RESEARCH ARTICLE

Evaluation of some distributional downscaling methods as applied to daily maximum temperature with emphasis on extremes

John R. Lanzante1 | Dennis Adams-Smith1,2 | Keith W. Dixon1 | Maryjo Nath1 | Carolyn E. Whitlock1,3

1National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL), Princeton, New Jersey
2University Corporation for Atmospheric Research (UCAR)/Cooperative Programs for the Advancement of Earth System Science (CPAESS), Boulder, Colorado
3SAIC, Reston, Virginia

Correspondence
John R. Lanzante, NOAA/Geophysical Fluid Dynamics Laboratory, 201 Forrestal Road, Princeton, NJ 08542.
Email: john.lanzante@noaa.gov

Abstract
Statistical downscaling methods are extensively used to refine future climate change projections produced by physical models. Distributional methods, which are among the simplest to implement, are also among the most widely used, either by themselves or in conjunction with more complex approaches. Here, building off of earlier work we evaluate the performance of seven methods in this class that range widely in their degree of complexity. We employ daily maximum temperature over the Continental U.S. in a “Perfect Model” approach in which the output from a large-scale dynamical model is used as a proxy for both observations and model output. Importantly, this experimental design allows one to estimate expected performance under a future high-emissions climate-change scenario. We examine skill over the full distribution as well in the tails, seasonal variations in skill, and the ability to reproduce the climate change signal. Viewed broadly, there generally are modest overall differences in performance across the majority of the methods. However, the choice of philosophical paradigms used to define the downscaling algorithms divides the seven methods into two classes, of better versus poorer overall performance. In particular, the bias-correction plus change-factor approach performs better overall than the bias-correction only approach. Finally, we examine the performance of some special tail treatments that we introduced in earlier work which were based on extensions of a widely used existing scheme. We find that our tail treatments provide a further enhancement in downscaling extremes.

KEYWORDS
bias-correction, distributions, perfect model evaluation, statistical downscaling, tail values

1 | INTRODUCTION

In the quest to assess the possible impacts of climate change, the output from Global Climate Models (GCM’s) plays a central role. Increasingly though, refinement, in terms of greater spatial resolution, is being achieved by Regional Climate Models (RCM’s), in order to better represent local conditions. In spite of their utility, physical models such as GCM’s and RCM’s, as well as reanalysis systems, suffer from two important drawbacks: their expense and residual biases.
Increasingly statistical downscaling (SD) techniques are being employed to address deficiencies in dynamical models. Such techniques are typically much less costly to implement, have the potential to mitigate biases, and, in theory, can be applied to any arbitrary spatial scale. When used in a climate change setting, SD methods operate by first constructing a statistical model that is trained via comparison of observations and physical model output from some past period. By applying the statistical model to output from a physical model pertinent to some future climate change scenario, it is possible to generate “future observations” which can then be compared to actual observations from the past to infer possible effects of climate change.

One of the crucial implicit assumptions is that the statistical model created from a historical period is applicable to a future scenario in which the climate has fundamentally changed. The validity of this assumption of “statistical stationarity” is difficult to assess lacking actual observations from future climates of interest. Instead, analysts sometimes resort to “Perfect Model” (PM) frameworks in which the output from physical models is used to generate both pseudo-observations and pseudo-model output for both historical and future climates. In such a setting, the analyst can compare the SD products with a proxy for the “truth.”

In the burgeoning field of SD, there are a plethora of approaches ranging from simple to highly complex. Unfortunately, critical assessment of such techniques has not kept up with their proliferation. Recently in this regard more systematic efforts have arisen, the most comprehensive of which, with a European focus, is the VALUE network (Maraun et al., 2015). More humble efforts involving the authors of this work, members of the GFDL (Geophysical Fluid Dynamics Laboratory) Empirical Statistical Downscaling (ESD) team (https://www.gfdl.noaa.gov/esd_eval) and collaborators have recently focused on the Continental United States (CONUS) (Dixon et al., 2016; Lanzante et al., 2018; Lanzante et al., 2019, hereafter referred to as L19).

Although there is no unique way to classify SD methods Lanzante et al. (2018) propose four classes:

1. direct transfer function,
2. distributional mapping,
3. spatial mapping, and
4. weather generators.

Examples from (1) include linear regression and neural nets, from (2) quantile mapping, from (3) analogue approaches and from (4) stochastic models. In practice a particular SD method may incorporate techniques from more than one category. Distributional methods are perhaps the most widely employed and are not only used by themselves but are frequently part of more complex algorithms. For example, methods that collectively account for a major share of U.S. downscaled data available on public servers all incorporate a distributional transform (Lanzante et al., 2018). Note that distributional methods are trained using the output from a physical model they also fall under the classification of Model Output Statistics (MOS; Wilks, 2006).

Distributional methods are appealing because most are relatively simple, both conceptually as well as computationally, and can potentially mitigate biases not only in the mean and variance but also in higher-order properties across the distribution. These methods operate by utilizing frequency distributions of: historical observations ($O_h$), historical model values ($M_h$), and for some methods future model values ($M_f$). Application results in generation of downscaled and bias-corrected future values ($D_f$) which serve as proxies for future observations ($O_f$).

Additional appeal of distributional methods stems from the fact that they are trained asynchronously – they do not require observed and model data to share the same weather sequencing. Many types of dynamical model simulations that by design are not suitably constrained by observations lack such synchronicity. Given multiple advantages, distributional methods have perhaps the most widespread reach in SD. While they are often applied by themselves, in addition they are frequently embedded within more complex methods (e.g., BCSD (Wood et al., 2004), BCCA (Maurer et al., 2010), MACA (Abatzoglou and Brown, 2012), and LOCA (Pierce et al., 2014), to name a few). Part of the reason for this is that distributional methods applied by themselves may introduce artefacts due to a spatial scale mismatch between model and observations (Maraun, 2013). Combining methods with complementary strengths is an attempt to “get the best of both worlds.”

A complication arises in that there is no a priori way to justify any particular means by which to combine the various distributions as the “best” approach. To better understand differences between methods it is useful to consider two distinct paradigms: bias-correction versus change-factor (or delta) (Ho et al., 2012; Lanzante et al., 2018). These two paradigms differ in that bias-correction methods adjust model output for consistency with observations, whereas change-factor methods adjust observations for consistency with model-simulated changes over time. A particular distributional SD method may apply one or both of these paradigms. Although distributional methods have been found to be generally useful (Fowler et al., 2007; Maraun et al., 2010; Gutierrez et al., 2018) in certain situations poor performance may be traced to reliance on one of these paradigms (Lanzante et al., 2018).

This study builds upon the evaluation approach developed recently (L19) aimed at assessing one of the methods (Michelangeli et al., 2009) in the tails. The tails of distributions are of particular interest since many climate impacts are disproportionately affected by extremes and since tail
evaluations have been less common than overall evaluations (L19). Here, although we emphasize tail performance, we extend our work to also consider performance over the entire distribution, as well as seasonality and the ability to replicate the climate change signal. Most importantly we extend our evaluation from one to seven distributional SD methods.

2 | DATA AND METHODOLOGY

2.1 | Data

We employ data from the GFDL-HiRAM-C360 model which were recently analysed by Lanzante et al. (2018) and L19. We only briefly describe them here and refer the reader to Dixon et al. (2016) for more details and to L19 (Appendix A) for data availability. These data are PM daily maximum temperatures in a rectangular region covering the CONUS, but excluding oceanic points (Pacific and Atlantic, and Gulf of Mexico).

