


RESEARCH ARTICLE

Evaluation and improvement of tail behaviour in the cumulative distribution function transform downscaling method

John R. Lanzante¹  | Mary Jo Nath¹ | Carolyn E. Whitlock^{1,2} | Keith W. Dixon¹ | Dennis Adams-Smith^{1,3}

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

²Engility Inc., Dover, New Jersey

³University Corporation for Atmospheric Research (UCAR)/Cooperative Programs for the Advancement of Earth System Science (CPAESS), Boulder, Colorado

Correspondence

John R. Lanzante, NOAA/Geophysical Fluid Dynamics Laboratory, 201 Forrestal Road, Princeton, NJ 08540-6649.
Email: john.lanzante@noaa.gov

Funding information

United States Geological Survey, South Central Climate Adaptation Science Center, Grant/Award Number: G13AC00387

The cumulative distribution function transform (CDFt) downscaling method has been used widely to provide local-scale information and bias correction to output from physical climate models. The CDFt approach is one from the category of statistical downscaling methods that operates via transformations between statistical distributions. Although numerous studies have demonstrated that such methods provide value overall, much less effort has focused on their performance with regard to values in the tails of distributions. We evaluate the performance of CDFt-generated tail values based on four distinct approaches, two native to CDFt and two of our own creation, in the context of a “Perfect Model” setting in which global climate model output is used as a proxy for both observational and model data. We find that the native CDFt approaches can have sub-optimal performance in the tails, particularly with regard to the maximum value. However, our alternative approaches provide substantial improvement.

KEYWORDS

bias correction, distributions, Perfect Model evaluation, statistical downscaling, tail values

1 | INTRODUCTION

As societal actions to curb the forcing agents responsible for anthropogenic climate change (ACC) have failed to keep pace with those generally deemed necessary to limit the most adverse effects (Nature Editorial, 2018), increasing attention is being devoted to the development of plans aimed at adapting to changing climate, both reactively and proactively (Bierbaum *et al.*, 2014). The effects of ACC cut across diverse sectors such as agriculture, human health, water resources, transportation, energy, and ecosystems, to name just a few. While ACC is global in nature, adaptation typically occurs on much smaller spatial scales. In devising coping strategies policymakers often rely on specific local information. Although global climate models (GCMs) are the most useful resource for projecting the future effects of ACC, they have limitations in resolving more detailed local

effects. As a result, a wide variety of techniques have been developed to provide finer details via “downscaling.” In this paper we examine the performance of one such method in producing extreme values.

The cumulative distribution function transform (CDFt) statistical downscaling method (Michelangeli *et al.*, 2009) is one in a category of approaches based on distributional transformations and has been utilized in numerous studies (e.g., Oettli *et al.*, 2011; Colette *et al.*, 2012; Lavaysse *et al.*, 2012; Tisseuil *et al.*, 2012; Vrac *et al.*, 2012; 2016; Flaounas *et al.*, 2013; Vautard *et al.*, 2013; Vigaud *et al.*, 2013; Famién *et al.*, 2017) as well as commercially (<https://theclimate-datafactory.com>). The transformations utilized by the methods in this category typically involve the frequency distributions of observations in a historical period (O_h), model values in a historical period (M_h), and sometimes model values in a future period (M_f). The result is the generation of

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2018 The Authors. International Journal of Climatology published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.

downscaled and bias-corrected values for a future period (D_f). The motivation for applying these methods includes the reduction of systematic model biases and/or rendering information on a smaller spatial scale than provided by physical models such as GCMs, regional climate models (RCMs), or reanalysis systems. Attaining the latter goal is more challenging (Maraun, 2013).

Although downscaling methods in this class have been found to be broadly useful in an overall sense (Fowler *et al.*, 2007; Maraun *et al.*, 2010; Gutierrez *et al.*, 2018) far less attention has been given towards evaluations of the tails of the distributions (e.g., Kallache *et al.*, 2011; Gutmann *et al.*, 2014; Cannon *et al.*, 2015). Ironically, much of the focus in assessing impacts of climate change is focused on extremes, which in many applications are considered to have a disproportionately large impact on natural systems.

