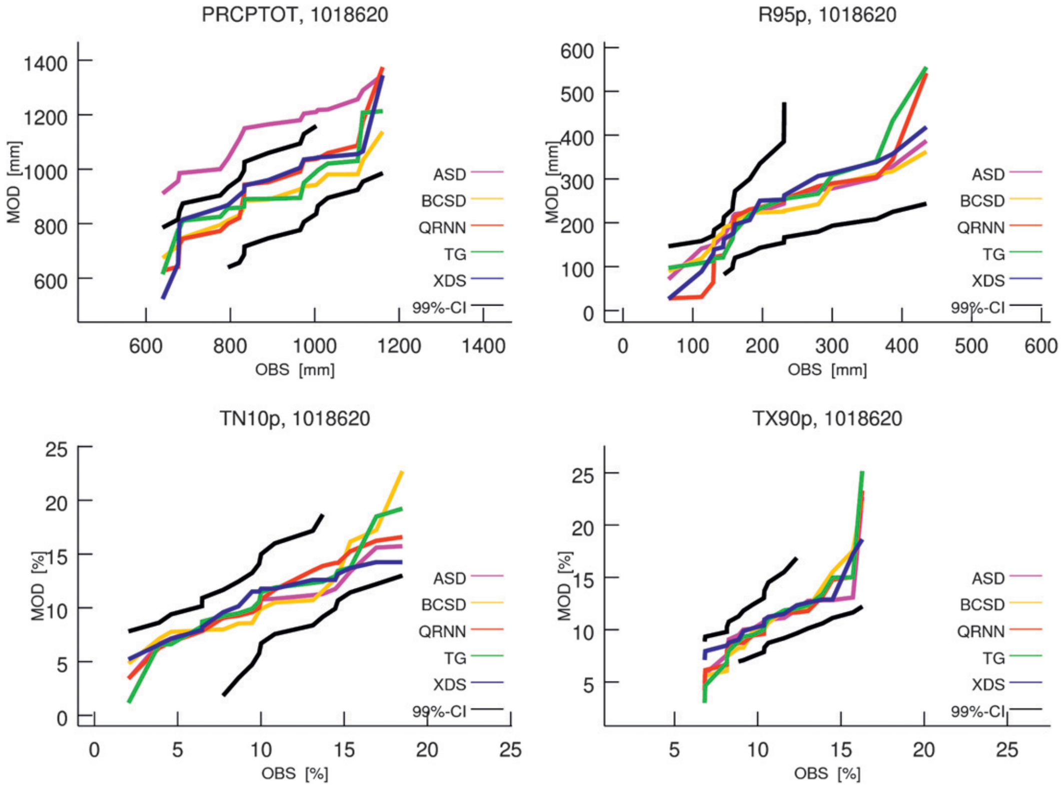

FIG. 4. As in Fig. 3, but for downscaling EH5OM/20C3M.

the longer period from 1971 to 2010 (not shown) reveals very similar results for most indices, indicative that overfitting is generally unlikely.

The passing of the distribution tests is shown in Fig. 6. Note that the tropical nights (TR), TN10p, and TX90p tests are passed by all methods. Additionally, BCSD passes for the most temperature- related indices and XDS for most precipitation-related indices. QRNN shows comparable performance for NCEP but drops sharply for the GCM.

### 2) MOUNTAINS

We show as an example the annual values for the maximum 5-day precipitation amount (RX5day) and simple daily intensity index (SDII) in Fig. 7, from the Mt. Fidelity station at 1875 m. XDS persistently overestimates RX5day, especially in the first half-period; QRNN, on the other hand, exhibits very little interannual variability. Both methods also overestimate SDII. With a few exceptions, the sensitivity to large-scale anomalies is weak, the exceptions being the year 1990 where all methods show positive anomalies of SDII. Note that none of the series shows a marked trend.

Figure 8 summarizes the results for test 1 for the near-mountaintop station at Mt. Fidelity (117CA90). As for the coastal region, XDS and QRNN have the largest number of significant correlations, which for XDS frequently approaches 0.9. The temperature-related indices pass the test for most methods.

### 3) TAIGA

For the single taiga station at Fort Nelson we show as an example annual values of minimum $T_n$ (TNn) and

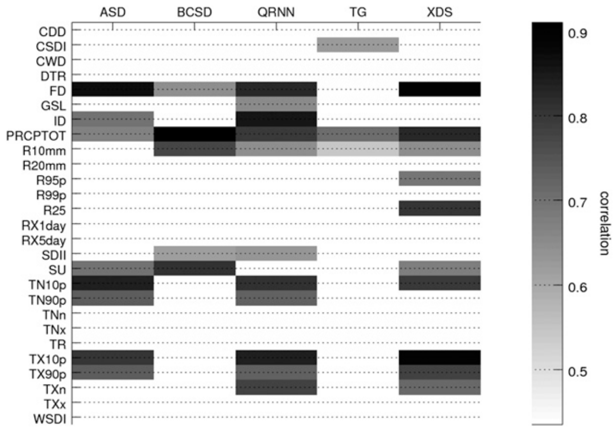

FIG. 5. Results of test 1 for the coastal region (Victoria). We show significantly nonzero correlations between NCEP-downscaled and observed indices. For the stochastic methods (ASD, QRNN, and TG), the average over all realizations is shown.