Historical (future) data correspond to 1979–2008 (2086–2095). There are 30 years of data for both historical (one ensemble member) and future (three 10-year members) time periods run under a high emissions scenario (RCP8.5). We downscale each ensemble separately and average verification statistics over the three members. As in L19, downscaling is performed separately for each gridpoint and each of 12 months. The \( O_h \) and \( O_t \) values are simply the raw GCM output (spatial resolution ~25 km). The \( M_h \) and \( M_t \) values are spatially averaged versions of the \( O_h \) and \( O_t \) values, respectively, yielding a spatial resolution of ~200 km. The mismatch in spatial resolution, which poses a challenge to SD methods, is typically what an impacts analyst would face in real-world applications using the current generation of GCM’s. For simplicity, here we refer to \( O_h \) and \( O_t \) (\( M_h \) and \( M_t \)) as “observations” (“model output”) in our PM world, even though all data values are technically derived from a GCM.

2.2 | Downscaling methodology

We employ seven techniques whose SD is accomplished via distributional transforms based on R-language code developed in-house or modified versions of code obtained from external collaborators. The SD techniques range from very simple, whose core algorithms require only a few lines of code, to more complex, requiring hundreds of lines of code. Below we briefly describe each SD method, referring the reader to other works for more details.

In L19 one of the focal points was tail values that were produced under “out-of-bounds” (OOB) conditions, which are expected to increase in frequency due to climate change for some variables such as daily maximum temperature. For some distributional methods there are conditions under which the base algorithm is not able to produce downscaled values, starting from some point in the distribution out to the end of one tail. For any sample of data OOB conditions may occur for one, both or neither tails – it then becomes necessary to utilize an alternate scheme. The typical remedy has been the approach of Deque (2007), hereafter D07, described in more detail and mathematically by L19, of using a “constant correction.” For the upper (lower) tail this defines an offset as the difference, \( D_t \) minus \( M_t \) when the base downscaling algorithm is evaluated at the maximum (minimum) value of \( M_h \). This offset is added to \( M_t \) for any values that are OOB to obtain \( D_t \). In L19 we extended this definition by using either a user-specified number (i.e., NPT) of the largest or smallest values and averaging the differences to generate the offset.

Based on L19, for those methods subject to OOB failure (noted below, with one exception) here we employ both the traditional D07 approach with NPT = 1 (which we refer to as the “basic” approach) as well as our alternative simple tail adjustment (SIM) with NPT = 10. Additionally we utilize the limited tail adjustment scheme (LIM) introduced by L19 which is an extension of the SIM scheme. Unlike SIM, which is applicable only to methods subject to OOB conditions, and then only to values which are deemed OOB by the base algorithm, the LIM approach can be applied to any downscaling method with the intent on improving the representation of tail values. For LIM the user decides a priori how many values at the end of the tails are subject to LIM adjustment via the parameter tail-length (TLN). Then the TLN values at the end of a given tail are downscaled using the “constant correction” defined as the average of the NPT differences, \( D_t \) minus \( M_t \). In this study we apply LIM adjustment to all of our downscaling methods, and following L19 we use NPT = 10 and TLN = 10.

2.2.1 | BCQM

The most widely used SD method is quantile mapping (QM), devised by Panofsky and Brier (1968). Inspired by Ho et al. (2012), and following Lanzante et al. (2018) we refer to it as bias correction quantile mapping (BCQM) since the term “quantile mapping” is often used more generically in reference to a variety of SD techniques from the distributional class. Mathematically it can be expressed as:

\[
F_D(x) = F_O [F^{-1}_{MB}(x)]
\]  

where \( F \) represents the cumulative distribution function (CDF) and \( F^{-1} \) its inverse.

For each value to be downscaled BCQM determines its relative position in the \( M_h \) distribution (e.g., its percentile)
and then finds the value from the $O_h$ distribution corresponding to the same relative position and assigns this as the downscaled value. Since it employs only the distributions of $O_h$ and $M_h$ it follows the bias-correction paradigm. The OOB conditions occur when future values of $x$ (i.e., $M_t$ values) lay outside of the bounds of the $M_h$ distribution. In our basic implementation of BCQM we employ SIM with NPT = 1.

### 2.2.2 | BCQMa

Using a very simple modified scheme we can create an alternative version of BCQM that incorporates both the bias-correction and change-factor paradigms. This is accomplished by using anomaly inputs created by subtracting smooth daily climatological means of the GCM from the raw data. This climatology is created by applying a moving 7-day averaging window to the raw daily climatological means. Note that the mean of $M_h$ ($M_t$) is subtracted from values of both $O_h$ and $M_h$ (values of $M_t$). After transformation via quantile mapping, we add the means back to the result.

In Lanzante et al. (2018) it was demonstrated that in certain situations inclusion of the change-factor paradigm can have a large effect, either favourable or unfavourable. In some circumstances it is appropriate to subtract (and later add back) a trend rather than a mean in order to incorporate the change-factor paradigm. Here we use the anomaly approach since our future experiments were conducted as “time-slice” experiments characterized by constant external forcing and since we do not expect much of a trend over the historical period. For multi-decadal transient experiments, over which there is a large trend, de-trending would be more appropriate. Note that removal and subsequent replacement of a constant or time-varying mean (e.g., a trend) attempts to preserve changes in the dynamical model mean but not individual quantiles.

### 2.2.3 | PARM

Here we employ a modified version of the Asynchronous Regional Regression Model (ARRM) of Stoner et al. (2013) which we refer to as piecewise asynchronous regression model (PARM). Mechanistically ARRM operates similar to BCQM with one major exception. Downscaling via BCQM can be thought of as exploiting a quantile–quantile (qq) plot with $M_h$ ($O_h$) on the abscissa (ordinate). Since by definition the abscissa and ordinate values of a qq plot are ordered, one can enter an $M_t$ value to be downscaled on the abscissa and retrieve a $D_f$ value from the ordinate (interpolating between adjacent values if the $M_t$ value does not correspond exactly to an available value of $M_h$).

In the case of ARRM the relationship defined by the points on the qq plot are smoothed by fitting a series of piecewise continuous line segments (up to a maximum of seven) and using the values from these lines, instead of the original points, to perform the transformation. The intent is to smooth out small-scale noise due to random sampling variability. Like BCQM, ARRM employs the bias-correction paradigm.

Our PARM code is based on code kindly supplied by the creators of ARRM, with a few noteworthy differences. The set of adjustments described in section 2.5 of Stoner et al. (2013) that operate on the tails and extreme values were replaced by our SIM scheme with NPT = 1. The use of hard-coded knots in ARRM, specified to exist well outside the bounds of the range of the sample data was generalized in PARM. We removed some rarely-invoked checks for physically impossible values as well as internal data consistency. Finally, whereas ARRM uses an expanded training window (i.e., the month of interest plus half a month on either side), PARM uses data only from the calendar month of interest in its training step, for consistency with other methods in this study. Although this choice might not be optimal for real-world applications, it simplifies interpretation for our evaluations.