In this paper we assess the tail behaviour of data downscaled using the CDFt method. This is part of a broader ongoing effort by the GFDL (Geophysical Fluid Dynamics Laboratory) Empirical Statistical Downscaling (ESD) team (https://www.gfdl.noaa.gov/esd_eval) and collaborators aimed at the evaluation of statistical downscaling methods. While this project represents the initiation of efforts aimed at evaluating the tails of distributions of CDFt downscaled output, the immediate follow-up will consider more than a half dozen of the commonly used distributional methods. The reasons for first focusing on the CDFt method are threefold: (a) in our early applications of the method we visually noticed a tendency for it to produce outliers, (b) the mechanics of the method overall are in some sense fundamentally different from the other methods in its category, and (c) one of CDFt's options for producing tail values is novel.

Here we expand on some of the unique aspects of the CDFt method [(b) and (c) above]. Most of the methods in this category can be considered quantile mapping methods in which quantiles and their associated inverses (i.e., values of their cumulative distribution functions [CDFs]) are cross-evaluated in the distributions of O_h , M_h , and M_f . Thus, a single value to be downscaled (M_f) directly yields a downscaled value (D_f) based on a predetermined set of mappings, which vary from method to method. Illustrations of these mappings for some of the common methods are found in Pierce *et al.* (2015). On the other hand, the CDFt method first empirically estimates, via transformations, the downscaled CDF and then evaluates it for the specific set of M_f values to be downscaled. The downscaled CDF is estimated by way of a set of nested transformations, first from M_f to M_h , and then from this result to O_h . Mathematically this is expressed as (Michelangeli *et al.*, 2009):

$$F_{D_f}(x) = F_{O_h}\left(F_{M_h}^{-1}\left(F_{M_f}(x)\right)\right), \quad (1)$$

where F represents the CDF and F^{-1} its inverse.

Another unique aspect of CDFt in its default configuration is the manner in which it generates tail values in some

situations. For some distributional methods there are “out-of-bounds” (OOB) conditions for which the base algorithm is unable to produce downscaled values, starting from some point in the CDF out to the end of the tail. For any given case this may occur for either, both or neither tails. An explanation for the CDFt OOB condition is given by Pierce *et al.* (2015) and illustrated in their Figure 1c. In such a case some other approach is used—the most common (Deque, 2007, hereinafter D07) is based on a constant correction—to which Pierce *et al.* (2015) provide a cautionary note regarding its utility. In this approach the difference, $D_f - M_f$ for the “last good value,” that is, the value farthest out in the tail for which the base algorithm is able to perform the downscaling, is used as a constant correction factor. This difference is added to each M_f value that has been deemed OOB to produce a corresponding D_f value. Although this simple scheme can often yield reasonable values, in some cases it can produce highly erroneous results (Lanzante *et al.*, 2018). However, as we describe below, the CDFt method has a very different and unique manner in which to deal with OOB conditions.

Here we evaluate several alternative schemes for producing CDFt tail values. While some of the approaches have been used before, others are our creations. We demonstrate the superiority of the new methods that we introduce.

2 | DATA AND METHODOLOGY

2.1 | Data

We employ a “Perfect Model” (PM) framework in which GCM data serve as a proxy for both observations and model. The advantage is that the PM provides not only data for a historical period, but also for a future state affected by considerable climate change. As such, the presence of future observations, which are not available in traditional retrospective studies, allows for a more rigorous evaluation via potential violation of the “stationarity assumption” (Dixon *et al.*, 2016) implicit to all statistical downscaling methods. Recent work (Lanzante *et al.*, 2018) highlights the value of this approach by demonstrating that some serious problems that appear in PM evaluations of future climate cannot be detected in retrospective studies involving only historical data. The Perfect Model input data (O_h , M_h , and M_f) along with the validation data (O_f) are available from the GFDL ftp site as detailed in Appendix A.

Here we only briefly describe our PM data; the interested reader is referred to Dixon *et al.* (2016) for more details. Historical data (O_h and M_h) for daily maximum temperature (T_{\max}) correspond to the time period 1979–2008. Future data (O_f and M_f), based on a high-climate-sensitivity model under a high emissions scenario (RCP8.5), correspond to the time period 2086–2095 and consist of three ensembles of 10 years each. We downscale each ensemble separately and average

verification statistics over the three members. The “observations” are simply the raw gridpoint values produced by the GCM whereas the “model data” are the raw values which have been spatially smoothed, yielding a mismatch in spatial resolution typical of those found in real-world applications of downscaling (~ 25 km vs. ~ 200 km). The spatial domain (Figure 2; Dixon *et al.*, 2016) is a rectangular area centred on the conterminous 48 United States; here we exclude oceanic points (Pacific and Atlantic, as well as the Gulf of Mexico). In this paper we generically refer to O_h and O_f as “observations” and M_h and M_f as “model” or “GCM,” which is appropriate in the PM world, even though all data values are technically derived from a GCM.

