Received: 15 May 2020 | Revised: 12 January 2021 | Accepted: 12 January 2021 | Published on: 1 February 2021

DOI: 10.1002/joc.7013

RESEARCH ARTICLE

International Journal of Climatology RMetS

Evaluation of some distributional downscaling methods as applied to daily precipitation with an eye towards extremes

John R. Lanzante[1] | Keith W. Dixon[1] | Dennis Adams-Smith[1,2] | Mary Jo Nath[1] | Carolyn E. Whitlock[1,3]

[1]National Oceanic and Atmospheric Administration (NOAA), Geophysical Fluid Dynamics Laboratory (GFDL), Princeton, New Jersey

[2]University Corporation for Atmospheric Research (UCAR), Cooperative Programs for the Advancement of Earth System Science (CPAESS), Boulder, Colorado

[3]SAIC, Reston, Virginia

Correspondence
John R. Lanzante, National Oceanic and Atmospheric Administration, Geophysical Fluid Dynamics Laboratory, 201 Forrestal Road, Princeton, NJ 08542.
Email: john.lanzante@noaa.gov

Abstract

Statistical downscaling (SD) methods used to refine future climate change projections produced by physical models have been applied to a variety of variables. We evaluate four empirical distributional type SD methods as applied to daily precipitation, which because of its binary nature (wet vs. dry days) and tendency for a long right tail presents a special challenge. Using data over the Continental U.S. we use a 'Perfect Model' approach in which data from a large-scale dynamical model is used as a proxy for both observations and model output. This experimental design allows for an assessment of expected performance of SD methods in a future high-emissions climate-change scenario. We find performance is tied much more to configuration options rather than choice of SD method. In particular, proper handling of dry days (i.e., those with zero precipitation) is crucial to success. Although SD skill in reproducing day-to-day variability is modest (~15–25%), about half that found for temperature in our earlier work, skill is much greater with regards to reproducing the statistical distribution of precipitation (~50–60%). This disparity is the result of the stochastic nature of precipitation as pointed out by other authors. Distributional skill in the tails is lower overall (~30–35%), although in some regions and seasons it is small to non-existent. Even when SD skill in the tails is reasonably good, in some instances, particularly in the southeastern United States during summer, absolute daily errors at some gridpoints can be large (~20 mm or more), highlighting the challenges in projecting future extremes.

KEYWORDS

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

1 | INTRODUCTION

Globally the effects of climate change on a variety of physical variables have been well documented (IPCC, 2013). Physical models in the form of global climate models (GCMs) and regional climate models (RCMs) are primary tools used in projecting climate change. Although both temperature and precipitation are

changing there are some fundamental differences. While temperature is eventually expected to rise everywhere the sign of the precipitation response varies by location, season, and between models. However, different RCM forced by the same GCM can yield opposite signed precipitation responses (Karmalkar, 2018; Holtanova *et al*., 2019). Hence, projecting changes in precipitation is particularly challenging, especially at the regional scale.

Variations in temperature and precipitation differ fundamentally in that temperature varies more smoothly both spatially and temporally. Precipitation is often discontinuous in both space and time. Accordingly, the distribution of precipitation is often highly asymmetrical and skewed whereas temperature is usually more well-behaved. Additionally, precipitation is characterized by two aspects: the (binary) occurrence and the actual distribution of amount on wet days, which makes statistical modelling of precipitation much more difficult.

Of particular relevance is the historical and projected increase in extreme precipitation events (IPCC, 2013). Compounding this is the disproportionate contribution of rare heavy events (Pendergrass and Knutti, 2018) such that globally—the wettest 12 days typically account for approximately half the annual total, with this concentration projected to increase in the future.

To mitigate the deficiencies of physical climate models and provide information for policymakers better suited for local areas, a wide variety of statistical downscaling (SD) techniques have been developed (Maraun and Widmann, 2018). Recently, this author team, members of the Geophysical Fluid Dynamics Laboratory (GFDL) Empirical Statistical Downscaling (ESD) team (https://www.gfdl.noaa.gov/esd_eval) has focused on evaluating some of these techniques. For an SD overview and our evaluation approach philosophy see our earlier works and cited references (Dixon *et al*., 2016; Lanzante *et al*., 2018; Lanzante *et al*., 2019a, hereafter L19a; Lanzante *et al*., 2019b, hereafter L19b). As a caution we note that even the best SD methods will fail to produce credible results if the driving physical climate model is flawed in its representation of circulation features (Hall, 2014; Maraun *et al*., 2017). Furthermore, large-scale models such as GCMs may not realistically represent sub-grid processes needed to simulate extreme precipitation (Giorgi *et al*., 2016; Maraun *et al*., 2017) in which case high resolution models may be needed.

We use a Perfect Model (PM) approach to test the 'stationarity assumption' inherent to all SD methods which implicitly assume that relationships defined during a historical period, intended to calibrate the method against observations, are valid for use in a future epoch when the climate has changed. The PM provides 'future observations' which do not exist in the real-world. However, as discussed below (see 2.1), it is important to note that the idealized nature of our PM design does not allow us to assess all sources of non-stationarities.

This is a follow-up to our recent SD work which assessed and improved representation of tails (L19a) and assessed daily maximum temperature (L19b). The methods we consider here and previously (L19a; L19b) are from a class of SD techniques operating distributionally, thus the expectation of better suitability than other SD techniques for reproducing extremes (i.e., tails). It is worth noting that this exercise is at a severe disadvantage (Maraun, 2013) because deterministic methods, such as those used here, cannot bridge the scale mismatch between GCM and observations, particularly for precipitation, having considerable sub-grid-scale variability (Chen and Knutson, 2008). Nevertheless there is value in our assessment because: (a) SD output from these methods is widely used in impact studies, (b) SD methods can provide bias correction, and (c) SD methods are often embedded in multivariate methods capable of bridging the scale mismatch.

2 | DATA AND METHODOLOGY

2.1 | Data

GFDL-HiRAM-C360 model data were introduced by Dixon *et al*. (2016) and used by Lanzante *et al*. (2018), L19a and L19b. We provide only a brief description as the reader is referred to theses earlier works, especially Dixon *et al*. (2016), for more details as well as appendix A of L19a for data availability.

Daily precipitation covering a rectangular region surrounding the Continental United States (CONUS), excluding oceanic points, constitute our PM data. Thirty years of data from a GCM driven by historical forcings cover the period 1979–2008. Thirty years of data based on three 10-year ensembles driven by forcings from a high emissions scenario (RCP8.5) cover the period 2086–2095.

Via our shorthand we refer to historical (future) observations as O_h (O_f), historical (future) model data as M_h (M_f), and future downscaled output as D_f, abbreviating observations (downscaled) as OBS (DWN). In our PM setup O_h and O_f (which can be considered *pseudo observations*) are raw GCM output at a spatial resolution ~25 km, while M_h and M_f are spatially averaged versions of O_h and O_f, respectively, yielding a resolution of ~200 km. The mismatch in spatial resolution is typical of that for real-world applications providing SD methods with a realistic challenge. In our PM world we refer to O_h and O_f (M_h and M_f) as OBS ('model' or 'GCM' output) even though all are GCM data.

Note that SD methods are typically faced with two challenges: (a) the spatial scale mismatch between model

and OBS and (b) model biases. Often in common usage the term 'statistical downscaling' is a misnomer, lacking explicit mention of (b). The fact that model values represent spatial averages results in (a). Strictly speaking the SD methods we use are bias correction methods. Since we use a single physical model to generate both model and pseudo-observations, our PM design explicitly introduces only challenge (a), thus we are not able to assess non-stationarities resulting from model biases in mean state or in climate change signals. More complex PM's, deriving 'OBS' and 'GCM' values from two different physical models would also explicitly introduce challenge (b). However, by way of spatial averaging our approach can introduce biases implicitly by altering distributions. We have chosen our simpler design in initial work as it facilitates easier diagnosis.

2.2 | Downscaling methodology

Guided by L19b, we use the two best performing methods for daily maximum temperature, one conceptually simple, quantile delta mapping (QDM) (Cannon *et al*., 2015) and one more complex, Kernel density distribution mapping (KDDM) (McGinnis *et al*., 2015). We also use Bias correction quantile mapping (BCQM) (Lanzante *et al*., 2018) which is both conceptually simple and very widely used. Finally, we consider PresRat, a modification of QDM designed specifically for precipitation (Pierce *et al*., 2015). Note that two of L19b's methods utilized the anomaly approach which is inappropriate for precipitation, a positive definite quantity. Below we briefly introduce the four methods—the interested reader is referred to L19b and references therein for more details.

2.2.1 | BCQM

BCQM, one of the most widely used SD methods, is often referred to as 'quantile mapping', although some use this term more generally in reference to various distributional SD methods. It is computed as:

$$F_{Df}(x) = F_{Oh}\left[F^{-1}{}_{Mh}(x)\right] \quad (1)$$

where F is the cumulative distribution function (CDF), F^{-1} its inverse and x the M_f value to be downscaled. Equation (1) is not applicable for novel values, that is, for M_f values outside the range of M_h values. When this occurs we use a modification of the standard extrapolation (Deque, 2007), detailed in L19a.

2.2.2 | KDDM

KDDM uses a complex, multi-step algorithm involving kernel density estimation to smooth O_h and M_h which have first been standardized to zero mean and unit variance, separately for each month of each year. Subsequent integration of the generated distribution functions followed by inverse standardization yields the desired transfer functions. We use R code kindly supplied by the KDDM authors (https://github.com/sethmcg/climod).