### 2.2.4 | PARMa

Using the same anomaly approach as used to create BCQMa from BCQM, we can create an alternate version of PARM known as PARMa. Like BCQMa, PARMa employs both the bias-correction and change-factor paradigms. Similar to BCQMa, removal and subsequent replacement of the mean (trend) attempts to preserve changes in the dynamical model mean but not individual quantiles.

### 2.2.5 | CDFt

The cumulative distribution function transform (CDFt) method of Michelangeli et al. (2009) was evaluated extensively regarding tail behaviour by L19. Here we employ the code in the CRAN repository (https://CRAN.R-project.org/package=CDFt) which is the source cited in the paper that introduced CDFt (Michelangeli et al., 2009). The distributional transformations involved in CDFt are more complex than the other methods introduced thus far, as discussed in more detail by L19. Mathematically CDFt can be expressed as:

\[
F_{Df}(x) = F_{Oh}(F^{-1}_{Mb}(F_{Mf}(x)))
\]

Because the CDFt method contains both a bias-correction mapping as well as anomaly normalization it employs both
the bias-correction and change-factor paradigms. Unlike the other methods above, CDFt has its own distinct way of handling out of bounds values. As discussed in more detail by L19, by default it borrows the end of the \( O_h \) distribution and appends it to the \( D_f \) distribution when values are out of bounds. Here we adopt this approach as the base CDFt algorithm. Note that due to its internal normalization it attempts to preserve changes in the dynamical model mean but not individual quantiles.

### 2.2.6 | QDM

The quantile delta mapping (QDM) method introduced by Cannon et al. (2015) was shown by those authors to be mathematically equivalent to the Equi-Distant Quantile Mapping (EDQM) method of Li et al. (2010). We use code supplied by the QDM authors (https://github.com/cran/MBC). Mathematically QDM can be expressed as:

\[
D_f(x) = M_f(x) + \left[ F^{-1}_{Oh} (F_{Mf}(M_f(x))) - F^{-1}_{Mh} (F_{Mf}(M_f(x))) \right],
\]

(3)

Since there are no OOB conditions the basic version of QDM does not employ SIM. Since it utilizes the \( O_h, M_h, M_f \) distributions, it operates via both the bias-correction and change-factor paradigms. Note that unlike the other methods that employ both paradigms, it preserves changes in individual quantiles but in general not changes in the mean.

### 2.2.7 | KDDM

The Kernel Density Distribution Mapping (KDDM) method (McGinnis et al., 2015) is philosophically similar to ARRM in that it aims to exploit a smooth version of the q-q relationship. However, instead of applying a smoothing to the raw values on a q-q plot as is the case for ARRM, it first smooths the \( O_h \) and \( M_h \) distributions using kernel density estimation (KDE) to generate probability density functions (PDFs). These PDFs are integrated to produce CDFs which are then used to create the q-q relationship. Note that prior to KDE the data are standardized to zero mean and unit variance based on means and variances estimated for each month of each year. This procedure has the effect of removing the climate change signal. After mapping, these processes are reversed. As is the case for QDM, KDDM is not subject to OOB conditions. KDDM operates via both the bias-correction and change-factor paradigms. Note that removal and subsequent replacement of the mean on a month-by-month basis attempts to preserve changes in the dynamical model mean but not individual quantiles. The code we employ is a modified version of that kindly supplied by the creators of the method (https://github.com/sethmcg/climod).

### 2.3 | Evaluation approach

Each downscaling method is applied separately for each gridpoint, each of 12 months, and each of three 10-year future ensembles. Following L19 we compute the mean absolute error (MAE) using the biweight mean (Lanzante, 1996) for future values of a given downscaling treatment \( (D_f) \) and for the corresponding raw GCM values \( (M_f) \). The biweight approach is a way to guard against the ill-effects of outliers by adapting to the data to act more like the arithmetic mean when data are “well-behaved” and more like the median otherwise. The MAE is averaged over all months and ensembles for \( D_f \) and \( M_f \) and then converted to a skill score (Wilks, 2006; L19). Finally, the skill is averaged over all non-ocean gridpoints.

Similar to L19, statistics are computed separately for each of nine distributional categories (CAT1-CAT9) and treatment. The distributional categories constitute portions of the tail of a given distribution, for example, CAT1 (CAT9) corresponds to the lowest (highest) value in the sample, with other categories defined as given in Table 1, which serves as a reference for our most common abbreviations. While CAT1-CAT4 (CAT6-CAT9) pertain to the left (right) tails, CAT5 includes all values in the distribution. Additionally, some results are given as averages over certain tail categories (weighted by the number of values in each category).

Statistics are presented by treatment, where a treatment consists of a combination of a downscaling approach (one SD method for either the base algorithm, SIM or LIM) and an evaluation criterion (EVAL). When we refer to a method without designation, or equivalently as the base algorithm, it should be noted that in some cases (as given above) the D07 tail scheme is used, which is equivalent to our SIM with NPT = 1. When we refer specifically to SIM it is implied that NPT = 10 and to LIM, NPT = 10 and TLN = 10. For simplicity we have fixed the values of the parameters based on the relative insensitivity and optimality of choices found in L19. In addition, we have three evaluation criteria designated numerically by 0 (not OOB), 1 (OOB), and 2 (all tail values). Note that sample sizes for different categories vary.

In order to test the significance of the difference in mean skill levels between two treatments we employ the same procedure as L19. Here we give a brief overview and the reader is referred to L19 (especially Appendix B) for a detailed account. For a given test, separate means are computed over maps for two distinct treatments. In the first step, we account for the fact that gridpoint values are not independent of one another pertaining to issues raised by Livezey and Chen (1983). In the first step we perform a Monte Carlo analysis...
in which a random translational shift of the gridpoint values is applied across the spatial domain for each trial. Based on the sampling distribution of pattern correlations between the original and translated maps we estimate an effective “block size” which is a spatial domain analogue of the temporal blocks in the bootstrap approach of Kunsch (1989). In the second step we randomly permute blocks of gridpoint values between the two treatments as per Preisendorfer and Barnett (1983) and use the resulting distribution of difference in mean skill between the two permuted maps to estimate a significance level. To account for uncertainties in the procedure, we report three significance levels based on conservative and liberal objective approaches as well as a very conservative subjective choice (L19).

### 3 | RESULTS

#### 3.1 | Overview of our approach

In L19 we examined in detail the performance of CDFt in the tails of the distribution. We considered a variety of configurations involving different tail treatments, both native to CDFt as well as our own creations. Here we use a similar approach to evaluate seven downscaling methods including CDFt but we limit the variety of alternate tail treatments based on lessons learned in L19. Here we compare results from three different treatments of tail values.