2.2 | Downscaling methodology

Our implementation of CDFt is based on code in the R-language CRAN repository (<https://CRAN.R-project.org/package=CDFt>) which is the source cited in the paper that introduced CDFt (Michelangeli *et al.*, 2009). One curious aspect is that the scheme used in the code to handle OOB conditions is not documented in the paper, or in any other source that we could find. The scheme is unique in that instead of a simple ad-hoc correction it involves “slicing off” the tail of the O_h distribution and appending it to the end of the D_f distribution. The slice point is the value of the CDF beyond which OOB conditions occur. We note that a number of subsequent papers by the original authors and collaborators cite the use of the more common D07 (i.e., constant correction) approach to handle OOB conditions.

In this paper we are concerned with evaluation of the tails of the output from CDFt based on several different tail treatments: (a) the base CDFt algorithm (i.e., that is used in the absence of OOB conditions), which we refer to as the CDFt base algorithm (CBA), (b) the CDFt internal tail scheme (CIT) based on appending part of the O_h distribution to the D_f distribution, (c) the constant correction approach (D07) and our own variants of it, which we refer to as simple tail adjustment (SIM), and (d) our own approach which we refer to as limited tail adjustment (LIM). Our codes for the SIM and LIM tail schemes are provided in Supporting Information along with some sample inputs and outputs. Given the complexity of our nomenclature, as an aid to the reader, the most common shorthand notions introduced to this point as well as subsequently are detailed in Table 1.

In the case of (c) we have extended the original D07 algorithm via the introduction of a parameter “lastN-points” (NPT). Instead of using a single value, that is, “the last good point” to determine the additive correction factor, we allow for an arbitrary number of points. Thus, a correction factor is computed individually for each of NPT points and then averaged, to produce the additive offset. The reasoning is that the average of several values should produce a result that is

more representative of the difference between the M_f and D_f CDFs at the boundary of OOB conditions.

As an example of SIM adjustment suppose that we have a set of $M_f(i)$ values, sorted from low to high, where $i = 1, n$ from which we wish to generate a corresponding set of $D_f(i)$ values. Also, suppose that the first k values of $M_f(i)$ have been deemed as OOB (i.e., less than the minimum M_h value) so that the corresponding values of $D_f(i)$ are undefined. Given the user selected value of NPT we compute a correction factor by averaging the differences between corresponding values of D_f and M_f :

$$\Delta = \sum_{i=k+1, k+NPT} [D_f(i) - M_f(i)] / NPT, \quad (2)$$

The downscaled values for $i = 1, k$ are:

$$D_f(i) = M_f(i) + \Delta. \quad (3)$$

While this example is illustrative for the left tail, the procedure for the right tail is analogous. Note that in the original D07 algorithm $NPT = 1$ always.

Our LIM adjustment borrows one of the aspects of SIM adjustment in that it utilizes the parameter NPT. But in addition it introduces a parameter tail-length (TLN) which is a user-defined tail length. In the case of SIM the values to be adjusted are determined by the CDFt OOB conditions. If no values are OOB then SIM cannot be applied. On the other hand, LIM ignores the OOB conditions and always performs adjustment to a tail of length TLN—therefore when LIM is used both tails are always adjusted as specified by the parameters. For SIM, either or both of the left and right tails may or may not be adjusted depending on OOB conditions in each tail. In summary, LIM adjustment is carried out in the same fashion as SIM adjustment except that instead of the value of k being determined by the OOB conditions the user selects the tail length $k = TLN$.

2.3 | Evaluation approach

Application of downscaling is performed separately for each gridpoint, each of 12 months, and each of three 10-year future ensembles. We compute the mean absolute error (MAE), using the biweight mean (Lanzante, 1996), which is resistant to outliers, for both a given downscaling treatment (D_f) as well as for the raw GCM (M_f). We average the MAE over all months and ensembles and then combine the MAE values for D_f and M_f yielding the skill (Wilks, 2006) which is computed with respect to M_f :

$$\text{Skill} = [(MAE_{M_f} - MAE_{D_f}) / MAE_{M_f}] \times 100\%. \quad (4)$$

Finally, we average skill over all non-ocean gridpoints.