2.2.3 | QDM

QDM can be thought of conceptually as using M_f as a first guess, but modifying it via a correction factor, which varies by position in the distribution. The correction factor, while additive for most variables, is multiplicative for precipitation. Its additive form is:

$$D_f(x) = x + \left[F^{-1}{}_{Oh}\left(F_{Mf}(x)\right) - F^{-1}{}_{Mh}\left(F_{Mf}(x)\right)\right], \quad (2)$$

and its multiplicative form is:

$$D_f(x) = x\,X\left[F^{-1}{}_{Oh}\left(F_{Mf}(x)\right)/F^{-1}{}_{Mh}\left(F_{Mf}(x)\right)\right], \quad (3)$$

We use R code made available by the QDM authors (https://github.com/cran/MBC).

2.2.4 | PRAT

PresRat, hereafter referred to as PRAT, is a simple modification of QDM which preserves the model-predicted future change in mean precipitation. Each value of D_f computed from (3) is multiplied by a calendar-month specific correction factor K:

$$K = \left(\bar{M}_f/\bar{M}_h\right)/\left(\bar{D}_f/\bar{O}_h\right) \quad (4)$$

where bars indicate climatological means over a specific calendar-month.

2.2.5 | Configuration options

Although some authors have fixed specific options in their particular implementation, here we make a distinction between SD methods and configuration options. We consider an SD method to be associated with overarching principle(s) whereas configuration options to be specific

choices made in implementation. We evaluate configuration choices since in much prior work consequences of these choices have not been considered. Some of our motivation stems from Vrac *et al*. (2016), hereafter V16, who did assess some configuration options, although not in a PM setting.

Added complexity of precipitation yields several additional choices, four of which we consider. The first is whether to use an additive (A) versus multiplicative scaling (M). Conventional wisdom has dictated the latter for precipitation as additive correction can yield negative values—although these can be reset to zero. We consider the additive option since we are not aware of any prior studies that have evaluated this approach for precipitation.

Another option is frequency adjustment (freqadj) in which a threshold (above the US trace value of 0.01 in.) is chosen yielding the same fraction of dry days in M_h as found in O_h, by setting M_h values below the threshold to zero. The same threshold is applied in the future. Frequency adjustment is aimed at accounting for the well-known GCM 'drizzle bias' (Stephens *et al*., 2010), hereafter DBIAS, the widespread tendency for GCMs to simulate an excess number of small amounts of precipitation compared with real-world observations.

A fundamental issue in precipitation SD is how to deal with days having zero precipitation. One could ignore them and apply SD only to non-zero values or SD could be applied to all values, including the zeros. We introduce the option ignore0, which when on (off) applies to the former (latter) case. A further issue arises with ignore0 off, namely the potential for a large number of identical values (i.e., zeros). Cannon *et al*. (2015) added a small random value (distributed uniformly over [0, trace/2]) to each zero before downscaling. We refer to this option as below trace noise (BTN), which can either be on or off.

Note that the four configuration options are not applicable to all of our SD methods. The choice of additive versus multiplicative is not applicable to BCQM or KDDM by nature of their algorithms. Furthermore, freqadj and ignore0 are 'baked into' the complex KDDM code. For QDM and PRAT all four options are viable, which is an extension to the methodology given by the original authors.

Although we consider four configuration options, our list is not exhaustive. For example, time windows used for SD training and evaluation can vary: 12-monthly or 4-seasonally non-overlapping, or overlapping moving windows of different lengths, to name a few possible choices. There is a tradeoff: wider windows yield larger sample sizes for training but narrower ones are better able to resolve seasonally varying relationships. Window choice may introduce artefacts (Dixon *et al*., 2016; especially Figure 4b). Hence, configuration options beyond those considered here may have consequences.

2.2.6 | Tail schemes

Special attention is warranted to tails, which present a greater challenge than the remainder of the distribution. We examined this issue in considerable detail and have devised special procedures (L19a) which were evaluated extensively (L19b). We use the limited tail adjustment scheme (LIM) as per L19a and L19b. LIM is applied only to tail values after initial application of any arbitrary SD method.

For application of LIM the user decides a priori how many values at the end of the distribution are to be modified by specification of the parameter tail-length (TLN). The user also specifies, via parameter NPT, the number of values to be used in performing the tail adjustment. Previously (L19b) we found TLN = 10 and NPT = 10 yield good results for temperature. While smaller values of NPT produce poorer results, increasing NPT beyond 10 generally yields little gain.

Conceptually LIM operates by computing a constant correction factor from the NPT points and applies it to the TLN points. For example, with NPT = 5 and TLN = 10 to apply LIM to the right tail the correction factor is computed using the 11th through 15th largest values and then applied to the 10 largest values. The correction factor is either additive, used for most variables such as temperature, or multiplicative, assumed more appropriate for precipitation. Here our LIM results are multiplicative, following conventional wisdom, with NPT = 10 and TLN = 10, based on L19a. We only apply LIM to the right tail as left-tail values are small, marginally larger than the trace value.

2.3 | Evaluation procedure

Evaluation statistics are presented by treatment, which we define as a combination of an SD approach (one SD method for either the base algorithm, or additionally with LIM adjustment in the right tail) and a set of configuration options as detailed in section 2.2.5. Each of the three 10-year future ensembles is downscaled separately and verification statistics are averaged over the ensembles. SD is performed separately at each gridpoint and for each of four standard seasons DJF (December, January, February), MAM, JJA, and SON. Results are presented mostly as averages over the four seasons with some limited results given for DJF and JJA. Use of seasons rather than months (as in our earlier works for daily temperature) is aimed to provide adequate sample sizes given that for some approaches, the presence of dry days reduces available sample size, sometimes substantially.

Our primary metric is the mean absolute error (MAE) which we use in two different ways. In the traditional synchronous application MAE is based on differences between values of D_f and O_f occurring on the same day. We also apply it asynchronously (MAE-ord) such that paired values of D_f and O_f represent the same order statistics. While MAE is a measure of how well SD represents day-to-day weather variations, MAE-ord measures how well SD reproduces the statistical distribution of values. MAE-ord is motivated by guidance provided by Maraun and Widmann (2018) and Maraun *et al.* (2019) regarding statistical model evaluation. Contrasting results based on these two similar but distinct metrics will help illustrate issues raised by Maraun (2013) regarding the stochastic nature of precipitation. After computing MAE or MAE-ord over a season, or averaging the four seasonal values, it is converted to a standard skill score (Wilks, 2006; L19a; L19b):

$$\text{Skill} = [(\text{MAE}_{\text{Mf}} - \text{MAE}_{\text{Df}})/\text{MAE}_{\text{Mf}}]\,\text{X}\,100\% \qquad (5)$$

We then average skill over all non-ocean gridpoints. Averaging utilizes the biweight mean (Lanzante, 1996) which guards against effects of outliers. The biweight is data adaptive behaving more like the arithmetic mean for 'well-behaved' data or more like the median otherwise.

As in L19a and L19b we compute separate verification statistics for different portions of the distribution referred to as distributional categories (CAT's): CAT 1 (CAT 9) consists of the lowest (highest) value in the sample, CAT 2 (CAT 8) the second-third lowest (highest) values, CAT 3 (CAT 7) the fourth-sixth lowest (highest) values, and CAT 4 (CAT 6) the 7th–10th lowest (highest) values. Finally, CAT 5 consists of all values in the sample. When considering extremes we devote our attention to the right tail, as values in the left tail are very small. Some results are presented as averages over the entire right tail (CAT 6–9), weighted by the number of values in each category.

Because of the binary nature of precipitation occurrence we utilize an additional metric in the form of a fractional error in the number of dry days, computed separately for M_f and D_f. It is computed as the number of days in error divided by the total number of days. For example, in the case of D_f, if D_f and Of are both wet days or both dry days there is no error. On the other hand, if one is wet and the other dry we count this as an error. As above, given fractional errors for both M_f and D_f we compute a skill.

In order to estimate statistical significance we use the same procedure developed previously, referring the reader to L19a (especially appendix B) for details, with only a brief overview here. We first compute the mean skill over our domain, separately for two SD approaches of interest. The difference between these two skills is the quantity for which significance is sought. We perform two separate Monte Carlo simulations, with 1,000 trials each to derive a distribution of differences in skill. The position of the actual difference in this randomly derived distribution determines the significance level.

The first step in the process is to estimate the spatial degrees of freedom in order to account for the fact that gridpoint values are not independent of one another. For each trial we apply random translational shifts in both the north–south and east–west directions to the pair of maps. Next we pattern correlate the original and shifted pairs of maps and use the distribution of correlations to infer an effective block size. In the second step, for each trial we randomly shuffle blocks of gridpoints between the two original maps and compute a difference in mean skill between the permuted maps. The distribution of these differences is used to assess significance. In order to ensure robust results we report three significances based on conservative and liberal objective estimates as well as a very conservative subjective choice (L19a). We modify the subjective choice of effective number of blocks of 4×2 (latitude by longitude gridpoints) used in our earlier works for temperature to 7×4 for precipitation based on Huang *et al.* (1996) and Richman (1986).

3 | RESULTS

3.1 | Evaluation of skill over the entire distribution

Table 1 summarizes skill (based separately on MAE and MAE-ord) averaged over the entire domain for various SD treatments. Rather than considering every possible combination, we limit to a manageable number from which we can draw conclusions considered representative of the class of SD methods examined in our PM framework. Our shorthand for treatments uses the first letter of the SD method followed by the ordinal row number from the first column of Table 1 which specifies configuration options. During discussion we also refer to Table 2, with results from a limited number of significance tests, pairwise between two treatments. For convenience, Table 2 lists group numbers (i.e., G1–G7) for sets of related significance tests.