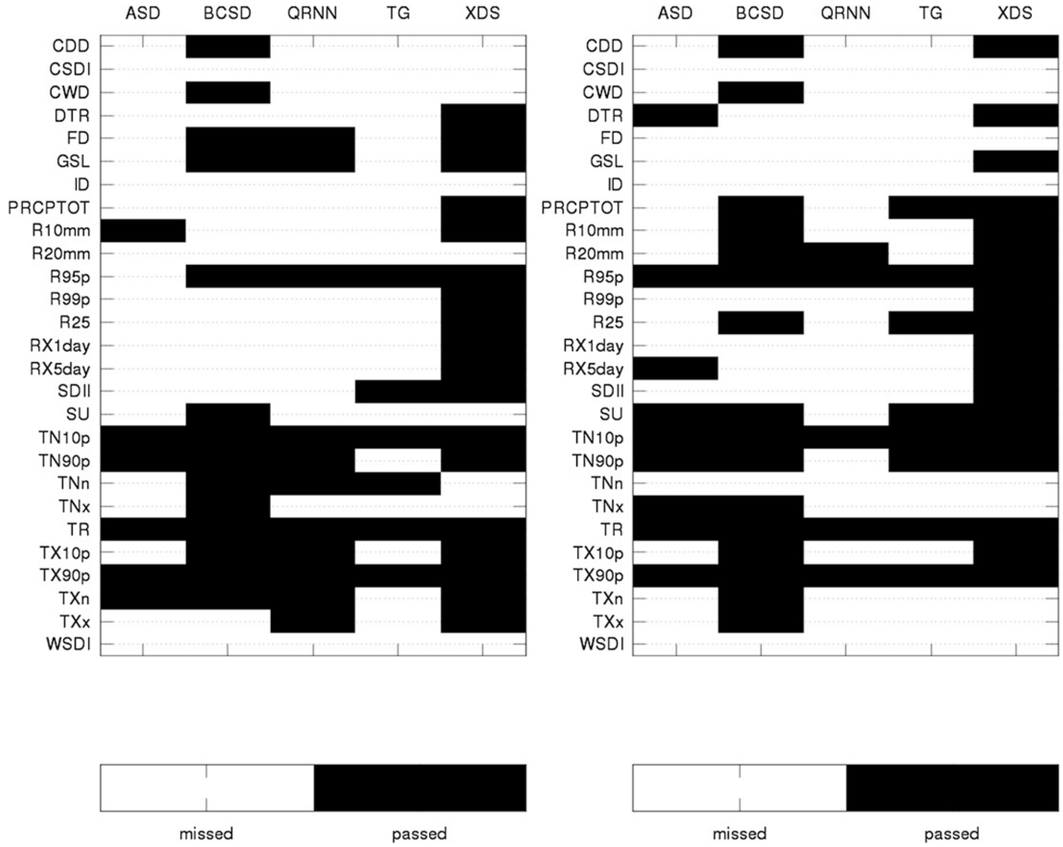

FIG. 6. Results of (left) test 2 (from NCEP) and (right) test 3 (from GCM) for the coastal region (Victoria).

maximum $T_x$ (TXx) in Fig. 9. It appears that for TNn, BCSD is least sensitive, especially in the early part. The cold anomalies of 1996 and 2008 are well reproduced by all methods; the persistent warming of the late 1980s, as well as the isolated warm year 1993, is visible in all simulations except BCSD. For TXx, XDS, and partly ASD, show a marked positive bias; sensitivity to interannual variations is weak in all methods.

Regarding actual anomalies for test 1, performance is lower than in the mountain region, as shown in Fig. 10. Again XDS and QRNN perform best, with XDS having some advantages for $P$-derived values and QRNN for those derived from $T_x$. The temporally more complicated indices [consecutive dry days (CDD), cold spell duration (CSDI), consecutive wet days (CWD), and diurnal T range (DTR)] do not pass for any of the methods.

### b. ClimDEX

Calculating the rate of passed tests relative to all methods and regions reveals the average results for each individual ClimDEX index. Figures 11a,b show for the single and combined tests that temperature-related indices are more easily downscaled (by our five methods) than those coming from precipitation (~50% for TX90p versus ~5% for R95p for the combined tests), which of course simply reflects that the corresponding raw daily series have a closer relation to the large-scale atmosphere. Note the symmetry between TN10p and TN90p on the one hand and TX90p and TX10p on the other.

Note that based on the combined tests, none of our methods is skillful for either some indices of very extreme events (R99p and TXx), or for indices representing a more complicated temporal pattern, such as those based on consecutive days.

### c. Methods

Figure 12a shows the rate of passed single tests across all indices and regions, the latter being weighted equally (to account for different station count). It is obvious that XDS shows the best performance, followed by, in that order, BCSD, QRNN, ASD, and TG. XDS is particularly good for the coastal region with about 60% of all tests passed. With about 50% of tests passed, BCSD shows best performance at the taiga station. The combined tests (cf. Fig. 12b) reveal stronger differences between regions and methods. XDS is best again, followed

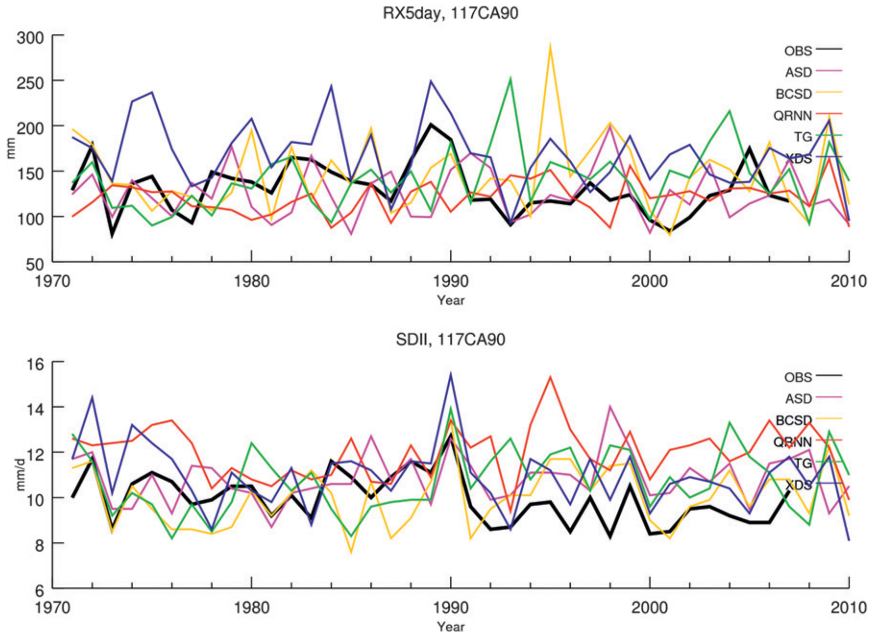

FIG. 7. The RX5day and SDII indices for the mountainous region (Mt. Fidelity).