We report the area-averaged skill scores separately for each of nine distributional categories (CAT1–CAT9) and each treatment. The distributional categories represent portions of the tail of a given distribution (i.e., order statistics). Specifically, CAT1 (CAT9) corresponds to the lowest

TABLE 1 Shorthand notation used in this paper

Shorthand in text	Description	Portion of distribution applicable to	Figure key shorthand
CBA	CDFt base algorithm	All non-OBB points	C0
CIT	CDFt internal tail adjustment	Only OBB points	C1
SIM	Simple tail adjustment	Only OBB points	SIM_{NPT}
LIM	Limited tail adjustment	TLN number of points in each tail	LIM_{NPT}_{TLN}
OOB	Out of bounds. Points for which the CDFt method determines the CBA 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		SIM_{NPT}; LIM_{NPT}_{TLN}
TLN	“Tail length” (applicable only to LIM) Number of tail points to be adjusted regardless of whether the CDFt method considers them to be OBB or not		LIM_{NPT}_{TLN}
Eval criterion “0”	Evaluation computed using only non-OBB points	All non-OBB points	
Eval criterion “1”	Evaluation computed using only OBB points	Only OBB points	
Eval criterion “2”	Evaluation computed using all points (i.e., union of non-OBB “0” and OBB “1” points)	All points	
CAT1	Minimum value at end of left tail	Left tail	
CAT2	Second and third smallest values	Left tail	
CAT3	Fourth, fifth, and sixth smallest values	Left tail	
CAT4	Seventh, eighth, ninth, and tenth smallest values	Left tail	
CAT5	All points in the distribution	All points in the distribution	
CAT6	Seventh, eighth, ninth, and tenth highest values	Right tail	
CAT7	Fourth, fifth, and sixth highest values	Right tail	
CAT8	Second and third highest values	Right tail	
CAT9	Maximum value at end of right tail	Right tail	

(highest) value in the sample, CAT2 (CAT8) the second–third lowest (highest), CAT3 (CAT7) the fourth–sixth lowest (highest), and CAT4 (CAT6) the seventh–tenth lowest (highest) values. Although not the focus of this paper, all values in the sample are included in CAT5 as a point of interest. Some of our results are presented as averages over all or most of the categories (weighted by number of values each category), excluding CAT5.

Our evaluations are geared towards treatments, which we define as the combination of a particular downscaling approach (CBA, CIT, SIM, or LIM) and an evaluation criterion (EVAL). Treatments involving SIM or LIM have associated parameter values (NPT for SIM; NPT and TLN for LIM). We have three evaluation criteria designated numerically by 0 (not OOB), 1 (OOB), and 2 (all). Thus, we can separately evaluate treatments for instances in which CDFt did (1) or did not (0) designate some of the tail values as being OOB, or we can ignore such designations and use all values (2) regardless of whether they were OOB or not. Note that because the existence and number of OOB values varies from case to case, the sample sizes for the different categories vary.

We adopt a short-hand to designate treatments using an alphanumeric sequence in which the first character is a letter (C for CDFt, S for SIM, and L for LIM) and the second an evaluation criterion (0, 1, or 2). For SIM and LIM, after the

first two characters is an underscore, followed by the two-digit value of NPT, and for LIM another underscore followed by the two-digit value of TLN. For example, C0 for CBA; C1 for CIT; S1_05 for SIM, with EVAL = 1 and NPT = 5; L2_10_20 for LIM with EVAL = 2, NPT = 10 and TLN = 20.

We have generated results for values of 1, 3, 5, 10, and 20 for NPT and 1, 10, and 20 for TLN. Economy of presentation is facilitated by the smooth, monotonic and sometimes slight dependence of skill on these parameters. Additionally, some combinations are nonsensical (e.g., CBA with EVAL = 1 and CIT with EVAL = 0 are null sets).