The most noteworthy aspect of Table 1 is the clear distinction between skills based on MAE versus MAE-ord, with the former substantially lower. Higher skill for MAE-ord reflects the fact that while quantile mapping substantially improves the distribution of values compared with

TABLE 1 Skill (%) averaged over the domain for various SD methods and configurations (T/F/I/B) referenced by treatment number (first column). CAT 5 is for the entire distribution whereas CAT 6–9 is averaged over the right tail. CAT 6–9 L is based on the LIM adjustment averaged over the right tail

			MAE			MAE-ord		
	Method	**T F I B**	**CAT 5**	**CAT 6–9**	**CAT 6–9 L**	**CAT5**	**CAT 6–9**	**CAT 6–9 L**	**D-DRY**
1	BCQM	– – I –	−1.7	−42.0	−33.2	1.7	0.0	7.4	13.3
2	BCQM	– F – –	22.2	−20.1	−9.0	59.1	24.5	28.8	32.6
3	BCQM	– F I –	20.4	−22.5	−12.0	52.5	22.0	28.1	30.3
4	BCQM	– – – –	22.9	−20.0	−9.3	60.7	24.6	30.0	34.2
5	KDDM	– F I –	21.4	−18.4	−8.6	56.3	27.2	31.4	34.3
6	QDM	M – – B	21.8	−15.6	−6.2	56.2	30.9	33.9	32.6
7	QDM	M F I –	20.4	−16.3	−7.5	52.8	28.9	31.2	30.4
8	QDM	M F – B	20.7	−15.6	−6.5	53.0	30.9	31.9	26.8
9	QDM	M F – –	20.2	−15.6	−6.2	53.2	30.9	33.9	31.2
10	QDM	M – – B	21.5	−15.6	−6.3	55.9	30.9	33.9	31.5
11	QDM	M – – –	21.6	−15.6	−6.2	55.8	30.9	34.0	31.6
12	QDM	A F I –	20.2	−12.8	−6.3	51.8	30.2	30.1	30.4
13	QDM	A F – B	15.3	−13.6	−5.4	47.7	30.5	25.1	27.2
14	QDM	A F – –	14.7	−13.6	−6.3	47.6	30.5	31.6	32.2
15	QDM	A – – B	15.4	−13.6	−5.5	49.5	30.4	25.1	30.7
16	QDM	A – – –	15.4	−13.6	−6.3	49.3	30.5	31.5	30.2
17	PRAT	M F I –	21.1	−12.7	−5.4	53.8	30.2	30.4	30.5
18	PRAT	M F – B	22.9	−8.9	−2.3	58.6	33.2	32.3	31.4
19	PRAT	M – – B	23.8	−8.8	−2.2	60.7	33.0	32.1	32.0

Note: Skill is based on three different error metrics: MAE, MAE-ord or dry day frequency (D-DRY). Regarding configuration options: T (QDM and PRAT only) is type of approach, M (multiplicative) or A (additive); F is frequency adjustment (F for on, – for off); I is ignore0 (I for on, – for off); B (QDM and PRAT only) is below trace noise (B for on, – for off). All treatments use as the dry day cutoff the US trace value of 0.01 in. except Q6 which uses 0.05 mm, the default value from the authors of the QDM code.

the raw GCM output, it exhibits less skill in reproducing the day-to-day weather sequencing, as measured by MAE. This can be explained by the arguments made by Maraun (2013) that attempting to bridge the scale mismatch between OBS and GCM using a deterministic method such as quantile mapping yields results with a corrupted time sequence. As such, MAE skill levels are disappointingly low (~20–25%), about half that found for temperature in our earlier work (L19a; L19b). Sub grid-scale variability, which is at the core of the problem, is much less of an issue for temperature than it is for precipitation. It is worth noting that the relative ordering of skills by treatment are generally quite consistent between the two metrics—the correlation of MAE and MAE-ord skills across treatments exceeds 0.9 for CAT5 and CAT6-9. Thus, comparisons between treatments—one of the main motivations for this work—are not much affected by the choice of metric. For the remainder of this work we draw conclusions based primarily on MAE-ord.

Next we examine overall skill (CAT 5) for four different configurations of BCQM (B1–B4) based on combinations of freqadj on/off and ignore0 on/off. BCQM was chosen for this purpose since it is the simplest of our methods, has the fewest possible options, and is a very commonly used distributional SD method. Three of the configurations (B2–B4) yield skill of the same order whereas B1 performs much more poorly. Significance tests in Table 2 for group G1 indicate that the outlier treatment (B1) is highly significantly worse than the others while the two better treatments (B2 and B4) are not different from one another. The key factor is that the better treatments have ignore0 off. However, use of freqadj can substantially mitigate the effect of having ignore0 on (B3), although this treatment is still significantly worse than the two better treatments. Below (3.5.2) we explore the reasons for this behaviour, explaining why it is configuration-specific but not SD method-specific. Finally, we consider K5 which has skill a bit lower than B4, but not significantly so.

Next we examine results based on a variety of treatments for QDM, chosen for this purpose since previously (L19b), for temperature, it was found to be as good or

TABLE 2 Significance level (%) testing the hypothesis that there is no difference in skill between the two specified SD treatments

Group	Nblon × Nblat			Significance (%)			Treatments	Categories
G1	7 × 4	2 × 1	3 × 2	0.0	6.0	0.0	B4 × B1	5
G1	7 × 4	4 × 3	7 × 4	0.2	5.1	0.2	B4 × B3	5
G1	7 × 4	6 × 3	10 × 5	62.9	68.7	50.1	B4 × B2	5
G1	7 × 4	6 × 3	8 × 5	11.3	20.6	6.4	B4 × K5	5
G2	7 × 4	5 × 3	9 × 5	85.2	92.1	82.8	Q12 × Q14	5
G2	7 × 4	6 × 3	10 × 5	10.9	20.2	3.7	Q6 × Q7	5
G2	7 × 4	7 × 4	12 × 7	30.7	30.7	9.8	Q6 × Q9	5
G2	7 × 4	6 × 3	9 × 5	6.6	14.6	3.3	Q6 × Q12	5
G2	7 × 4	6 × 4	10 × 6	14.2	15.4	3.6	B4 × Q6	5
G3	7 × 4	5 × 3	7 × 5	0.5	2.7	0.0	P19 × P17	5
G3	7 × 4	7 × 4	12 × 7	7.5	7.5	0.6	P19 × Q6	5
G3	7 × 4	6 × 4	9 × 6	5.3	6.7	1.5	P19 × K5	5
G4	7 × 4	2 × 2	5 × 3	0.0	9.7	0.1	B4 × B1	6–9
G4	7 × 4	4 × 3	10 × 5	32.5	54.6	28.9	B4 × B3	6–9
G4	7 × 4	4 × 3	10 × 5	41.5	61.8	34.7	B4 × B3	6–9 LM
G4	7 × 4	4 × 3	10 × 5	8.8	26.7	5.4	B4 × B4	6–9 vs. 6–9 LM
G5	7 × 4	5 × 2	7 × 5	50.0	68.6	44.3	P19 × Q6	6–9
G5	7 × 4	4 × 3	9 × 5	43.5	62.3	38.8	P19 × B4	6–9 LM
G5	7 × 4	5 × 2	7 × 5	50.6	72.1	46.9	P19 × Q6	6–9 LM
G6	7 × 4	2 × 1	4 × 3	0.0	0.1	0.0	B4 × B1	DD
G6	7 × 4	7 × 4	13 × 7	12.4	12.4	0.9	B4 × B3	DD
G6	7 × 4	7 × 5	15 × 8	52.4	44.8	22.2	B4 × B2	DD
G7	7 × 4	9 × 5	14 × 9	44.6	30.2	15.2	K5 × P19	DD
G7	7 × 4	9 × 5	14 × 9	45.9	32.8	16.4	B4 × P19	DD

Note: For MAE-ord, skill is either for the entire distribution (CAT 5) or as a weighted average over the right tail (CAT 6–9). Limited tail adjustment is indicated by LM. For DD, skill is based on dry-day metric (D-DRY). Paired comparisons are aggregated into groups which are considered together in the discussion in the text. For each comparison there are three probabilities (left to right) corresponding to three effective grid sizes. The grid sizes (Nblon × Nblat) are expressed as the number of blocks in the longitudinal and latitudinal dimensions. For example, 7 × 4 indicates a grid composed of 28 blocks. During Monte Carlo simulation each block (consisting of multiple gridpoints) is shuffled as a unit. The leftmost is a subjectively determined grid dimension and is likely too small. The centre (right) dimensions are based on more conservative (liberal) criteria derived from Monte Carlo analysis (see Section 2.3). Treatments consist of a downscaling method identified by the leading letter (B = BCQM, K = KDDM, Q = QDM, and P = PRAT) and configuration options in reference to the ordinal row number given in Table 1.

better than any of the tested distributional methods, and has the most configuration options. In the literature, conventional wisdom from a variety of SD methods suggests that multiplicative scaling rather than additive adjustment should be used for precipitation. One of the reasons for this is that the additive approach may result in negative precipitation—which then must be reset to zero. Since this assumption is rarely tested we have applied a number of such treatments with different configurations. Indeed, the multiplicative approach is superior in all cases. Not coincidently the best approach for QDM is Q6, the default configuration in the code made publicly available by the creators of QDM (Cannon *et al.*, 2015). However, while the multiplicative and additive QDM treatments tend not to differ significantly from one another, the best of the former group (Q6) is marginally significantly better than the best of the latter group (Q12) as seen in G2. Finally we note that the best versions of BCQM (B4) and QDM (Q6) do not differ significantly (G2).