by QRNN and BCSD, which show comparable performance, and ASD and TG. For the coast, XDS sticks out with about 25% of passed tests as compared to 5%–10% for the others. Interestingly, while performance for XDS drops for the other regions, it improves for the other methods, especially for BCSD and QRNN, with the latter reaching almost 20% of passed tests for the taiga.

## 6. Discussion

We have tested five different downscaling methods for the complex area of British Columbia. The methods cover a wide range of statistical downscaling, from a quantile mapping of monthly gridpoint data (BCSD), to quantile regressions using neural nets (QRNN), to ordinary, (automated) stepwise regression with variance adjustment (ASD) to ‘‘expanded’’ regression with full covariance preservation (XDS), and to a weather-type approach with stochastic resampling of within-type weather (TG). It was our goal to test the methods as they were used previously, including the selection of predictors. In some cases, however, the methods were altered in order to account for obvious shortcomings. For example, ASD needed extra corrections for cases when $T_n > T_x$, XDS employed a multivariate bias correction of the predictor fields, and TG included a within-type trend correction.

This setting leads to an average of roughly 10% (30%) of the better methods passing the combined (single) tests. As is shown, downscaling temperature extremes, whether on the cold or the hot end, can be done with moderate reliability for all test sites, regardless of method, and the most appropriate index to do so is TN10p (~60% passed tests) for monitoring cold and TX90p (~50%) for hot extremes. Both measure relatively moderate extremes, which should also leave enough room for parameter estimation and reduce the uncertainty for the downscaling of climate scenarios. For downscaling precipitation extremes no corresponding index exists that could be used regardless of region and method. To use RX5day, for example, one would need to inspect and find that it can only be downscaled reliably for one mountain station by using BCSD. The temporally more complex indices (CDD, CSDI, and

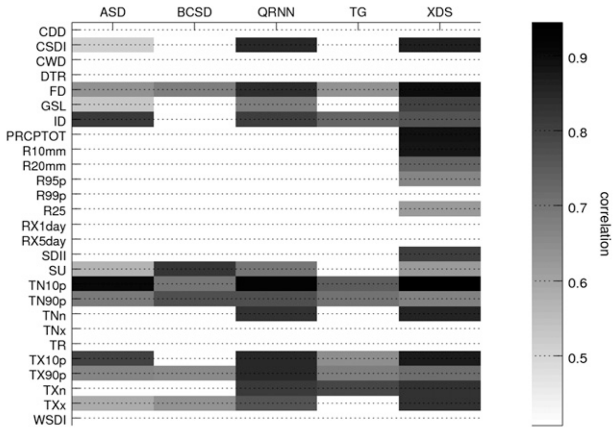

FIG. 8. As in Fig. 5, but for the mountainous region (Mt. Fidelity).

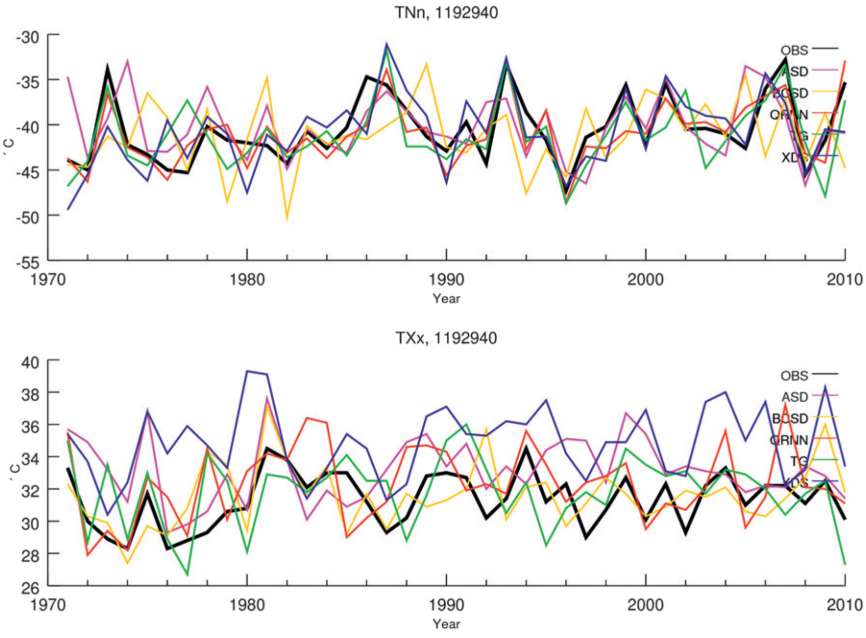

FIG. 9. The TNn and TXx index for the taiga region (Fort Nelson).

CWD), by passing none of the combined tests, cannot be downscaled by any of the methods at any test site with enough confidence.

With respect to regions one would expect a degradation of performance with increasing topographic disturbance, so that the best performance is achieved at coastal or mountaintop stations. This is true, however, only for XDS (and partly ASD), as evident from Figs. 8 and 12b. The opposite is true for QRNN, which actually follows from the fact that performance at the coast is poor for all methods except XDS. We have no explanation for this. Apart from that, improved performance for the taiga region is noticeable for TG and, in particular, QRNN. This is mainly based on the skill in temperature downscaling, as exemplified by Fig. 10.