3 | RESULTS

3.1 | Skill by distributional category and treatment

Figure 1 displays the skill as a function of distributional category where each curve represents a different treatment; line patterns indicate downscaling method and colours indicate evaluation criteria. Ignoring for the moment CAT5 (All) and CAT9, we see that the poorest results are for CIT (solid red, labelled C1) with negative or relatively small positive skills. By comparison, CBA (solid cyan, labelled C0) yields considerably better results. This indicates that the undocumented “slicing” method performs poorly and perhaps is a reason

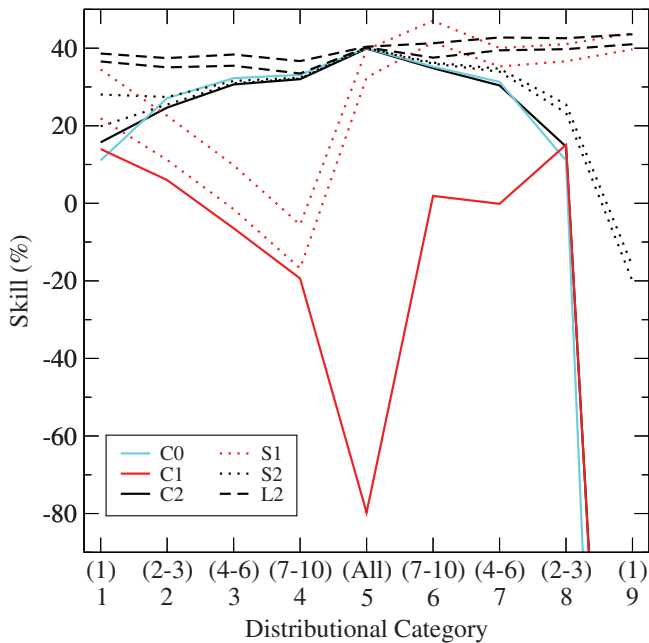


FIGURE 1 Skill (%) as a function of distributional category for various downscaling treatments: C0 (solid cyan), C1 (solid red), C2 (solid black), S1₀₁/S1₁₀ (lower/upper dotted red), S2₀₁/S2₁₀ (lower/upper dotted black), and L2₀₁₁₀/L2₁₀₁₀ (lower/upper dashed black). The abscissa axis label gives the distributional category number (1–9) along with the corresponding ordinal point numbers in parentheses on the left/right indicating the first (1), second–third (2–3), fourth–sixth (4–6), and seventh–tenth (7–10) lowest/highest values from the sample. The “All” category includes all available values for the given treatment (i.e., non-OOB values when EVAL = 0, OOB values when EVAL = 1, and all values when EVAL = 2). The skill shown is the biweight mean over all non-ocean gridpoints, three ensembles, and 12 months [Colour figure can be viewed at wileyonlinelibrary.com]

why a number of authors employed the D07 approach. The two dotted red curves correspond to SIM with NPT = 1 and 10, the former of which is the D07 approach. Comparing the dotted and solid red curves it is apparent that the constant correction approach is uniformly superior. A further noteworthy point is that our modification, that is, the introduction of the NPT parameter, here using a value of 10, also yields a modest improvement over D07.

The performance of CDFt for CAT9, that is, the maximum value, is noticeably different from all other results as very large negative skills are found for both CIT and CBA (~–289 and ~–387, respectively, not shown). The fact that this behaviour occurs even when OOB conditions are not experienced (CBA, solid cyan line) suggests that there is a problem with the CDFt code specific to the maximum value, separate from any issues related to the tail scheme. However, regardless of the root cause the alternative tail schemes, SIM and LIM, are able to remedy the problem with CAT9 as seen by the fact that for S1 and L2 skill for CAT9 is comparable to that for other points in the right tail. Note that for S2 CAT9 skill is depressed because it is based on C0 for non OOB conditions as well as S1 for OOB conditions.

The solid black curve, based on CDFt for all tail values (OOB and not OOB; the union of CIT and CBA), is not much different from the solid cyan curve, based on CDFt not OOB because the number of OOB cases is generally in the minority. This can be seen in Table 2, which gives the frequency of occurrence of the invocation of the tail schemes. In the more interior categories where OOB is relatively rare, the base algorithm dominates, but at the very ends of the distribution (CAT1, CAT8, and CAT9) the internal scheme is invoked much more frequently (and likewise SIM) due to the dominance of OOB conditions. Nevertheless, CDFt overall performs very poorly for the maximum value (~–288, not shown). In comparison, LIM with TLN = 10 and NPT = 1 (bottom) and 10 (top) dashed black curves indicate an improvement, moderate for most categories but very large for the maximum. In addition, SIM evaluated for all tail values (black dotted curves) lies between CDFt and LIM, with the biggest advantage over CDFt for the maximum.