The final method under consideration is PRAT, which (see above) is a variant on QDM. Having used QDM to explore various configurations we limit the number of PRAT treatments. Not surprisingly the best (P19) has the same configuration as the best QDM (Q6). Although P19 is not better than B4, it is marginally significantly better than both Q6 and K5 (see G3), and

TABLE 3 Comparison of MAE-ord skill (%) across V16 configuration scenarios: Positive correction (PC), threshold adaptation (TA), direct approach (DA), and singularity stochastic removal (SSR)

Scenario	Treatment	BCQM	QDM	PRAT
	F I B			
PC	– I –	1.7 (B1)		
TA	F I –	52.5 (B3)	52.8 (Q7)	53.8 (P17)
DA	– – –	60.7 (B4)	55.8 (Q11)	
SSR	– – B		55.9 (Q10)	60.7 (P19)

Note: Skill values have been extracted from Table 1 for configurations that match those of V16 with corresponding treatments from Table 1 given in parentheses.

certainly the poorest version of PRAT (P17) which has ignore0 on.

Next we consider the effect of the BTN option. Although we have not made extensive tests, there are several pairings that differ only in the BTN setting (Q6/Q11, Q8/Q9, Q13/Q14, and Q15/Q16). In general there is very little difference.

Finally, Table 3 gives a limited comparison between our findings and those involving four configuration scenarios of V16: Positive Correction (PC), Direct Approach (DA), Threshold Adaptation (TA), and Singularity Stochastic Removal (SSR). These are analogous to our configurations: PC (freqadj = off, ignore0 = on), DA (freqadj = off, ignore0 = off, BTN = off), TA (freqadj = on, ignore0 = on), and SSR (freqadj = off, ignore0 = off, BTN = on). Although V16 utilized a single downscaling method (CDFt; Michelangeli *et al.*, 2009), not used here because of sub-par performance (L19b), our use of multiple methods enables us to demonstrate much greater sensitivity of results to configuration rather than SD method.

In agreement with V16 we find that, of the methods tested here using our PM experimental design, PC is by far the worst approach with the other three yielding fairly similar results, although TA may be slightly poorer than DA and SSR. We also agree that SSR (also used by Zhang *et al.*, 2009 and Cannon *et al.*, 2015) is preferable because of its flexibility in dealing equally well when M_h has more wet days than O_h (i.e., the DBIAS) as well as the inverse. Recall that use of freqadj can only remedy the DBIAS not the inverse. The SSR approach is also simpler in that it avoids having separate corrections for occurrence and amount.

3.2 | Evaluation of skill in the tails of the distribution

Skill based on MAE-ord averaged over the right tail using the base algorithm (CAT 6–9 in Table 1) is about half that

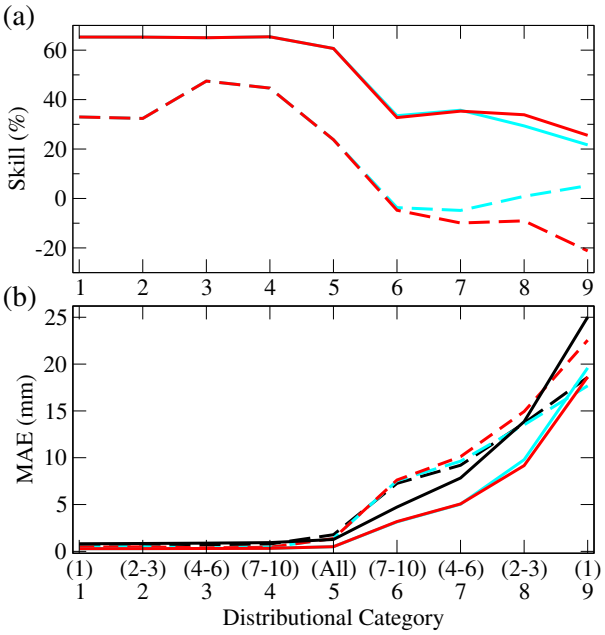

FIGURE 1 Skill (%) (a) based on MAE and MAE-ord and (b) MAE and MAE-ord (mm) for P19 as a function of distributional category. Curves depict basic algorithm (red), LIM (cyan), and GCM (black), for both MAE-ord (solid) and MAE (dashed). The abscissa axis label gives the distributional category number (1–9) along with the corresponding ordinal point numbers in parentheses on the left/right indicating the first (1), second to third (2–3), fourth to sixth (4–6), and seventh to tenth (7–10) the lowest/highest values from the sample. The 'all' category includes all available values from the distribution. The skill, MAE or MAE-ord shown is the biweight mean over all non-ocean gridpoints, three ensembles, and 12 months [Colour figure can be viewed at wileyonlinelibrary.com]

for CAT 5 (~25–30% vs. 50–60%). Other than the B1 outlier, differences in CAT 6–9 skill between treatments are generally small and few if any are likely to be significantly different as evidenced by the B4 versus B3 comparison (G4) which exhibits fairly typical differences. Application of the LIM adjustment yields small improvements which are likely to be mostly insignificant with again not much difference between treatments (G4). Furthermore, tail skill for our preferred treatment (P19) does not differ significantly from that of other leading treatments (G5).

Skill based on MAE shows a very different pattern. While CAT 6–9 skill is much poorer than for CAT 5, LIM yields substantial improvement. However, skill in the tails for both the basic algorithm as well as LIM adjustment is negative, indicating they provide no improvement over the raw GCM. Here the results in the right tail are strikingly different than was found previously for temperature (L19b) for which skill in the tails was comparable to the whole distribution.

For further perspective on tail performance Figure 1 shows MAE and MAE-ord, along with their associated

skills, for P19 as a function of distributional category. High skill in the left tail is of little practical importance since the values there are quite small. Aside from consistency with the general points made above, this figure demonstrates the rapid increase in error going farther out in the right tail. Although P19, likely as good as any of the tested treatments, shows considerable improvement over the raw GCM, errors in the tail (representing domain averages) are still quite large ~5–20 mm.

3.3 | Evaluation of dry-day skill

The right-most column of Table 1 gives skill based on the dry-day error metric. Other than for B1, which is highly significantly worse than other treatments (G6), differences tend to be not too dissimilar. This is indicated by the fact that while P19 differs by more than two from both B4 and K5, these differences are not even close to being significantly different (G7). Note that the actual fractional error (not shown) does not differ much between SD treatments, with that for M_f ~0.27 and that for D_f ~0.18.

3.4 | Spatial patterns of skill and MAE

The pattern of DJF skill for CAT 5 in Figure 2a shows high skill over much of the eastern and western portions of the domain with lower skill in central portions. However, in the mountainous west there are some locations with very high skill, often in close proximity to locations with low or negative skill. This phenomenon is explored in detail below (3.5.1). For CAT 5 during JJA the areas of high skill in the western U.S. are greatly reduced, with too little precipitation for analysis in portions of this region, and moderate skill in much of the Midwest. A curious feature is a strip of lower skill extending from the Gulf coast of Mexico up through the mid-Atlantic. Our earlier work (Dixon *et al*., 2016; Lanzante *et al*., 2018)