Choosing ClimDEX for the tests had the folloing two advantages: a) by measuring ‘‘moderate’’ extremes they represent important and relevant tendencies of the climate system, which are at the same time statistically manageable; moreover, passing ClimDEX is a prerequisite for testing the far tail of a distribution that corresponds to singularly catastrophic events; b) because the statistical methods are trained on (daily) temperature and precipitation and not on ClimDEX itself, and most indices are a complicated composition of these base variables, performance of ClimDEX is more independent from the calibration time period and less prone to artificial skill from overfitting. This was verified in additional tests (not shown), which showed similar results with calibration data included.

Why do most methods perform so poorly at the coast? Comparing Figs. 12a,b for the coast, one sees a dramatic drop from the single to the combined testing, for at least BCSD and QRNN. For BCSD, Tables 5 and 6 show that this is mainly due to BCSD performing well for the distribution tests (test 2 and 3) and underperforming for the sensitivity test (test 1), and in particular for the temperature indices. This points to the fact that BCSD simulates $T_n$ and $T_x$ as a mere proxy, based on daily mean temperature and climatological diurnal temperature range. The quantile mapping may adjust for deficiencies in the distribution of daily mean temperature, but not for the actual minima and maxima. Work is under way to implement the direct downscaling of $T_n$

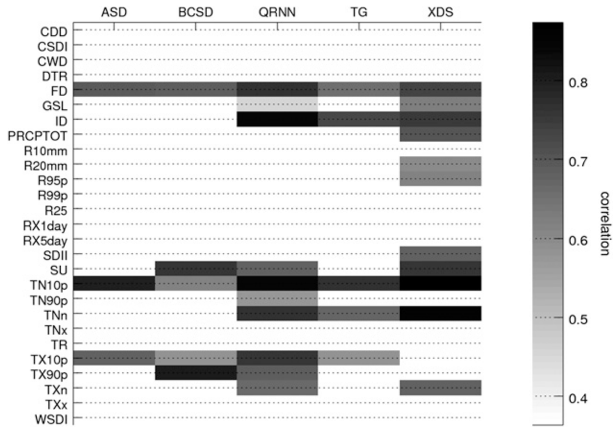

FIG. 10. As in Fig. 5, but for the taiga region (Fort Nelson).

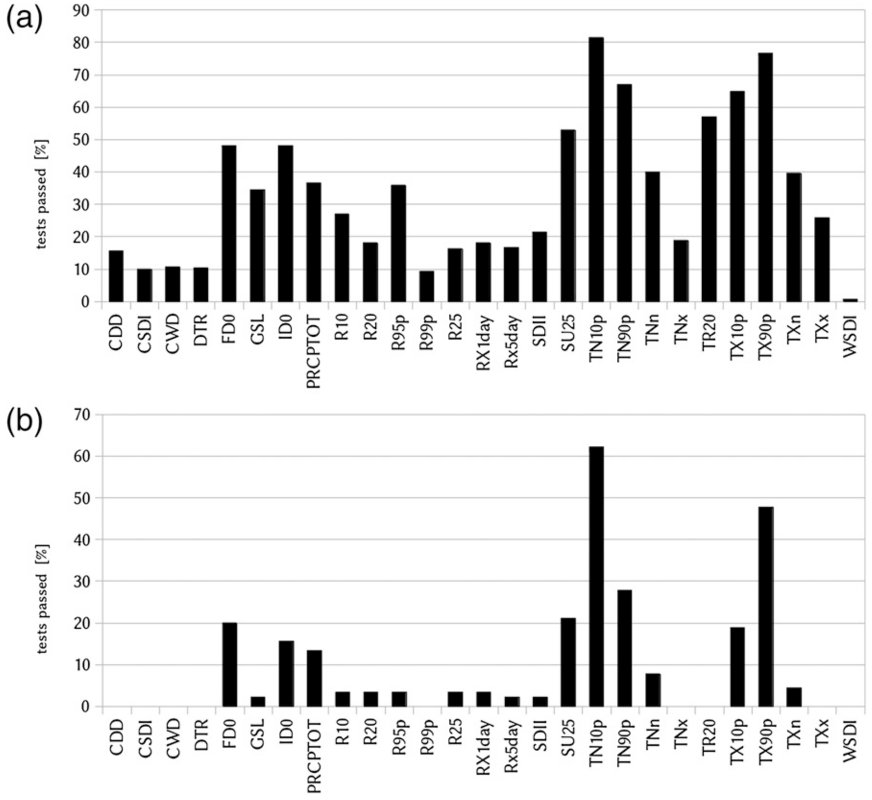

FIG. 11. (a) For each index, rate of single tests passed across regions and methods. Regions are weighted equally. (b) As in (a), but for the combined tests.

and $T_x$ into BCSD. QRNN, on the other hand, does a relatively good job for the NCEP-based tests 1 and 2, but does poorly in the GCM downscaling test 3. This may in part be due to its use of a relatively simple linear bias correction for GCM predictors as compared to the probit and quantile mapping transformations of XDS and BCSD. But as indicated, this is mere speculation and has, moreover, no particular coastal characteristic.

It is not easy to get a handle on the different performance statistics of the methods. ASD and XDS, for example, are both regression-based but show rather different skill, with XDS passing about 2.5 times as many tests as ASD for each region. We have run a couple of sensitivity experiments by exchanging certain modeling components, including the set of predictors, so that ASD operates on the larger predictor set of XDS, but we could not bring the methods into any better agreement. There are more advanced forms of ASD, making use, for example, of alternative predictor selection methods (e.g., partial correlation) and regression schemes (e.g., ridge regression), but based on the above experience we doubt that this is the root cause of the differences. It seems the main methodological difference, at least for precipitation, lies in the use of different predictands: whereas ASD predicts the probability of precipitation and simulates precipitation stochastically, XDS predicts precipitation directly.