The first set of treatments involves the configuration employed most commonly in the literature which we refer to as the base or basic configuration. For those methods that employ the D07 tail scheme in their base state (which corresponds to our SIM with NPT = 1) we also examine as a second set of treatments, SIM with NPT = 10, as it was found in L19 that the latter yielded a moderate and statistically significant improvement over the former. On the other hand, increasing NPT beyond 10 (i.e., going from 10 to 20) yielded only a small, but insignificant improvement. The third set of treatments corresponds to our LIM approach with NPT = 10 and TLN = 10, again guided by results from L19’s CDFt experiments that suggested a relative insensitivity of results to the choice of parameter values.

| Shorthand in text | Description | Portion of distribution applicable to |
|-------------------|-------------|--------------------------------------|
| SIM               | Simple tail adjustment | Only OOB points |
| LIM               | Limited tail adjustment | TLN number of points in each tail |
| OOB               | Out of bounds. Points for which an SD method determines the base algorithm does not apply, hence a tail scheme is needed to produce output. | |
| NPT               | “lastN-points” (applicable to SIM and LIM.) The number of “good points” (i.e., those adjacent to the portion of a tail to be adjusted) averaged to determine the tail adjustment factor. | |
| TLN               | “Tail length” (applicable only to LIM.) Number of tail points to be adjusted regardless of whether the SD method considers them to be OOB or not. | |
| Eval criterion “0”| Evaluation computed using only non-OOB points | All non-OOB points |
| Eval criterion “1”| Evaluation computed using only OOB points | Only OOB points |
| Eval criterion “2”| Evaluation computed using all points (i.e., union of non-OOB “0” and OOB “1” points) | All points |
| CAT 1             | Minimum value at end of left tail | Left tail |
| CAT 2             | 2nd and 3rd smallest values | Left tail |
| CAT 3             | 4th, 5th, and 6th smallest values | Left tail |
| CAT 4             | 7th, 8th, 9th, and 10th smallest values | Left tail |
| CAT 5             | All points in the distribution | All points in the distribution |
| CAT 6             | 7th, 8th, 9th, and 10th highest values | Right tail |
| CAT 7             | 4nd, 5th, and 6th highest values | Right tail |
| CAT 8             | 2nd and 3rd highest values | Right tail |
| CAT 9             | Maximum value at end of right tail | Right tail |

1576  | Lanzante et al.
All of our downscaling methods have both a basic and LIM configuration. However, the methods for which the SIM approach is applicable are BCQM, BCQMa, PARM, and PARMa. SIM is not applicable for two of the methods (QDM and KDDM) which do not have special tail schemes since their methodologies do not have any OOB conditions. The CDFt method is a special case in that it has been applied in the literature using both its own unique tail scheme as well as that of D07. Here for the base configuration we employ the unique CDFt tail scheme, but as an alternative we employ SIM with NPT = 10.

### 3.2 Skill by distributional category

Figure 1 displays skill as a function of distributional category for the base treatment (solid curves) and LIM (dashed curves) with different downscaling methods represented by different colours. In both cases the evaluation criterion is 2, corresponding to both OOB and non-OOB values in the category. In the majority of cases there is not a large variation in skill with position within the tail. However, there is a tendency for those that do relatively well in the left tail to do relatively poorer in the right tail, and vice versa.

As an exercise we average MAE separately for the left and right tails (weighting by the number of values in each distributional category). For downscaled results we further average the MAE from the 4 best methods (KDDM, PARMa, BCQMa and QDM, see below). For downscaled (GCM) we get an average MAE of 0.86 and 0.95 (1.33 and 1.72), for the left and right tails, respectively. The resulting downscaling skill for the left and right tails is 35.4 and 44.5%, respectively. Thus, although downscaling error is larger in the right tail skill is greater there because GCM error is proportionately larger too. This asymmetry between tails is a consequence of the GCM forcing scenario and variable of interest (i.e., temperature).

Although there is a clustering of curves near the top of Figure 1, a few of the treatments stand out from the rest. In their base configurations both PARM and CDFt perform a bit worse in the left tail and much worse in the right tail than the other methods. The poorer performance of CDFt was detailed in L19. Note that because our PARM algorithm is a modified version of the ARRM, we cannot say with certainty that results for ARRM as they appear in the literature would behave similarly. BCQM also performs somewhat more poorly than the majority of methods. Note however that the anomaly approach (BCQMa and PARMa), which incorporates the change-factor paradigm, yields substantial improvement (over BCQM and PARM, respectively) at least partially due to the “coastal effect” (Dixon et al., 2016; Dixon et al., 2018) as seen in maps of skill (not shown). However, the substantial improvement of PARMa over PARM in the right tail is widespread without any tilt towards coastal regions.

For the best performing methods LIM yields a very slight improvement over the base approach in the right tail and a modest improvement in the left tail. For CDFt LIM also yields a noticeable improvement as shown by L19. In the right tail LIM results in a substantial improvement for PARM but a negligible degradation for BCQM. In the left tail the difference between LIM and the base algorithm is not worth noting given the extremely small sample size (see Figure 2).

We do not show skill by distributional category for the other two evaluation criteria (OOB and non-OOB) but instead, in the next section show category average skill. On the one hand skill by category for non-OOB conditions is in most cases very similar to that for all cases as shown in...
Figure 1 since the special tail schemes are usually invoked relatively infrequently (Figure 2). On the other hand, in a minority of cases the tail schemes are invoked much of the time (for BCQM, PARM in the right tail, and CDFt at the very ends of both tails) so that overall results (Figure 1) are similar to that for OOB.

3.3 | Skill averaged over distributional categories

In this section, we compare the different downscaling methods by way of skill averaged over tail categories for several different scenarios (Figure 3). We consider the base algorithm (B), SIM (S) and LIM (L) tail schemes and report skill averaged over the left and right tail values, weighted by the number of values in each distributional category. In addition to a particular tail treatment, each scenario is associated with an evaluation criterion (0, 1, or 2 as defined above). For completeness we also include one scenario (A2) based on all values in the distribution (corresponding to CAT 5) using the base algorithm. For a few cases we plot additional averages based on excluding some outlying values. For CDFt we eliminate CAT 9 which was shown to be highly problematic by L19. For BCQM and PARM we eliminate CAT 1–4 since the sample sizes are extremely small (see Figure 2), leading to unreliable results. In conjunction with the average skills we utilize results from a limited number of tests of statistical significance (Table 2). The significance testing was conducted as outlined in Section 2.3 and discussed in greater detail in L19.

In Figure 3 average skill values are plotted as circles with each colour representing a different downscaling method. The first column (B2) shows results for the basic algorithm for all tail values (i.e., both OOB and non-OOB). Four of the methods yield very similar results (KDDM, PARMa, BCQMa and QDM), with BCQM (CDFt and PARM) yielding somewhat (much) poorer results. By excluding the problematic CAT 9 (open circle) results for CDFt are considerably better, although still worse than BCQM.