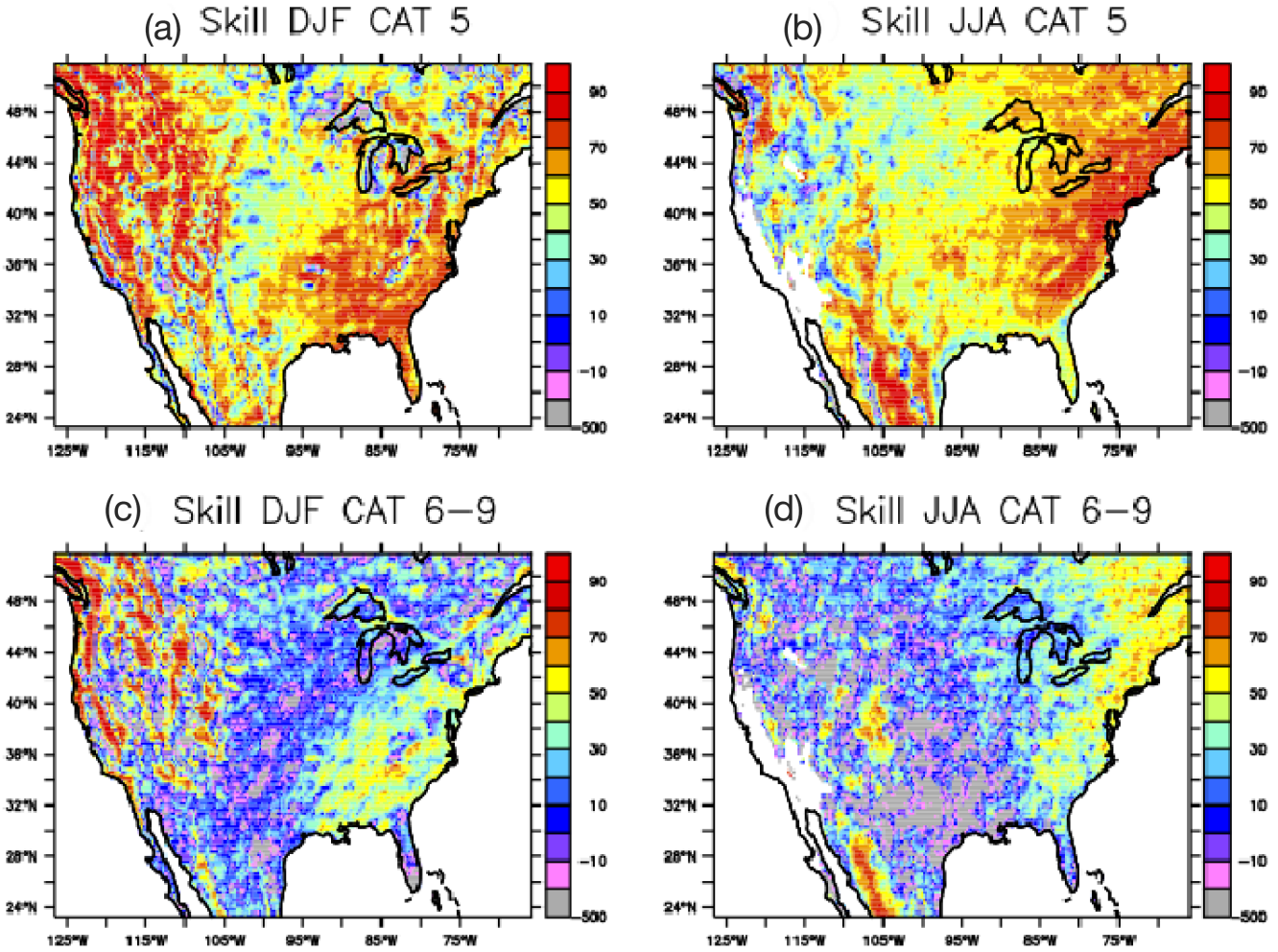

FIGURE 2 Maps of P19 skill (%) based on MAE-ord for CAT 5 (a and b) and the average of CAT 6–9 (c and d) for DJF (a and c) and JJA (b and d). White areas have too little data for analysis [Colour figure can be viewed at wileyonlinelibrary.com]

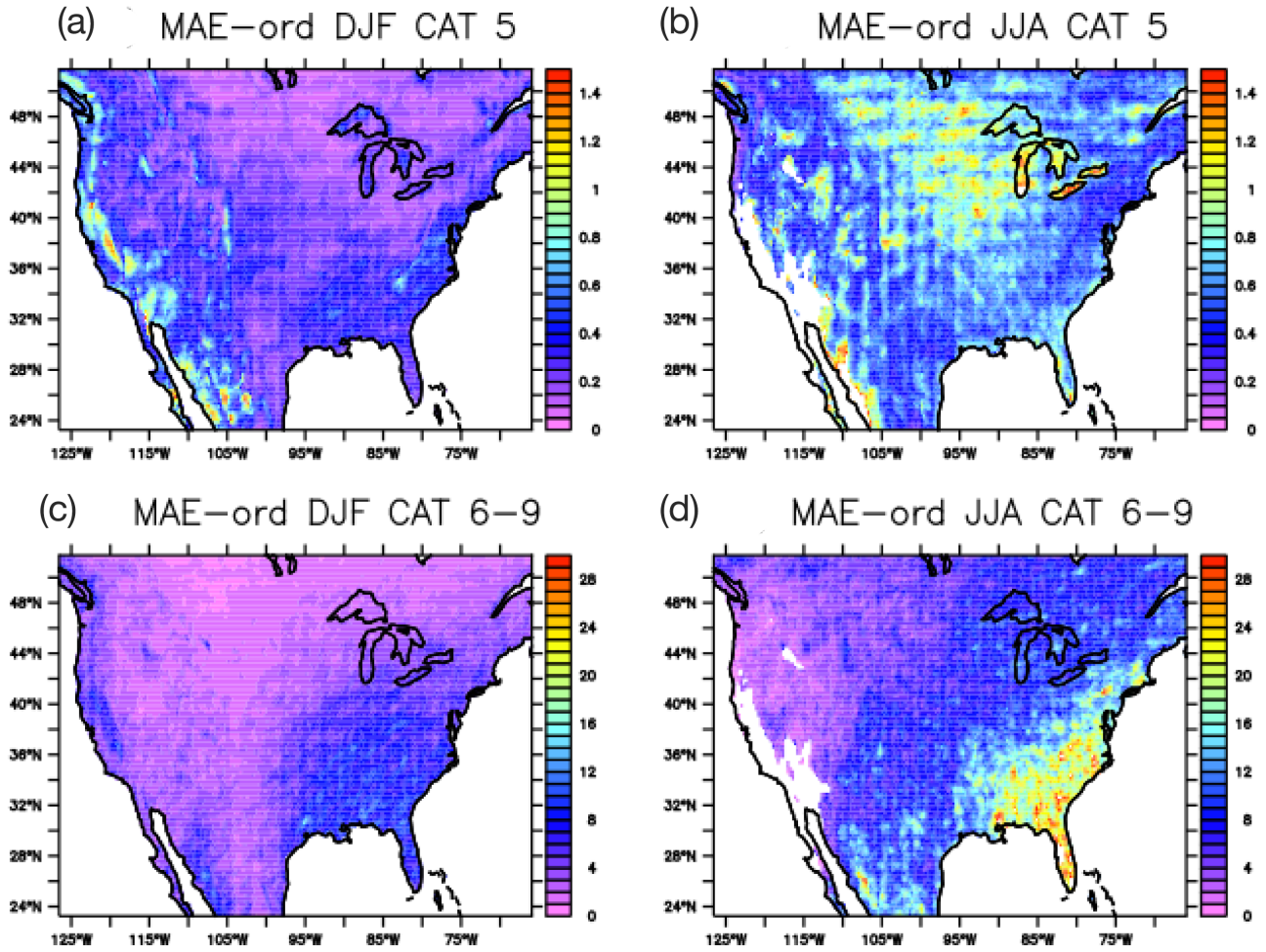

FIGURE 3 As in Figure 2 except for MAE-ord (mm) [Colour figure can be viewed at wileyonlinelibrary.com]

identified lower skill for temperature downscaling in coastal regions. Although diagnosis of this feature is beyond the scope of this work the mechanistic explanations from our earlier work would not seem applicable here. In the tails (Figure 2c,d) the patterns are roughly similar to their corresponding CAT5 patterns, but with a considerable reduction in overall skill. As a consequence, substantial areas of the domain have near-zero or negative skill.

For MAE-ord the DJF patterns for CAT 5 and CAT 6–9 (Figure 3a,c) both have higher values over the Southeast and along the extreme West Coast, which correspond roughly to the DJF climatology (not shown). For JJA while the CAT 6–9 pattern (Figure 3d) also corresponds reasonably well to the JJA climatology (not shown), the CAT 5 pattern (Figure 3b) does not, with largest errors in the higher latitudes of interior North America. The domain-averaged MAE-ord seen in Figure 1 of ~5 mm for CAT 6–9 masked the strong regionality seen in Figure 3d where the tail errors along the East Coast, and particularly the Southeast are typically several times larger ~20 mm, even in areas exhibiting considerable skill (Figure 2d).

There is an interesting contrast between Figures 2b and 3b that relates to precipitation frequency. While lowest skill is found both in the mountainous West and the Northern Interior, the former (latter) has relatively low (high) MAE-ord. As illustrated below (3.5.2), having a very large number of dry days (i.e., zero values) complicates distributional mapping. The extreme aridity of the Far West results in either an insufficient sample for analysis or very small errors, while the Interior has just a bit more total and frequency of precipitation to trigger the complications.

3.5 | Case studies

3.5.1 | Jupiter and Coaldale

The mountainous West has some distinctive variations in skill and MAE-ord characterized by sharp gradients