All methods that are founded on a transfer function with parameters estimated from some principle of error minimization, such as regression or neural networks, have to face the fact that the simulated variability is reduced relative to observations. There are two ways to overcome this deficit—a stochastic way and a deterministic way. The former creates the missing variability purely stochastically by invoking specially calibrated statistical distributions, such as Gamma or Weibull. The latter does it by incorporating this requirement directly into the transfer function definition itself. For example, any quantile mapping approach that employs the ''empirical transformation'' of Panofsky and Brier (1958), such as the bias correction part of BCSD, is of this type; XDS does it by imposing an extra constraint on the error minimization; and for the analog method (Zorita and von Storch 1999) an entire list of candidate global and local fields is built into the transfer function (which therefore becomes *very* complicated).

The deterministic approach, specifically XDS and its simpler cousin ''inflation,'' has been denounced, on the

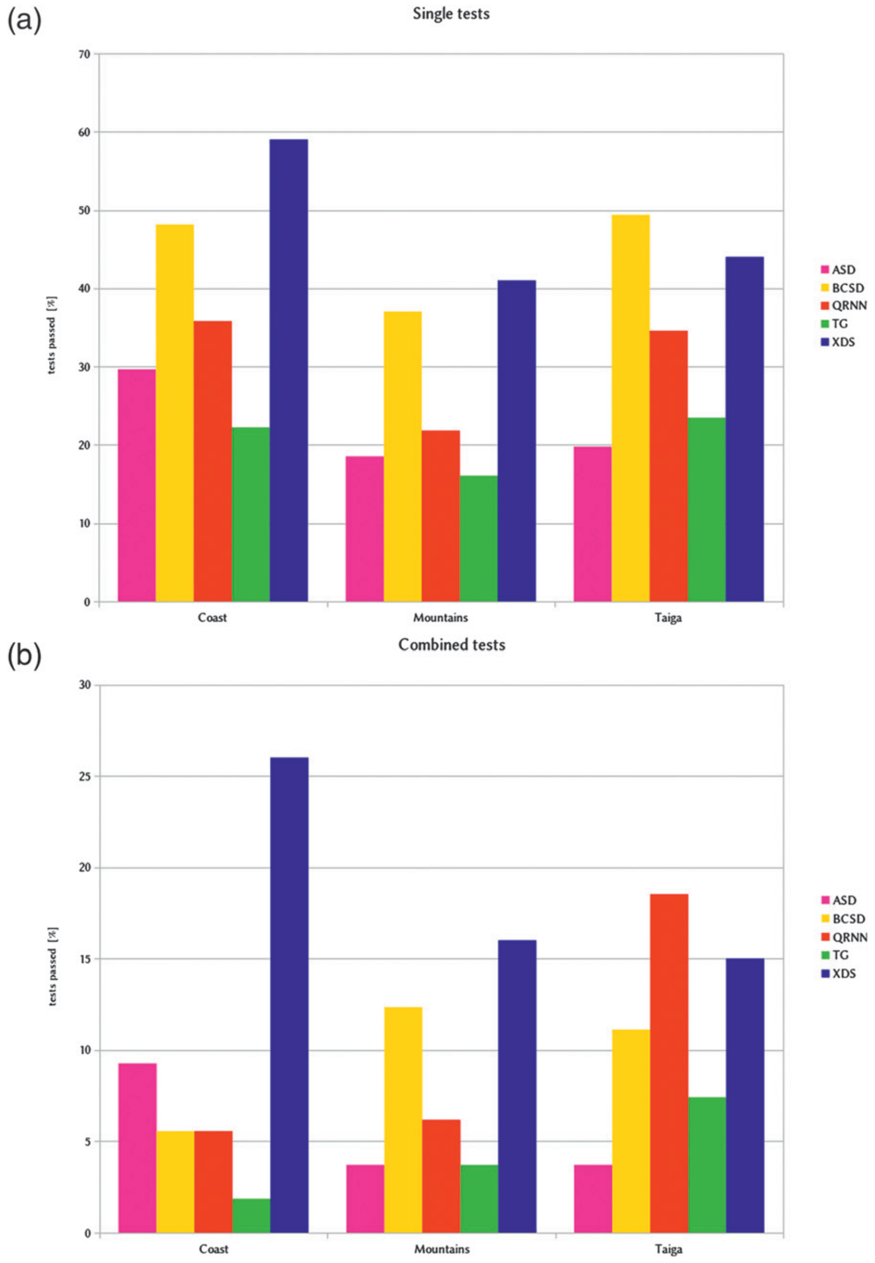

FIG. 12. (a) Rate of single tests passed for all regions and methods. (b) As in (a), but for combined tests.

other hand, by von Storch (1999) as being inappropriate for downscaling in general because all variability is inherited from the larger scales. He advocates ''randomization'' instead, where stochastic noise is *added* to the regression model. This was briefly discussed previously in Bürger and Chen (2005), and here we also provide a short appendix where the main statistical characteristics of both methods are illustrated, including a simple example. We stress that any variance adjustment, whether it is stochastic or deterministic, is in conflict with the (unconstrained) regression approach and will inevitably lead to a larger model error, as exemplified in the appendix. Much of the confusion, it seems, is caused by the notion of variability, which ought to be taken as a climatic amplitude measure and not, as von Storch (1999) does, in the sense of individual fluctuations and the related concept of explained variance. Inflation does not aim to explain any variations that were left unexplained by the regression. On the contrary, as shown in the appendix (and as obvious from the optimality of regression), any posterior variance adjustment explains less variance than regression; this is simply the price for preserving variance. And the regression model degrades quickly when variance is adjusted. If predictor–predictand correlations fall below ~0.7, then the explained variance of the randomization model starts to vanish (becomes negative); for inflation this happens only for values below ~0.5. However, with no variance explained from the large scales (equivalent to forecasting climatology) such models are not any better than the most basic weather generators. This should be borne in mind when applying variance adjustment techniques to future scenarios, and it is also the main reason why in test 1 we test the sensitivity to actual climate anomalies. Finally, if XDS or any other deterministic method was inappropriate for downscaling this would likely have been revealed in our tests.