To ascertain the degree to which differences in average skill levels for B2 in Figure 3 are meaningful we can consult with Table 2 containing significance levels for tests of...
| Nhlon x Nblat | Probability (%) | Treatments | Categories |
|---------------|-----------------|------------|------------|
| 4 x 2 5 x 4 11 x 6 | 42.1 31.4 9.6 | K2 x Q2 | 1-4 |
| 4 x 2 14 x 7 19 x 12 | 78.6 54.7 41.8 | K2 x Q2 | 6-9 |
| 4 x 2 6 x 3 10 x 5 | 4.4 1.3 0.0 | K2 x B2 | 1-4 and 6-9 |
| 4 x 2 5 x 2 8 x 4 | 0.7 0.2 0.0 | K2 x C2 | 1-4 and 6-8 |
| 4 x 2 8 x 4 12 x 7 | 74.4 71.2 56.0 | Pa0 x Ba0 | 1-4 and 6-8 |
| 4 x 2 4 x 3 7 x 5 | 1.0 0.5 0.0 | Pa0 x B0 | 1-4 and 6-8 |
| 4 x 2 6 x 3 10 x 5 | 1.8 0.8 0.0 | Pa0 x B0 | 1-4 and 6-9 |
| 4 x 2 14 x 9 23 x 14 | 69.1 42.8 26.7 | B1 x Ba1 | 6-9 |
| 4 x 2 14 x 9 23 x 3 | 7.0 0.0 0.0 | B1 x Pa1 | 6-9 |
| 4 x 2 13 x 7 21 x 12 | 0.0 0.0 0.0 | B1 x C1 | 6-8 |
| 4 x 2 3 x 2 7 x 4 | 0.0 0.0 0.0 | Pa0 x Pa1 | 1-4 and 6-9 |
| 4 x 2 4 x 2 7 x 5 | 0.0 0.0 0.0 | Ba0 x Ba1 | 1-4 and 6-9 |
| 4 x 2 8 x 4 13 x 7 | 0.0 0.0 0.0 | B0 x B1 | 6-9 |
| 4 x 2 4 x 3 10 x 6 | 0.0 0.0 0.0 | C0 x C1 | 1-4 and 6-8 |
| 4 x 2 2 x 1 4 x 3 | 0.0 5.9 0.0 | P0 x P1 | 6-9 |
| 4 x 2 7 x 5 12 x 8 | 87.7 83.5 75.0 | QL2 x PaL2 | 1-4 and 6-9 |
| 4 x 2 8 x 4 13 x 7 | 34.9 20.1 5.4 | QL2 x CL2 | 1-4 and 6-9 |
| 4 x 2 4 x 2 7 x 4 | 0.2 0.2 0.0 | QL2 x BL2 | 1-4 and 6-9 |
| 4 x 2 2 x 1 4 x 3 | 0.0 13.3 0.0 | QL2 x PL2 | 1-4 and 6-9 |
| 4 x 2 6 x 3 10 x 6 | 58.7 51.6 28.2 | K2 x KL2 | 1-4 |
| 4 x 2 14 x 8 22 x 12 | 75.0 45.9 33.5 | K2 x KL2 | 6-9 |
| 4 x 2 6 x 3 10 x 6 | 58.9 51.3 28.1 | Pa2 x PaL2 | 1-4 |
| 4 x 2 12 x 8 19 x 12 | 58.9 21.8 9.6 | Pa2 x Bal2 | 6-9 |
| 4 x 2 5 x 3 9 x 6 | 45.0 40.0 12.5 | Ba2 x Bal2 | 1-4 |
| 4 x 2 12 x 7 17 x 11 | 64.1 30.0 18.0 | Ba2 x Bal2 | 6-9 |
| 4 x 2 5 x 4 11 x 6 | 24.0 14.5 1.3 | Q2 x QL2 | 1-4 |
| 4 x 2 11 x 7 17 x 10 | 48.1 16.3 6.0 | Q2 x QL2 | 6-9 |
| 4 x 2 7 x 4 13 x 8 | 36.7 16.0 1.4 | B2 x BL2 | 1-4 |
| 4 x 2 10 x 6 16 x 10 | 81.7 72.9 62.5 | B2 x BL2 | 6-9 |
| 4 x 2 5 x 3 8 x 5 | 2.6 1.1 0.0 | C2 x CL2 | 1-4 and 6-8 |
| 4 x 2 7 x 4 13 x 8 | 36.0 14.3 1.3 | P2 x PL2 | 1-4 |
| 4 x 2 12 x 6 19 x 11 | 0.0 0.0 0.0 | P2 x PL2 | 6-9 |
| 4 x 2 6 x 4 11 x 6 | 81.7 73.4 61.9 | Pa2 x Q2 | 5 |
| 4 x 2 4 x 4 11 x 6 | 61.6 50.8 22.1 | Pa2 x B2 | 5 |
| 4 x 2 4 x 2 6 x 4 | 1.1 1.1 0.0 | Pa2 x B2 | 5 |
| 4 x 2 3 x 2 6 x 4 | 0.1 1.1 0.0 | Pa2 x P2 | 5 |

Notes: For each comparison there are three probabilities (left to right) corresponding to three effective grid sizes. The grid sizes (Nhlon x Nblat) are expressed as the number of blocks in the longitudinal and latitudinal dimensions. The leftmost is a subjectively determined grid dimension and is almost certainly too small. The centre (right) grid dimensions are based on more conservative (liberal) criteria derived from Monte Carlo analysis (see Section 2.3). Treatments consist of a downscaling method, optionally a tail treatment (basic if not specified), and an evaluation criterion. Downscaling methods: B (BCQM), Ba (BCQMA), P (PARM), Pa (PARMa), C (CDFt), Q (QDM), and K (KDDM). Tail treatments: S (SIM), L (LIM), or basic if not specified. Evaluation criteria: 0 (no-OOB), 1 (OOB), and 2 (no-OOB + OOB). Categories as defined in Table 1. Abbreviations: BCQM, bias correction quantile mapping; CDFt, cumulative distribution function transform; KDDM, Kernel density distribution mapping; LIM, limited tail adjustment; OOB, out-of-bounds; PARM, piecewise asynchronous regression model; QDM, quantile delta mapping; SIM, simple tail adjustment.
difference between pairs of average skill. We limit the number of paired comparisons to a manageable number. For example, the first row has significance levels of 42.1, 31.4, and 9.6% for the difference between KDDM and QDM averaged over the categories in the left tail (1–4). From the second row a similar test for the categories in the right tail (6–9) yields values of 78.6, 54.7, and 41.8%. Before declaring a measure of significance we require robust results from the three separate estimates. Thus, in this case we conclude non-significance for both tails. Separate tests for left and right tails were motivated by the larger separation between KDDM and QDM shown in Figure 1 for the left as compared to the right tail.

Continuing with rows three and four in Table 2 we find a significant difference between KDDM (black) and both BCQM (violet) and CDFt (brown). Although we have not tested for the difference between KDDM and PARM (green) we anticipate this is highly significant since the distance between these two points is much larger than between KDDM and CDFt which was highly significant. Furthermore, even eliminating the problematic CAT 9 value for CDFt (brown, open circle) the separation is greater than for BCQM (violet) which was significantly different. In summary, we have found that comparing performance of the basic tail schemes for all methods, four of them are indistinguishable (KDDM, PARMa, BCQMa and QDM) and these were significantly better than the remaining three methods (BCQM, CDFt and PARM).

Next we examine the performance of methods for tail values for non-OOB conditions (B0 in Figure 3). This is a test of the basic algorithm’s ability to downscale tail values rather than that of the specialized tail schemes. Note that KDDM and QDM are excluded from this exercise since they have no OOB conditions. In this case PARMa and BCQMa are best, yielding statistically indistinguishable results. BCQM, and by inference PARM and CDFt perform significantly worse. Note that even if we eliminate the troublesome CAT 9 value, CDFt is significantly worse.

The third column in Figure 3 (B1) displays results for skill in downscaling tail values using the special tail schemes for OOB values. The ranking of methods is fairly similar to that for B0, with BCQMa and PARMa the best, with BCQM, CDFt and PARM performing much more poorly. However, eliminating the sparsely sampled left tail (open circles, Figure 2) results in an improvement for the three poorer performing methods, especially BCQM. Perhaps the most important point to be made is that comparing B0 and B1 skill for a given method is significantly better for B0, with the magnitude of the differences in skill ~20–60. This indicates that when values are OOB the specialized tail schemes do not perform very well.