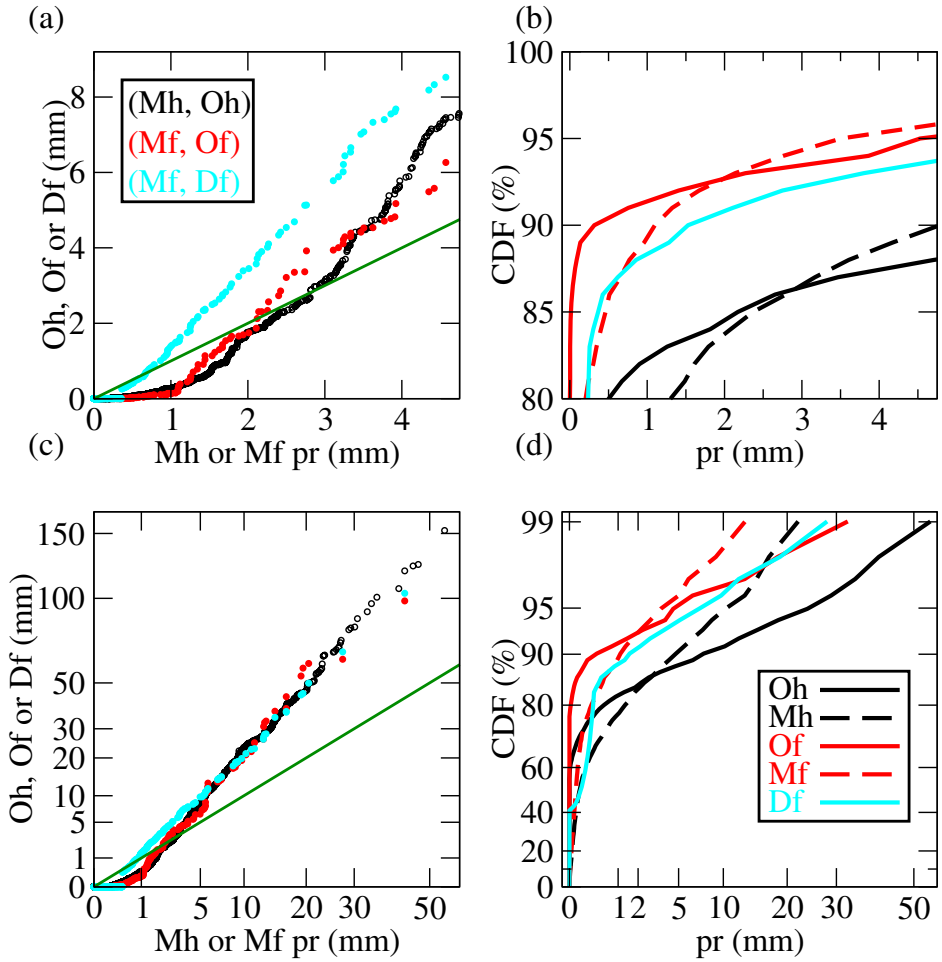

FIGURE 4 Quantile-quantile (qq) plots (a) and (c) and cumulative distribution function (CDF) plots (b) and (d) for Jupiter, CA (38.1°N, 120.2°W; member three during SON; ~1,065 m). For the qq plots the abscissa has precipitation (mm) for the GCM (M_h or M_f) and the ordinate OBS or DWN (O_h, O_f or D_f). Black circles have coordinates (M_h, O_h), red circles (M_f, O_f), and cyan circles (M_f, D_f) where D_f is based on P19. The plots on the right side have CDFs for O_h and M_h (black), O_f and M_f (red) and D_f (cyan), with dashed for model and solid for observed. For clarity the plots on the top omit ~10% of the largest abscissa values while those at the bottom cover the entire range of values. The CDF plots in (b) cover only the upper portion of the ordinate values to focus on the portions having the largest displacement between the curves given that much of the lower portions of the CDFs pertain to dry days. The green line represents y = x. For (c) the x and y axes are scaled using a square-root transformation while for (d) the x axis uses a square-root transformation while the y axis uses an inverse error function transformation [Colour figure can be viewed at wileyonlinelibrary.com]

aligning with topography (Figures 2a and 3a). In the West, the basic nature of this pattern for CAT 5 is similar among the four SD methods and all seasons (not shown) except JJA which differs somewhat due to climatologically much drier conditions.

These variations can be explained in terms of interaction between topography and nonstationarity introduced by climate change. To illustrate this, two relatively nearby points are chosen for which the behaviour is strikingly different. Near Jupiter, CA (Figure 4), CAT 5 MAE-ord skill for P19 is highly negative (−62.0%) yet at a gridpoint near Coaldale, NV (Figure 5), skill is impressively high (90.9%). Here the use of different seasons and members (see captions Figures 4 and 5) was made to accentuate the disparity, but is not crucial to the conclusions.

Jupiter is upwind of the Sierra Nevada in a region of rapidly rising elevation from west to east whereas Coaldale is downwind on the plateau. As a result, Jupiter is located near an orographically forced climatological local maximum of precipitation while Coaldale is near a climatological minimum. Accordingly, for Jupiter (Coaldale) the much larger GCM footprint encompasses areas of less (greater) precipitation in the surrounding areas. As seen in Table 4 at Jupiter (Coaldale) mean precipitation of OBS is greater (less) than GCM. Although the climate change signal at both locations is that of drying, this effect is more

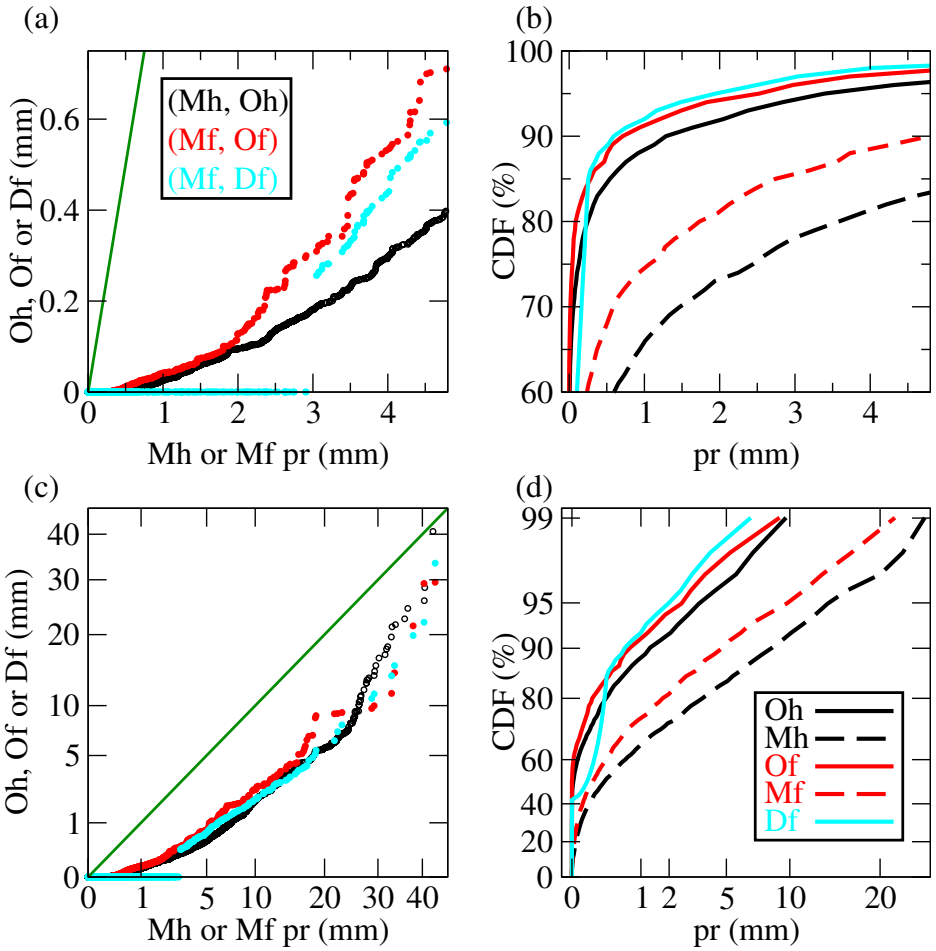

FIGURE 5 Same as Figure 4 except for Coaldale, NV (37.9 °N, 117.7°W; member two for DJF; ~1,540 m). Note that in (a) cyan values pinned at zero correspond to red values below the trace—thus these two sets are indistinguishable as 'dry days' [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Climatological mean daily precipitation (mm·day^{-1}) at Jupiter CA (38.1°N, 120.2°W; member three during SON) and Coaldale NV (37.9°N, 117.7°W; member two for DJF)

	M_h	O_h	M_f	O_f
Jupiter	1.63	3.23	0.80	1.39
Coaldale	2.60	0.59	1.70	0.50

pronounced at Jupiter, which, since it is wetter, has more to lose via climate change.

A better understanding is had by examining quantile-quantile (qq) plots at the two locations. These plots are like traditional x-y plots except that instead of each x-y pair coming from the same point in time, they come from the same relative location in their respective distributions. For example, the right-most (left-most) point consists of the maximum (minimum) x-value paired with the maximum (minimum) y-value. To complement the qq plots we also show corresponding plots of CDF curves. For added clarity, both types of plots are shown zoomed in (top) on the most salient features as well as over the full range (bottom) of values (with nonlinear scaling).

The qq plots (Figures 4a,c, 5a,c) consist of three curves, each with GCM on the abscissa and OBS or DWN on the ordinate; black (red) depicts the historical (future) relationship. Thus, black represents what is available to 'train' the downscaling method while red represents 'truth'. The cyan curve represents what the SD method generated as its rendering of the future. One can think of the green line, with a slope of 1, as the starting point—downscaling moves from it to the cyan curve—with the best possible result lying on the red curve. The point at which curves cross y = x shows where the bias changes sign. For points above (below) the green line OBS is greater (less) than GCM. Changes from black (historical) to red (future) indicate non-stationarity.

In the case of Jupiter (Figure 4a,c) the downscaled (cyan) departs significantly from the truth (red) for most of the distribution. To understand the downscaling operation we use QDM—the results for which (not shown) are quite similar, only slightly worse. However, the basic principles apply to any of the methods used herein operating via analogous bias correction principles. Downscaling via QDM consists of using M_f as a 'first guess' and then modifying this through a multiplicative factor

expressed as the ratio of values O_h/M_h for a given percentile in their respective distributions (see Equation (3)). The percentile is that of M_f from its distribution.

At first glance it seems curious that downscaling does so poorly given that the red curve is not that different from the black one. But these curves only represent relative relationships. The key lies in the fact that the distributions have shifted significantly towards lower values in the future (Figure 4b,d) due to drying (Table 4). Another important factor is that the black curve shifts from a local ratio (OBS to GCM) much larger than 1 (i.e., above the green line) at the high end of the distribution to values less than 1 at the low end. This ratio is proportional to the multiplicative correction factor applied by QDM. Note that ratio values less than 1 at the very low end of the distribution are to be expected as per the well-known DBIAS which results from the GCM having a larger footprint than the OBS.

Consider an arbitrary value of Mf which is to be downscaled. The correction factor (Equation (3)) is determined by applying the percentile of this value from the M_f distribution to the O_h and M_h distributions. However, because of considerable drying from historical to future periods the quantile (i.e., amount of precipitation) corresponding to this percentile will be higher in the historical than the future periods. Thus, the correction factor which is applied will be biased towards the high end of the historical distributions. Since, as we noted above, the O_h/M_h ratio increases with increasing precipitation amount, the resulting correction factor will be too large.