Thus, should the stochastic component be responsible for some of the reduced skill that we see for ASD, QRNN, and TG? On top of that, and related to it, recall that to pass the distribution tests only the small fraction of $\alpha = 1\%$ of all realizations to fail the test were allowed. A slight misrepresentation of the stochastic component, hence, may easily lead to such a failure. The stochasticity may therefore seem somewhat unfair as compared to the deterministic methods BCSD and XDS. It is, however, just the flip side of the greater chance of a realization passing a test when in fact the method should not.

Of all methods BCSD uses the least large-scale information and, at least based on the single tests (Fig. 12a), performs comparatively well. Given the ''cheap'' input it is quite versatile and can be used to generate large ensembles of downscaled scenarios. Compared to that, all other methods require input that many of the GCMs do not offer (such as daily upper-level fields). BCSD can be improved to genuinely model $T_n$ and $T_x$ and can be expected to pass more of the temperature and, hence, the combined tests. Whether this new form compares favorably to the more expensive methods needs to be seen.

One would think that to do downscaling work in one of our study regions simply requires inspecting Tables 4–7, depending on the application, and picking for each index of interest one of the passing methods. However, despite the strict testing setup all results still depend on the fixed framework in which they were derived, most notably the use of GR-1 for the reanalysis fields [which is now superseded by GR-2 (cf. Kanamitsu et al. 2002)] and EH5OM run 1 for the GCM. Whether the results are robust against the use of alternate GCMs is unknown. It is in fact doubtful whether they persist for the much more abundant GCMs of lower resolution.

*Acknowledgments.* Dave Spittlehouse initiated this study 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 who provided useful comments to the manuscript, as well as to Hailey Eckstrand for preparing the figures.

## APPENDIX A

### Inflation versus Randomization

In the simplest of all cases, the predictor variable $x$ and the predictand variable $y$ are related as follows:

$$y = xr + \sqrt{1 - r^2}\varepsilon. \tag{A1}$$

We assume without loss of generality that $x$, $y$, and $\varepsilon$ (which represents the unresolved error) are zero mean Gaussian and that $x$ and $\varepsilon$ have unit variance and are uncorrelated. The correlation between $x$ and $y$ then simply equals $r$. If one observes a sample of realizations of $x$ and $y$ one can regress $y$ on $x$ (RGR), which then leads to the well-known regression coefficient $a = x\backslash y$ and which approaches $r$ with increasing sample size. The variance of the regression-simulated predictand $\hat{y} = xa$ is, accordingly, $a^2$ (in the limit $r^2$), so that $1 - a^2$ measures the amount of variance unexplained by the linear regression.

To adjust the missing variance two methods are in use, inflation (IFL) and randomization (RND)

$$\begin{aligned} \text{IFL:}\, \hat{y}_{\mathbf{I}} &= \hat{y}\sigma_y/\sigma_{\hat{y}} \\ \text{RND:}\, \hat{y}_{\mathbf{R}} &= \hat{y} + \sqrt{(1 - a^2)}\eta, \end{aligned} \tag{A2}$$

where $\eta$ denotes another $N(0, 1)$ process uncorrelated to $\hat{y}$. From the definitions (A2) the main simulation characteristics such as variance, correlation to observations ($\rho$), and explained variance (EV) are straightforward (noting that generally EV = $2\rho - 1$ if observed and simulated variance both equal 1). Along with an example of two time series of length 1000 with $r = 0.6$, they are shown in Table A1.

The first row expresses that variance is in fact adjusted by both methods; the second row means that local correlation is preserved by IFL but is degraded by RND in proportion to $r$; in terms of simulation error, or equivalently EV (third row), RGR is optimal followed by IFL followed by RND. IFL has positive EV values for $r > 0.5$ and RND for $r > \sqrt{1/2} \sim 0.7$. In real world cases when

TABLE A1. Regression, inflation, and randomization characteristics.

| Example: $r = 0.6$ | RGR | IFL | RND |
|---|---|---|---|
| Model | $\hat{y} = xa$ | $\hat{y}_\mathbf{I} = \hat{y}\sigma_y/\sigma_{\hat{y}}$ | $\hat{y}_\mathbf{R} = \hat{y} + \sqrt{(1 - a^2)}\eta$ |
| Variance | $r^2$ | 1 | 1 |
| True, estimate | 0.36, 0.35 | 1.00, 1.01 | 1.00, 1.00 |
| Correlation | $r$ | $r$ | $r^2$ |
| True, estimate | 0.60, 0.59 | 0.60, 0.59 | 0.36, 0.32 |
| EV | $r^2$ | $2r - 1$ | $2r^2 - 1$ |
| True, estimate | 0.36, 0.34 | 0.20, 0.17 | −0.28, −0.35 |

precipitation is the predictand, $r$ is usually in the range of 0.4–0.7. The table was produced from Octave/Matlab code (available online at http://xds.googlecode.com/git/ifl_rnd.m).

## REFERENCES

Allen, D., A. Cannon, M. Toews, and J. Scibek, 2010: Variability in simulated recharge using different GCMs. *Water Resour. Res.,* **46,** W00F03, doi:10.1029/2009WR008932.

Benestad, R. E., 2010: Downscaling precipitation extremes. *Theor. Appl. Climatol.,* **100,** 1–21.

Bronstert, A., 2004: Rainfall runoff modelling for assessing impacts of climate and land use change. *Hydrol. Processes,* **18,** 567–570.

Bürger, G., 1996: Expanded downscaling for generating local weather scenarios. *Climate Res.,* **7,** 111–128.

——, and Y. Chen, 2005: Regression-based downscaling of spatial variability for hydrologic applications. *J. Hydrol.,* **311** (1–4), 299–317.