In L19 it was found that for CDFt a simple modification of the D07 scheme yielded considerable improvement over either the internal (basic) tail scheme or the D07 scheme. Here we apply our modified D07 scheme (i.e., SIM with NPT = 10) to the five methods that experience OOB conditions. The results shown in Figure 3 (S1) yield a ranking of methods fairly consistent with prior results in which PARMa and BCQMa perform the best. However, for all methods the skill is considerably greater for S1 than B1 and highly significantly so in almost all cases. Nevertheless, even with our improved approach the results are worse than for B0 in all cases except one (CDFt), demonstrating the difficulty in downscaling OOB values.

In L19 we introduced another modified scheme (LIM) which can be applied to all tail values, not just those deemed OOB by a particular method. These values (L2) are directly comparable to those based on the basic Scheme (B2). The ordering of results is, with one exception (CDFt), fairly similar between L2 and B2. The four best methods (QDM, KDDM, BCQMa and PARMa) have almost identical levels of skill that are statistically indistinguishable from one another. Although these four methods each have slightly higher skill for L2 than B2, the differences are statistically indistinguishable. Two of the poorer performing methods (CDFt and PARMa) have significantly higher skill for L2 than B2 while the other (BCQM) has insignificantly lower skill. However, the improvement in CDFt is sufficiently large that its skill, although less than the cluster of the four best methods, is no longer significantly less.

Our final comparison of average skill levels in Figure 3 is based on all values in the distribution (A2, which is the same as CAT 5 in Figure 1) vs. just the tail values (B2). The ordering is very similar between the two with the exception of CDFt which for A2 resides in the cluster of best methods. Taken in conjunction with earlier results this indicates that CDFt does well outside of the tails, but is inferior for tail values regardless of OOB conditions or not.

### 3.4 Average error and skill by month

Up to this point the analyses conducted are similar to those performed by L19. Next we extend our analyses in order to examine seasonality. Figure 4 shows MAE by month for each downscaling method as well as the GCM. We include the GCM because skill (shown later) is derived from MAE for a given method in reference to that from the GCM. While the dashed lines indicate MAE for the whole distributions (CAT 5), solid lines are for just the tails.

Figure 4 shows that MAE varies widely between methods following the general ordering seen earlier in Figure 3. Similarly, MAE is larger for the tails than the whole distribution, with one exception (BCQM), consistent
with seasonal cycles of MAE with larger values during a broad plateau from April to September indicating that during this time area-averages (GCM) are less representative than point values (observations). This plateau can be explained largely by two factors: the “mountain snow effect” in spring and the “coastal effect” during summer (Lanzante et al., 2018). Both of these effects result from the larger footprint of the GCM compared to the observations. For the former there is a blending of snow-covered higher elevation sites with those from lower snow-free locations. For the latter, there is a blending of land and water-covered surfaces over land near boundaries with bodies of water.

For downscaling, the tails generally have larger MAE than for CAT 5, as expected, although this difference is predominantly negligible during spring because of the mountain snow effect. In the presence of a warming signal the most challenging values to downscale occur in the right tail in the form of “novel” values. The downscaling “mountain snow effect” occurs when snow is present in the historical training period but absent in the future application period (Lanzante et al., 2018). But since right tail values occur mostly during snow-free conditions in both the training sample as well as in the future this mitigates the “mountain snow effect,” thus degrading tail values much less compared to the rest of the distribution than would normally be expected. On the other hand, much of the remainder of the distribution is subject to the “mountain snow effect.”

The seasonality of downscaling skill is shown in Figure 5, which for display purposes omits curves for the poorer performing methods (PARM, BCQM, and CDFt). There is a marked difference between CAT 5 and the tails. For CAT 5 the downscaling spring maximum in MAE (Figure 4) is consistent with the left portion of the GCM plateau whereas the pronounced drop in downscaling MAE in summer contrasts with the right portion of the GCM plateau. This leads to a prominent CAT 5 peak in skill during the summer.

On the other hand, for the tails, the greater similarity in the shapes of the MAE curves between downscaling and GCM leads to much less of a seasonal difference in skill (~7% vs. 17%) between maximum and minimum. However, more subtle changes result in a general difference in skill between the first and second halves of the year due to the fact that the decline in MAE after the summer peak for downscaling leads that for the GCM. This decline is associated with the good downscaling performance in coastal regions, particularly along the west coast (not shown).

### 3.5 Rendering of the climate change signal

One additional evaluation performed here that was not performed by L19 regards the rendering of the climate change signal (CCS) which we examine in two distinct and complementary ways. First we examine the CCS as the temperature,
future minus historical periods, averaged over the spatial domain (Figure 6). Then we examine the pattern correlation of the spatial patterns of temperature change over the domain between each treatment and that for $O_f$ (Figure 7). Note that due to the different nature of these two metrics there is no a priori expectation that they will produce similar results.

Although we have assessed the significance of differences in CCS between treatments in a similar fashion as earlier we do not show the results since very few of the tests unambiguously indicate a significant difference. Lack of significance is due in large part to the small degrees of freedom in the CCS whose values vary much less from location to location than for MAE which is a much noisier field. Furthermore, for pattern correlations the significance levels are much more sensitive to the different estimates of degrees of freedom; lacking a clear result we are hesitant to declare significance. Therefore, instead of reporting all of the significance levels we make some limited, more qualitative statements.

Figure 6 shows the average CCS for the left and right tails as well as the entire distribution. For the left tail the three methods found earlier to perform most poorly (BCQM, PARM and CDFt) also perform most poorly by overestimating the warming by $-0.5$ to $-0.6$ C. The four best methods (QDM, PARMa, KDDM and BCQMa) do quite well with an average error of $0.06$ C or less, improving on the GCM which underestimates the warming $-0.13$ C. For the right tail the same four best methods plus CDFt underestimate the warming by $-0.1$ C while BCQM and PARM overestimate the warming, the latter by $-0.9$ C. Most of the methods fail to improve upon the GCM which overestimates the warming $-0.07$ C. For the entire distribution (CAT 5) BCQM and PARM overestimate the warming $-0.6$ to $-0.65$ C while the remaining methods and the GCM underestimate the warming $-0.06$ C. The only treatment with a statistically significant difference (0.01 level) with the observations is PARM for the right tail. The next most significant result ($-0.10$ level) is that of PARM for CAT 5.

The results based on the pattern correlation of the CCS (Figure 7) are similar to those based on the average CCS. PARM is the worst followed by BCQM and CDFt. The remaining methods and the GCM yield similar results. Regarding significance, for each of the three panels the greatest outliers (one or two treatments) are ambiguously significant at the 0.05 level; for these only one of the estimates of degrees of freedom yield significance.

Earlier studies showed that quantile mapping can potentially distort GCM trends in the mean (Hagemann et al., 2011; Themeßl et al., 2012; Maraun, 2013; Maurer and Pierce, 2014; Gobiet et al., 2015; Pierce et al., 2015) as well as in the quantiles (Cannon et al., 2015). However, lacking a future “truth” they were not able to determine the extent of

![Figure 6](image1.png)  
**Figure 6** Climate change signal (CCS) (C) by treatment for the basic scenario as a weighted average over the left tail (CAT 1–4), right tail (CAT 6–9) or the entire distribution (CAT 5). The CCS is the difference in temperature, future minus historical ($D_f-D_h$), averaged over the spatial domain. Treatments are represented by filled circles with colours as in Figure 1. In addition, the black asterisk is for the GCM ($M_f-M_h$) and the yellow square for observations ($O_f-O_h$), the latter representing the “truth.”