There is an equivalence between the explanations based on qq plots and CDFs (Figure 4b,d). Note in Figure 4a the cross-over of the ratio O_h/M_h occurs at ~3 mm where the black curve intersects the green line. In the CDFs the O_h and M_h curves cross at ~3 mm (having a CDF value ~86%), with the O_h curve to the left of the M_h curve for values less than ~86% and vice versa. Similarly, the D_f curve is to the left of the M_f curve for CDF values less than ~86% and vice versa.

Next consider Coaldale (Figure 5), located in the orographically induced down-wind 'rain shadow' region. The qq curves are below the green line due to the DBIAS but unlike Jupiter they do not cross above it for higher values because the DBIAS operates only for low values of precipitation. For larger amounts of precipitation, especially convective, a localized intense area of precipitation is more likely surrounded by less intense precipitation, leading to in effect an inverse of the DBIAS. The fact that Coaldale is so arid (compare the y axis extents in Figures 4a,c and 5a,c) precludes it from reaching the inverse DBIAS regime which Jupiter is able to attain. Furthermore, although there is drying, it is less dramatic than at Jupiter. The combination of the drying and percentile shift effect does force the cyan curve above the black curve, just as was the case for Jupiter, but the amount is much less. But this shift has a positive effect by pushing the cyan curve closer to the red curve, in contrast to Jupiter in which it pushed it away from the red curve. In passing we point out that the discontinuity in the cyan curve is a reflection of SD adjustment for the DBIAS (i.e., converting some wet days to dry days). It is more prominent here because of the very dry climate which results in a 'zooming in' on the low end of the distribution.

In summary, drying due to climate change and the inevitable DBIAS operate in opposite fashion between the two locations. At Jupiter they conspire to yield a poor downscaling result. Because the magnitudes are smaller at Coaldale they have a smaller effect, but fortuitously they combine to produce an exceptionally good result.

3.5.2 | Alpena

Motivation for this example comes from results shown earlier (Table 1) comparing four variants of BCQM. Treatment B1, having ignore0 on and frequency adjustment off produced much worse results. Maps not shown here, show a considerable similarity between patterns of seasonal variation in B1 skill and patterns of seasonal variation in climatological amount and frequency of precipitation. Namely, relatively low skill for B1 corresponds with climatological small amounts and daily frequencies of precipitation, with poorest performance during JJA. Furthermore, seasonal patterns of skill for our favoured approach of P19 are quite similar to those for the better performing variants of BCQM (B2-B4).

The qq plots in Figure 6a,c for a point near Alpena, KS typify the poor performance of B1. The relationship between OBS and GCM is similar between the historical (black) and future (red) periods. When configured optimally with ignore0 off, BCQM (B4) and PRAT (P19) yield similar and reasonable results (violet and cyan circles, respectively) that follow the O_h/M_h and O_f/M_f curves quite well. On the other hand, turning ignore0 on, with all other settings the same yields again similar, but this time extremely poor results for both BCQM and PRAT (violet and cyan pluses, respectively). It is striking that skill at this location for B1 is—26.9% while that for B4 is 52.9%.

In our PM experimental design, we find that poor performance with ignore0 on and frequency adjustment off is not downscaling method specific but instead is much more likely to occur when two conditions are met: (a) a large difference in dry day frequency between GCM and OBS and (b) high dry day frequency (~80% or greater). As

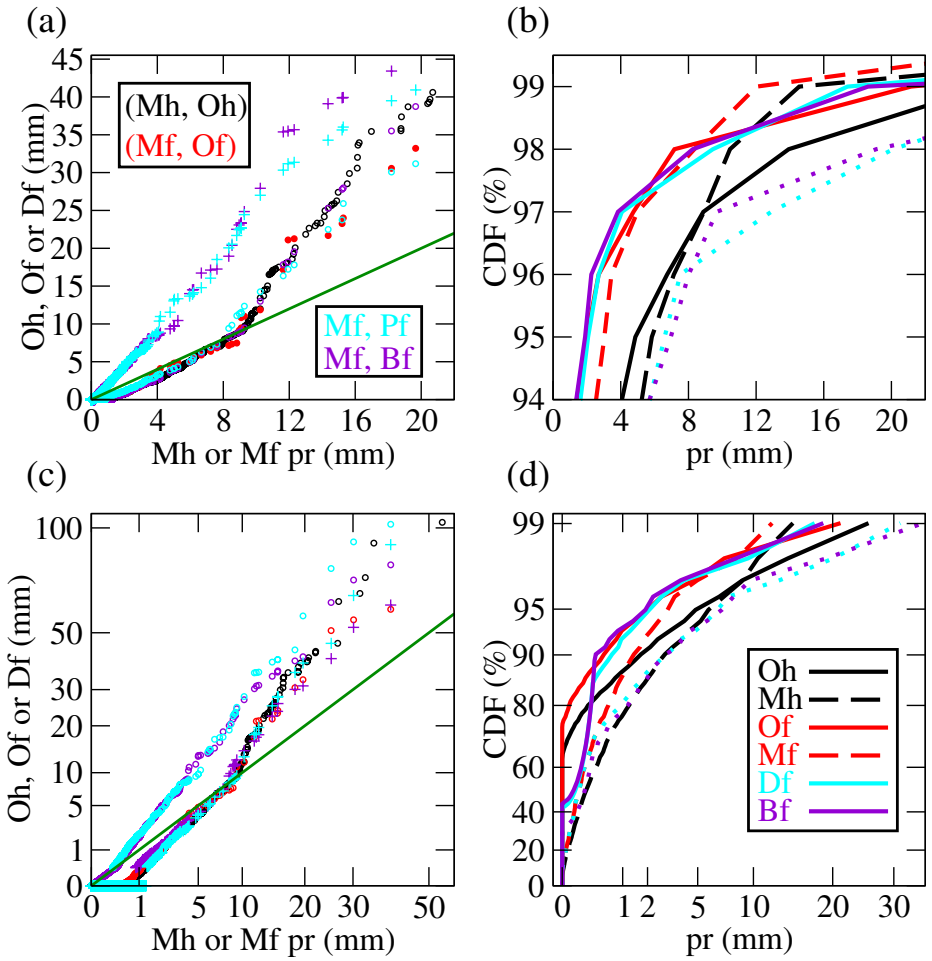

FIGURE 6 Same as Figure 4 except for Alpena, KS (39.9°N, 99.8°W; member one during JJA; ~730 m). In addition, symbols for (M_f, D_f) consist of violet circles for B4, violet plus signs for B1, and cyan plus signs for PRAT run with the same configuration as B1; note that the latter configuration (cyan plus signs) does not appear in Table 1 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 5 Climatological daily dry day frequency (%) at Alpena, KS (39.9°N, 99.8°W; member one during JJA) for OBS, GCM, and two configurations of BCQM (B1 and B4) during the historical and future periods

	OBS	GCM	B1	B4
Historical	81.1	57.2		
Future	87.4	69.1	64.8	89.5

shown in Table 5, at Alpena during both time periods condition (a) is met with a disparity of ~18–24%. Differences ~10–20% are common at a majority of gridpoints, and in some locations/seasons are as high as ~40–75%. Condition (b) is met since more than 80% of the days are dry for OBS in both the historical and future periods.

The reason for the problem (B1) is that when (a) is met mapping occurs between disparate portions of the OBS and GCM distributions. Condition (b) exacerbates the problem by reducing the sample used to define the distributions. For example (Table 5), at Alpena the upper 42.8% of the M_h distribution is mapped to the upper 18.9% of the O_h distribution with ignore0 on and frequency adjustment off. However, if frequency adjustment was invoked with ignore0 on, the upper 18.9% of both distributions would be used in the mapping. Finally, if both options were off then 100% of both distributions would be used in the mapping. As seen in Table 5, an additional consequence of the more equitable mapping (B4) is the good representation of precipitation frequency (89.5 vs. 87.4) as opposed to the mismatched case (B1) where a large DBIAS remains (64.8 vs, 87.4).

We can visualize the mechanisms for the poor performance of B1 via the qq plot for Alpena (Figure 6a,c). The SD 'training relationship' (black) is such that for the upper portion of the distribution the ratio of O_h to M_f monotonically increases. This ratio diminishes at the lower end to below 1 as per the DBIAS. Because there are many more wet days for M_h than O_h (Table 5), as per condition (a) above, during training, values of M_h are systematically mapped to inappropriately large values of O_h, for which the O_h/M_h ratio is biased too large. Consequently SD systematically overestimates precipitation for a given value of M_f. Note how this overestimation is similar to what was seen above at Jupiter (3.5.1), although

the ultimate cause was different (climate change, rather than a mismatch in precipitation frequency).

4 | DISCUSSION AND CONCLUSIONS

We have compared a number of distinct distributional downscaling methods as applied to daily precipitation in a Perfect Model context as a follow-up to our recent studies involving daily maximum temperature (L19a; L19b). Applying a more stringent metric (MAE) geared towards assessing agreement in day-to-day variability yields skill ~20–25% which is barely half of that found in earlier studies for temperature. Because of the more stochastic nature of precipitation (Maraun, 2013) we emphasize results based on a metric that assesses agreement of distributions (MAE-ord). By this metric skill is ~50–60% overall and about half of that in the right tail. This is distinctly different than for temperature for which skill in the tails could be boosted comparable to that for the remainder of the distribution (L19a; L19b).