——, D. Reusser, and D. Kneis, 2009: Early flood warnings from empirical (expanded) downscaling of the full ECMWF Ensemble Prediction System. *Water Resour. Res.,* **45,** W10443, doi:10.1029/2009WR007779.

Busuioc, A., R. Tomozeiu, and C. Cacciamani, 2008: Statistical downscaling model based on canonical correlation analysis for winter extreme precipitation events in the Emilia Romagna region. *Int. J. Climatol.,* **28,** 449–464.

Cannon, A. J., 2008: Probabilistic multisite precipitation downscaling by an expanded Bernoulli–Gamma density network. *J. Hydrometeor.,* **9,** 1284–1300.

——, 2011: Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Comput. Geosci.,* **37,** 1277–1284, doi:16/j.cageo.2010.07.005.

——, P. H. Whitfield, and E. R. Lord, 2002: Synoptic map-pattern classification using recursive partitioning and principal component analysis. *Mon. Wea. Rev.,* **130,** 1187–1206.

Chen, S. C., 2002: Model mismatch between global and regional simulations. *Geophys. Res. Lett.,* **29,** 1060, doi:10.1029/2001GL013570.

Christensen, N., and D. P. Lettenmaier, 2007: A multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the Colorado River Basin. *Hydrol. Earth Syst. Sci.,* **11,** 1417–1434.

——, A. W. Wood, N. Voisin, D. P. Lettenmaier, and R. N. Palmer, 2004: The effects of climate change on the hydrology and water resources of the Colorado River basin. *Climatic Change,* **62,** 337–363.

Dehn, M., and J. Buma, 1999: Modelling future landslide activity based on general circulation models. *Geomorphology,* **30** (1–2), 175–187.

——, G. Bürger, J. Buma, and P. Gasparetto, 2000: Impact of climate change on slope stability using expanded downscaling. *Eng. Geol.,* **55,** 193–204.

Department of Defense, cited 2002: DoD News Briefing—Secretary Rumsfeld and Gen. Myers. [Available online at http://www.defense.gov/transcripts/transcript.aspx?transcriptid=2636.]

Dibike, Y. B., and P. Coulibaly, 2006: Temporal neural networks for downscaling climate variability and extremes. *Neural Networks,* **19,** 135–144.

Dryden, I. L., and K. V. Mardia, 1998: *Statistical Shape Analysis.* Wiley, 376 pp.

Easterling, D. R., 1999: Development of regional climate scenarios using a downscaling approach. *Climatic Change,* **41,** 615–634.

Ebisuzaki, W., 1997: A method to estimate the statistical significance of a correlation when the data are serially correlated. *J. Climate,* **10,** 2147–2153.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi, 2007: Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.,* **27,** 1547–1578.

Gachon, P., and Y. Dibike, 2007: Temperature change signals in northern Canada: Convergence of statistical downscaling results using two driving GCMs. *Int. J. Climatol.,* **27,** 1623–1641.

Harpham, C., and R. L. Wilby, 2005: Multi-site downscaling of heavy daily precipitation occurrence and amounts. *J. Hydrol.,* **312** (1–4), 235–255.

Hay, L. E., and M. P. Clark, 2003: Use of statistically and dynamically downscaled atmospheric model output for hydrologic simulations in three mountainous basins in the western United States. *J. Hydrol.,* **282** (1–4), 56–75, doi:10.1016/S0022-1694(03)00252-X.

Hayhoe, K., and Coauthors, 2007: Past and future changes in climate and hydrological indicators in the US Northeast. *Climate Dyn.,* **28,** 381–407.

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby, and C. M. Goodess, 2006: Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios. *Int. J. Climatol.,* **26,** 1397–1415.

Hessami, M., P. Gachon, T. B. M. J. Ouarda, and A. St-Hilaire, 2008: Automated regression-based statistical downscaling tool. *Environ. Modell. Software,* **23,** 813–834.

Huth, R., 2004: Sensitivity of local daily temperature change estimates to the selection of downscaling models and predictors. *J. Climate,* **17,** 640–652.

IPCC, cited 2007: Scenario data for the atmospheric environment. [Available online at http://www.ipcc-data.org/sres/ddc_sres_emissions.html.]

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S. K. Yang, J. Hnilo, M. Fiorino, and G. Potter, 2002: NCEP–DOE AMIP-II Reanalysis (R-2). *Bull. Amer. Meteor. Soc.,* **83,** 1631–1644.

Khan, M. S., P. Coulibaly, and Y. Dibike, 2006: Uncertainty analysis of statistical downscaling methods. *J. Hydrol.,* **319** (1–4), 357–382.

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,** 3760–3788.

——, ——, X. Zhang, and G. C. Hegerl, 2007: Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. *J. Climate,* **20,** 1419–1444.

Kundzewicz, Z., and E. Stakhiv, 2010: Are climate models ''ready for prime time'' in water resources management applications, or is more research needed? *Hydrol. Sci. J.,* **55,** 1085–1089.

Kysely, J., 2002: Comparison of extremes in GCM-simulated, downscaled and observed central-European temperature series. *Climate Res.,* **20,** 211–222.

Mannshardt-Shamseldin, E. C., R. L. Smith, S. R. Sain, L. O. Mearns, and D. Cooley, 2010: Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data. *Ann. Appl. Stat.,* **4,** 484–502.

Maraun, D., and Coauthors, 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.,* **48,** RG3003, doi:10.1029/2009RG000314.

Maurer, E. P., and P. B. Duffy, 2005: Uncertainty in projections of streamflow changes due to climate change in California. *Geophys. Res. Lett.,* **32,** L03704, doi:10.1029/2004GL021462.

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

——, L. Brekke, T. Pruitt, and P. Duffy, 2007: Fine-resolution climate projections enhance regional climate change impact studies. *Eos, Trans. Amer. Geophys. Union,* **88,** 504, doi:10.1029/2007EO470006.