![Figure 7](image2.png)  
**Figure 7** Similar to Figure 6 except for the pattern correlation between the map of the climate change signal (future minus historical temperature) for each treatment ($D_f-D_h$) and the “truth” ($O_f-O_h$). Separate pattern correlations for each month and ensemble are averaged using the fisher-Z transformation (Wilks, 2006).
the distortion in a realistic setting. Furthermore, altering the GCM trends could be good if bias is removed or bad if the “true” trend is distorted – there are competing effects and it is not clear which way things should go. Here, using a PM experimental design that enables analysis of future projections, we show a general consistency with earlier studies in that BCQM and PARM, both of which operate using the bias-correction paradigm have deleterious effects on the CCS for the tails as well as overall (i.e., for the mean). In contrast, QDM, KDDM, CDFt or anomalies (BCQM and PARM) greatly reduce or eliminate the distortion of trends because of the additional change-factor component. Cannon et al. (2015) explained how QDM maintains the climate change signal at all quantiles whereas another approach (which is equivalent to BCQM) does not. This could help explain why for CAT 5 BCQM and QDM perform almost identically in reproducing the climate change signal whereas QDM is slightly better than BCQM for the tails.

4 | DISCUSSION AND CONCLUSIONS

We have compared the performance of seven distributional-type statistical downscaling methods in a Perfect Model experimental design for a high-emissions future scenario. We use an approach similar to that employed by L19 in which we pay particular attention to performance in the tails of the distribution in addition to overall performance. In spite of the very different algorithmic approaches of the various downscaling methods that we have examined, we find minimal difference in performance for the majority of methods and metrics when considering the whole distribution. The two methods (BCQM and PARM) based on the bias-correction paradigm performed consistently worse than the rest that also utilized the change-factor paradigm. One exception to this was the generally poorer performance of the CDFt method in the tails of the distribution which previously (L19) we determined to have pathological problems. We also found that special schemes that we devised earlier (L19) for application to CDFt yielded benefits to the handling of tail values for the other methods as well. Our modifications are an extension of the widely used D07 approach aimed at handling values that are out of normal operating bounds due to model-simulated climate change.

In examining the seasonality of performance of the downscaling methods we found generally larger errors (MAE) in the GCM as well as the downscaling methods during the warm season. This seasonality is largely due to factors identified previously (Lanzante et al., 2018) that are linked to spatial resolution differences between the GCM (coarse) and observations (fine). The spring skill minimum, referred to as the “mountain snow” effect, is related to a non-stationarity arising from the disappearance of snow during spring in a warmed climate, and is a particular challenge for methods that utilize a change-factor approach. Bias-correction methods that do not include a change-factor element are particularly susceptible to the summertime “coastal” effect, resulting from a very large number of OOB cases.

We have also evaluated the ability of the downscaling methods to reproduce the climate change signal. We found that the methods whose algorithms explicitly incorporate a change-factor signal out-perform those that incorporate only a bias-correction derived from the historical period. Furthermore, methods that perform more poorly regarding the temperature change averaged over all gridpoints also tend to perform more poorly in terms of the pattern correlation. Not only do those methods distort the strength of the signal (common to all gridpoints) but they also produce distortions that vary from gridpoint to gridpoint.

We find little meaningful difference in performance between most of the downscaling methods and across several fundamentally different evaluation metrics. At first thought this may be surprising given the substantial differences in the nature and complexity of the underlying downscaling algorithms. But perhaps this should not be surprising given that there is no rigorous theoretical reason for any particular method to be fundamentally better than the others. However, we find one philosophical choice that separates the methods into two classes of poorer versus better overall performance within the Perfect Model experimental design. The former are based on the bias-correction paradigm whereas the latter incorporates both the change-factor and bias-correction paradigms. Nevertheless, there may be more limited instances in which the generally poorer performing paradigm is superior (Lanzante et al., 2018).

Based on code provided to us by the developers of the methods PARM (our slightly simplified version of ARRM) and KDDM stand out as by far the most complex, as they invest much effort in smoothing the q-q relationship between Oh and Mf. On the other hand, BCQM and QDM, whose implementation is considerably simpler offer comparable overall performance. Although implementation of CDFt is relatively simple as well, it suffers from certain pathologies (L19).

We would be remiss in not caveating our findings. We use perhaps the simplest possible Perfect Model design in which the mismatch between observations and model is limited to spatial resolution. In more realistic settings a mismatch between the climatologies is both possible and likely. However, we believe our general findings would hold in that the mismatch in spatial resolution manifests itself as a mismatch in distributional characteristics, as would a mismatch in climatologies. In future work we intend to utilize GCM-
RCM pairings yielding a more realistic Perfect Model design.

Another limitation of our work is the analysis of only daily maximum temperature over the CONUS. As a general rule we believe that the challenges we have identified in our body of work to date apply to less well-behaved variables such as precipitation and other water-vapour related quantities. However, such variables may present additional challenges, for example in the form of less symmetry and/or discontinuous and bounded distributions. Work in progress is examining daily precipitation over the CONUS. Furthermore, climate regimes unlike those over the CONUS may have challenges we have yet to discover. Thus, we cannot rule out the possibility that some of the more complex methods that we have examined may have advantages yet to be highlighted when dealing in these other settings.

ACKNOWLEDGEMENT

We thank Adrienne Wootten and Andrew Ross for comments on an earlier draft of this article, as well as Katherine Hayhoe and Anne Stoner for the ARRM code, and Seth McGinnis for the KDDM code.

ORCID

John R. Lanzante https://orcid.org/0000-0002-1736-7170
Keith W. Dixon https://orcid.org/0000-0003-3044-326X

REFERENCES

Abatzoglou, J.T. and Brown, T.J. (2012) A comparison of statistical downscaling methods suited for wildfire applications. International Journal of Climatology, 32(5), 772–780. https://doi.org/10.1002/joc.2312.

Cannon, A.J., Sobie, S.R. and Murdock, T.Q. (2015) Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? Journal of Climate, 28(17), 6938–6959. https://doi.org/10.1175/JCLI-D-14-00754.1.

Deque, M. (2007) Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: model results and statistical correction according to observed values. Global and Planetary Change, 57(1–2), 16–26. https://doi.org/10.1016/j.gloplacha.2006.11.030.

Dixon, K.W., Lanzante, J.R., Nath, M.J., Hayhoe, K., Stoner, A., Radhakrishnan, A., Balaji, V. and Gaitan, C.F. (2016) Evaluating the stationarity assumption in statistically downscaled climate projections: is past performance an indicator of future results? Climatic Change, 135(3–4), 395–408. https://doi.org/10.1007/s10584-016-1598-0.

Fowler, H., Blenkinsop, S. and Tebaldi, C. (2007) Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. International Journal of Climatology, 27(12), 1547–1578. https://doi.org/10.1002/joc.1556.