Although downscaling overall yields useful MAE-ord skill (~30–35%) for values in the right tail of the distribution, there are considerable seasonal and regional variations. More importantly, even when skill is attained the magnitude of the errors can be considerable, for example ~15–25 mm in the southeastern U. S. during summer. We remind the reader that our Perfect Model design is somewhat idealized—accounting only for the mismatch in spatial scale between OBS and GCM but not for differences in the underlying climate states – thus, real-world downscaling performance may differ.

Compared with temperature, downscaling of precipitation via distributional methods is more complex having more configuration choices. Certain of these may be more consequential than the choice of SD method. These configuration choices result from the fact that precipitation consists of two aspects: (a) binary occurrence of precipitation (dry vs. wet days) and (b) distribution of precipitation conditional on occurrence. An equivalent but simplifying approach is to treat dry days as having zero precipitation, yielding a single distribution. Ultimately how these zero values are handled is crucial.

In our PM framework the poorest performance occurs when SD methods train and apply a transfer function to only the non-zero daily precipitation values, without a frequency adjustment to account for the DBIAS. However, the use of a frequency adjustment when downscaling only the non-zero values largely remedies the situation. The best configuration occurs when downscaling all values (zeros included) without a frequency adjustment. With regard to the use of configuration options our findings are in general agreement with V16.

Using optimal configurations, comparisons between several different downscaling methods do not always yield conclusive differences. PresRat, which involves a tweak to QDM was found to be comparable to BCQM while KDDM and QDM were found to be comparable to one another; the former pair were deemed marginally statistically significantly better than the latter pair.

For diagnostic purposes we have provided some examples which highlight the mechanisms responsible for good or bad performance in our PM framework. Poor performance can result from non-stationarity introduced via climate change, as was the case at Jupiter. Surprisingly, especially good performance can result when non-stationarity due to climate change by chance compensates for a deficiency in the downscaling method (Coaldale)—thus, two wrongs can make a right. Finally, when excluding all dry days from the SD, we demonstrated that for locations (such as Alpena) having infrequent precipitation (typically less than 20% of the days) the nearly ubiquitous DBIAS leads to an inappropriate mapping between the OBS and GCM distributions leading to very poor performance.

Our case studies identified an intrinsic weakness of distributional methods when applied to precipitation. Because of the DBIAS, for low values of precipitation the ratio of O_h to M_h will be less than 1. However, for larger values, often the bulk of the distribution, typically this ratio will be greater than 1 because of the spotty nature of precipitation and the larger GCM footprint. This effect is accentuated in convective regimes. Furthermore, this ratio often increases with increasing precipitation amount, again due to convection which tends to produce greater, more isolated bullseye values.

An inherent weakness of the class of quantile mapping methods is that while distributional methods operate via mappings between *relative positions* within distributions, there are certain physical constraints that operate with regard to *absolute amounts* of precipitation via spatial scale. A perturbing factor such as climate change (Jupiter) or excessively infrequent precipitation (Alpena) can distort the mapping in a manner that yields a physically inconsistent mapping—that is, where the DBIAS and its inverse get mapped to each other. Other perturbing factors, which have yet to be identified, may exist as well. The underlying characteristics of precipitation which often lead to poor results are not inherent to other better behaved variables such as temperature.

It is intriguing that for some seasons and locations errors in the right tail can be quite large even when downscaling has demonstrable skill. One wonders what

affect these errors might have on extreme value analysis (EVA) of precipitation, which has frequently been applied to raw GCM output? Recently Lopez-Cantu *et al.* (2020) performed EVA on CONUS precipitation and found large differences among five downscaled datasets. In future work we intend to explore this issue in a PM context.

Finally, as a bridge back to our earlier PM SD evaluations for daily maximum temperature (tasmax), which were based solely on MAE skill, here we have computed MAE-ord skill for tasmax for a limited number of cases from L19b. In this comparison we report only the averages of three SD methods that correspond most closely to B4, Q6 and K5. For the basic approach for CAT5, going from MAE to MAE-ord skill (%) increases from ~42 to 67 for tasmax compared with ~22 to 58 for precipitation (see Table 1). Using the LIM adjustment and averaging over the tail (CAT6-9) skill increases from ~46 to 57 for tasmax compared with approximately −8 to 32 for precipitation. Hence, the improvement in skill based on MAE-ord over that for MAE is greater for precipitation than temperature as expected given the more stochastic nature of the former. Furthermore, the improvement in the tails is much greater for precipitation, even though tail performance is much poorer compared with overall (CAT5) performance for precipitation than temperature.

ACKNOWLEDGEMENTS

We thank Andrew Ross and Adrienne Wootten for comments on an earlier draft of this manuscript. We are grateful to the anonymous reviewers for their many insightful comments which have led to a substantial improvement of this manuscript.

ORCID

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

REFERENCES

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.

Chen, C.T. and Knutson, T. (2008) On the verification and comparison of extreme rainfall indices from climate models. *Journal of Climate*, 21(7), 1605–1621. https://doi.org/10.1175/2007JCLI1494.1.

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.

Giorgi, F., Torma, C., Coppola, E., Ban, N., Schar, C. and Somot, S. (2016) Enhanced summer convective rainfall at alpine high elevations in response to climate warming. *Nature Geoscience*, 9 (8), 584–589. https://doi.org/10.1038/ngeo2761.

Hall, A. (2014) Projecting regional change. *Science*, 346(6216), 1461–1462. https://doi.org/10.1126/science.aaa0629.

Holtanova, E., Mendlik, T., Kolacek, J., Horova, I. and Miksovsky, J. (2019) Similarities within a multi-model ensemble: functional data analysis framework. *Geoscientific Model Development*, 12 (2), 735–747. https://doi.org/10.5194/gmd-12-735-2019.

Huang, J., van den Dool, H. and Georgakakos, K. (1996) Analysis of model-calculated soil moisture over the United States (1931-1993) and applications to long-range temperature forecasts. *Journal of Climate*, 9(6), 1350–1362. https://doi.org/10.1175/1520-0442(1996)009<1350:AOMCSM>2.0.CO;2.

IPCC. (2013) Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. In: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V. and Midgley, P.M. (Eds.) . Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, p. 1535.

Karmalkar, A.V. (2018) Interpreting results from the NARCCAP and NA-CORDEX ensembles in the context of uncertainty in regional climate change projections. *Bulletin of the American Meteorological Society*, 99(10), 2093–2106. https://doi.org/10.1175/BAMS-D-17-0127.1.

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%3C1197::AID-JOC89%3E3.0.CO;2-L.

Lanzante, J.R., Adams-Smith, D., Dixon, K.W., Nath, M.J. and Whitlock, C.E. (2019b) Evaluation of some distributional downscaling methods as applied to daily maximum temperature with emphasis on. *International Journal of Climatology*, 40(3), 1571–1585. https://doi.org/10.1002/joc.6288.

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. (2019a) Evaluation and improvement of tail behaviour in the cumulative distribution function transform downscaling method. *International Journal of Climatology*, 39(4), 2449–2460. https://doi.org/10.1002/joc.5964.

Lopez-Cantu, T., Prein, A.F. and Samaras, C. (2020) Uncertainties in future U.S. extreme precipitation from downscaled climate projections. *Geophysical Research Letters*, 47, e2019GL086797. https://doi.org/10.1029/2019GL086797.

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., Shepherd, T., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J.M., Hagemann, S., Richter, I., Soares, P.M.M., Hall, A. and Mearns, L.O. (2017) Towards process-informed bias correction of climate change simulations. *Nature Climate Change*, 7(11), 764–773. https://doi.org/10.1038/nclimate3418.

Maraun, D. and Widmann, M. (2018) *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge, UK: Cambridge University Press.

Maraun, D., Widmann, M. and Gutierrez, J.M. (2019) Statistical downscaling skill under present climate conditions: a synthesis of the VALUE perfect predictor experiment. *International Journal of Climatology*, 39(9), 3692–3703. https://doi.org/10.1002/joc.5877.

McGinnis, S., Nychka, D. and Mearns, L. (2015) A new distribution mapping technique for climate model bias correction. In: Lakshmanan, V., Gilleland, 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.

Pendergrass, A.G. and Knutti, R. (2018) The uneven nature of daily precipitation and its change. *Geophysical Research Letters*, 45 (21), 11980–11988. https://doi.org/10.1029/2018GL080298.

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.

Richman, M. (1986) Rotation of principal components. *Journal of Climatology*, 6(3), 293–335. https://doi.org/10.1002/joc.3370060305.

Stephens, G.L., L'Ecuyer, T., Forbes, R., Gettleman, A., Golaz, J.-C., Bodas-Salcedo, A., Suzuki, K., Gabriel, P. and Haynes, J. (2010) Dreary state of precipitation in global models. *Journal of Geophysical Research-Atmospheres*, 115(D24), 1–14. https://doi.org/10.1029/2010JD014532.

Vrac, M., Noel, T. and Vautard, R. (2016) Bias correction of precipitation through singularity stochastic removal: because occurrences matter. *Journal of Geophysical Research-Atmospheres*, 121(10), 5237–5258. https://doi.org/10.1002/2015JD024511.

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

Zhang, X., Zwiers, F.W. and Hegerl, G. (2009) The influences of data precision on the calculation of temperature percentile indices. *International Journal of Climatology*, 29(3), 321–327. https://doi.org/10.1002/joc.1738.

How to cite this article: Lanzante JR, Dixon KW, Adams-Smith D, Nath MJ, Whitlock CE. Evaluation of some distributional downscaling methods as applied to daily precipitation with an eye towards extremes. *Int J Climatol*. 2021;41: 3186–3202. https://doi.org/10.1002/joc.7013