Meehl, G. A., G. J. Boer, C. Covey, M. Latif, and R. J. Stouffer, 2000: The Coupled Model Intercomparison Project (CMIP). *Bull. Amer. Meteor. Soc.,* **81,** 313–318.

Menzel, L., and G. Bürger, 2002: Climate change scenarios and runoff response in the Mulde catchment (Southern Elbe, Germany). *J. Hydrol.,* **267** (1–2), 53–64.

——, A. H. Thieken, D. Schwandt, and G. Bürger, 2006: Impact of climate change on the regional hydrology–scenario-based modelling studies in the German Rhine catchment. *Nat. Hazards,* **38,** 45–61.

Min, S.-K., X. Zhang, F. W. Zwiers, and G. C. Hegerl, 2011: Human contribution to more-intense precipitation extremes. *Nature,* **470,** 378–381, doi:10.1038/nature09763.

Müller-Wohlfeil, D. I., G. Bürger, and W. Lahmer, 2000: Response of a river catchment to climatic change: Application of expanded downscaling to northern Germany. *Climatic Change,* **47,** 61–89.

Olsson, J., C. Uvo, and K. Jinno, 2001: Statistical atmospheric downscaling of short-term extreme rainfall by neural networks. *Phys. Chem. Earth,* **26B,** 695–700.

Panofsky, H. A., and G. W. Brier, 1958: *Some Applications of Statistics to Meteorology*. The Pennsylvania State University, 224 pp.

Peterson, T. C., 2005: Climate change indices. *WMO Bull.,* **54,** 83–86.

Quinonero-Candela, J., C. Rasmussen, F. Sinz, O. Bousquet, and B. Schölkopf 2006: Evaluating predictive uncertainty challenge. *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment,* J. Quinonero-Candela et al., Eds., Springer, 1–27.

Salathe, E. P., Jr., 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. *Int. J. Climatol.,* **27,** 1611–1621.

Sausen, R., K. Barthel, and K. Hasselmann, 1988: Coupled ocean–atmosphere models with flux correction. *Climate Dyn.,* **2,** 145–163.

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,** D04105, doi:10.1029/2005JD007026.

Schnorbus, M., K. Bennett, A. Werner, and A. Berland, 2011: Hydrologic impacts of climate change in the Peace, Campbell and Columbia Watersheds, British Columbia, Canada. Pacific Climate Impacts Consortium Hydrologic Modelling Project Final Rep., 175 pp. [Available online at http://pacificclimate.org/sites/default/files/publications/Schnorbus.HydroModelling.FinalReport2.Apr2011.pdf.]

Schubert, S., and A. Henderson-Sellers, 1997: A statistical model to downscale local daily temperature extremes from synoptic-scale atmospheric circulation patterns in the Australian region. *Climate Dyn.,* **13,** 223–234.

Seber, G. A. F., and A. J. Lee, 2003: *Linear Regression Analysis*. Wiley, 582 pp.

Sillmann, J., and E. Roeckner, 2007: Indices for extreme events in projections of anthropogenic climate change. *Climatic Change,* **86,** 83–104, doi:10.1007/s10584-007-9308-6.

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 Resour. Res.,* **44,** W02422, doi:10.1029/2007WR005956.

Taylor, J. W., 2000: A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *J. Forecasting,* **19,** 299–311.

Tebaldi, C., K. Hayhoe, J. M. Arblaster, and G. A. Meehl, 2006: Going to the extremes. *Climatic Change,* **79,** 185–211.

VanRheenen, N. T., A. W. Wood, R. N. Palmer, and D. P. Lettenmaier, 2004: Potential implications of PCM climate change scenarios for Sacramento–San Joaquin River Basin hydrology and water resources. *Climatic Change,* **62,** 257–281.

von Storch, H., 1999: On the use of ''inflation'' in statistical downscaling. *J. Climate,* **12,** 3505–3506.

Vrac, M., and P. Naveau, 2007: Stochastic downscaling of precipitation: From dry events to heavy rainfalls. *Water Resour. Res.,* **43,** W07402, doi:10.1029/2006WR005308.

Wilby, R. L., 2005: Uncertainty in water resource model parameters used for climate change impact assessment. *Hydrol. Processes,* **19,** 3201–3219.

——, L. E. Hay, and G. H. Leavesley, 1999: A comparison of downscaled and raw GCM output: Implications for climate change scenarios in the San Juan River basin, Colorado. *J. Hydrol.,* **225** (1–2), 67–91.

——, C. W. Dawson, and E. M. Barrow, 2002: SDSM—A decision support tool for the assessment of regional climate change impacts. *Environ. Modell. Software,* **17,** 145–157.

——, O. Tomlinson, and C. Dawson, 2003: Multi-site simulation of precipitation by conditional resampling. *Climate Res.,* **23,** 183–194.

——, S. P. Charles, E. Zorita, B. Timbal, P. Whetton, and L. O. Mearns, 2004: Guidelines for use of climate scenarios developed from statistical downscaling methods. IPCC Task Group on Data and Scenario Support for Impacts and Climate

Analysis, 27 pp. [Available online at http://www.narccap.ucar.edu/doc/tgica-guidance-2004.pdf.]

——, P. Whitehead, A. Wade, D. Butterfield, R. Davis, and G. Watts, 2006: Integrated modelling of climate change impacts on water resources and quality in a lowland catchment: River Kennet, UK. *J. Hydrol.,* **330** (1–2), 204–220.

Wood, A., E. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.,* **107,** 4429, doi:10.1029/2001JD000659.

Zhang, X., and Coauthors, 2011: Indices for monitoring changes in extremes based on daily temperature and precipitation data. *WIREs Climate Change,* **2,** 851–870, doi:10.1002/wcc.147.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate,* **12,** 2474–2489.