Gobiet, A., Suklitsch, M. and Heinrich, G. (2015) The effect of empirical-statistical correction of intensity-dependent model errors on the temperature climate change signal. Hydrology and Earth System Sciences, 19(10), 4055–4066. https://doi.org/10.5194/hess-19-4055-2015.

Gutierrez, J.M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., Roessler, O., Wibig, J., Wilcke, R., Kotlarski, S., San Martin, D., Herrera, S., Bedia, J., Casanueva, A., Manzanas, R., Ithurbi, M., Vrac, M., Dubrovsky, M., Ribalaygua, J., Portoles, J., Raty, O., Raisanen, J., Hingray, B., Raynaud, D., Casado, M.J., Ramos, P., Zerenner, T., Turco, M., Bossard, T., Stepanek, P., Bartholy, J., Pongracz, R., Keller, D.E., Fischer, A.M., Cardoso, R.M., Soares, P.M.M., Czerniecki, B. and Page, C. (2018) An intercomparison of a large ensemble of statistical downscaling methods over Europe: results from the VALUE perfect predictor cross-validation experiment. International Journal of Climatology, 39, 3750–3785. https://doi.org/10.1002/joc.5462.

Hagemann, S., Chen, C., Haerter, J.O., Heinke, J., Gerten, D. and Piani, C. (2011) Impact of a statistical bias correction on the projected hydrological changes obtained from three GCMs and two hydrology models. Journal of Hydrometeorology, 12(4), 556–578. https://doi.org/10.1175/2011JHM1336.1.

Ho, C.-K., Stephenson, D.B., Collins, M., Ferro, C.A.T. and Brown, S. J. (2012) Calibration strategies: a source of additional uncertainty in climate change projections. Bulletin of the American Meteorological Society, 93(1), 21–26. https://doi.org/10.1175/2011BAMS3110.1.

Kunsch, H. (1989) The jackknife and the bootstrap for general stationary observations. The Annals of Statistics, 17(3), 1217–1241. https://doi.org/10.1214/aos/1176347265.

Lanzante, J.R. (1996) Resistant, robust & non-parametric techniques for the analysis of climate data: theory and examples, including applications to historical radiosonde station data. International Journal of Climatology, 16(11), 1197–1226. https://doi.org/10.1002/(SICI)1097-0088(199611)16:11<1217::AID-JOC91>3.0.CO;2-L.

Lanzante, J.R., Dixon, K.W., Nath, M.J., Whitlock, C.E. and Adams-Smith, D. (2018) Some pitfalls in statistical downscaling of future climate. Bulletin of the American Meteorological Society, 99(4), 791–803. https://doi.org/10.1175/BAMS-D-17-0046.1.

Lanzante, J.R., Nath, M.J., Whitlock, C.E., Dixon, K.W. and Adams-Smith, D. (2019) Evaluation and improvement of tail behaviour in the cumulative distribution function transform downsampling method. International Journal of Climatology, 39(4), 2449–2460. https://doi.org/10.1002/joc.5964.

Li, H., Sheffield, J. and Wood, E.F. (2010) Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. Journal of Geophysical Research, 115(D10101), 1–20. https://doi.org/10.1029/2009JD012882.

Livezey, R. and Chen, W. (1983) Statistical field significance and its determination by Monte Carlo techniques. Monthly Weather Review, 111(1), 46–59. https://doi.org/10.1175/1520-0493(1983)111<0.3006:SFSAID3.0.CO;2.

Maraun, D. (2013) Bias correction, quantile mapping, and downscaling: revisiting the inflation issue. Journal of Climate, 26(6), 2137–2143. https://doi.org/10.1175/JCLI-D-12-00821.1.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themessl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., & Thiele-Eich, I. (2010). Precipitation downscaling under climate change: recent developments to
bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3), RG3003, 1–34, https://doi.org/10.1029/2009RG000314

Maraun, D., Widmann, M., Gutiérrez, J.M., Kotlarski, S., Chandler, R. E., Hertig, E., Wibig, J., Huth, R. and Wilcke, R.A.I. (2015) VALUE: a framework to validate downscaling approaches for climate change studies. *Earth’s Future*, 3, 1–14. https://doi.org/10.1002/2014EF000259.

Maurer, E.P., Hidalgo, H.G. and Das, T. (2010) The utility of daily large-scale climate data in the assessment of climate change impacts on daily stream-flow in California. *Hydrology and Earth System Sciences*, 14 (6), 1125–1138. https://doi.org/10.5194/hess-14-1125-2010.

Maurer, E.P. and Pierce, D.W. (2014) Bias correction can modify climate model simulated precipitation changes without adverse effect on the ensemble mean. *Hydrology and Earth System Sciences*, 18 (3), 915–925. https://doi.org/10.5194/hess-18-915-2014.

McGinnis, S., Nychka, D. and Mearns, L. (2015) A new distribution mapping technique for climate model bias correction. In: Lakshmanan, V., Gilceland, E., McGovern, A. and Tingley, M. (Eds.) *Machine Learning and Data Mining Approaches to Climate Science*. Cham: Springer. https://doi.org/10.1007/978-3-319-17220-0_9.

Michelangeli, P.A., Vrac, M. and Loukos, H. (2009) Probabilistic downscaling approaches: application to wind cumulative distribution functions. *Geophysical Research Letters*, 36, L11708. https://doi.org/10.1029/2009GL038401.

Panofsky, H.W. and Brier, G.W. (1968) *Some Applications of Statistics to Meteorology*. Pennsylvania: Pennsylvania State University Press, p. 224.

Pierce, D.W., Cayan, D.R., Maurer, E.P., Abatzoglou, J.T. and Hegewisch, K.C. (2015) Improved bias correction techniques for hydrological simulations of climate change. *Journal of Hydrometeorology*, 15(6), 2421–2443. https://doi.org/10.1175/JHM-D-14-0236.1.

Pierce, D.W., Cayan, D.R. and Thrasher, B.L. (2014) Statistical downscaling using localized constructed analogs (LOCA). *Journal of Hydrometeorology*, 15(6), 2558–2585. https://doi.org/10.1175/JHM-D-14-0082.1.

Preisendorfer, R. and Barnett, T. (1983) Numerical model-reality intercomparison tests using small-sample statistics. *Journal of the Atmospheric Sciences*, 40(8), 1884–1896. https://doi.org/10.1175/1520-0469(1983)040%3C1884:NMRITU%3E2.0.CO;2.

Stoner, A.M.K., Hayhoe, K., Yang, X. and Wuebbles, D.J. (2013) An asynchronous regional regression model for statistical downscaling of daily climate variables. *International Journal of Climatology*, 33 (11), 2473–2494. https://doi.org/10.1002/joc.3603.

Wilks, D.S. (2006) *Statistical Methods in the Atmospheric Sciences*, 2nd edition. San Diego, CA: Academic Press.

Wood, A.W., Leung, L.R., Sridhar, V. and Lettenmaier, D.P. (2004) Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, 62(1–3), 189–216. https://doi.org/10.1023/B:CLIM.0000013685.99609.9e.

How to cite this article: Lanzante JR, Adams-Smith D, Dixon KW, Nath M, Whitlock CE. Evaluation of some distributional downscaling methods as applied to daily maximum temperature with emphasis on extremes. *Int J Climatol.*, 2020;40:1571–1585. https://doi.org/10.1002/joc.6288