The results for CAT5 (All) must be interpreted with caution. They represent skill for *all* values in the distribution only for the black curves labelled C2 in Figure 1. However, most of the other treatments for CAT5 have nearly the same high level of skill, the exception being CIT, which is much lower. It is noteworthy that skill for tail values for CBA and especially CIT is considerably less than overall skill (CAT5). The fact that LIM alone yields skill values for all categories comparable to that for CAT5 suggests that the LIM approach is a remedy for the shortcomings of CDFt in the tails.

3.2 | Frequency distributions of daily errors

Figure 1 clearly demonstrates the serious problem CDFt has in downscaling the maximum value (CAT9) and that both of our alternate schemes (SIM and LIM) ameliorate this. In Figure 2 we show histograms of signed, daily error for some of the treatments. In so much as skill is a function of the error of a downscaling scheme as compared to the error from the GCM (without any downscaling), each panel displays the errors for both the GCM (black, shaded grey) and a downscaling approach (red). The top row presents error distributions for all tail distributional categories excluding CAT9. While the distributions for downscaling and GCM are roughly comparable, the degree to which the former distribution lies inside the latter is indicative of the skill. The

TABLE 2 Percentages of tail adjustment type (CBA or CIT) by distributional category

Category	1	2	3	4	6	7	8	9
Rank	1	2–3	4–6	7–10	7–10	4–6	2–3	1
CBA	38.2	83.2	94.7	97.3	94.1	84.4	53.5	16.2
CIT	61.8	16.8	5.3	2.7	5.9	15.6	46.5	83.8

Note. These occurrences are based on aggregating results over all gridpoints, three ensembles, and 12 months, for a total of 549,468 cases for each rank.

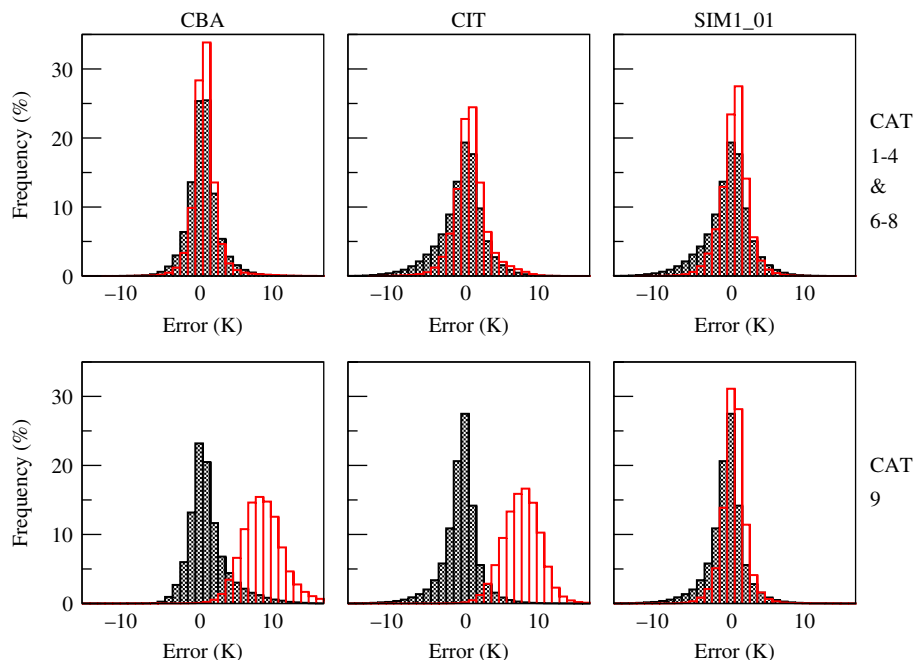


FIGURE 2 Histograms of daily error (K) for selected distributional categories and treatments. Top (bottom) for categories 1–4 and 6–8 (9). Left for CBA, middle for CIT, and right for SIM1_01. Errors for downscaling are red and for GCM black, shaded grey [Colour figure can be viewed at wileyonlinelibrary.com]

results for CAT9 (bottom) indicate that both CDFt tail schemes (CBA and CIT) produce large errors, almost all with positive sign, indicating that they systematically overestimate the maximum value. By contrast, even the poorest performing of our alternative schemes (SIM_01) corrects the problem and produces positive skill (Figure 1).

3.3 | Skill averaged over distributional categories

In examining the more detailed structure of skill by category in Figure 1 we found that differences between treatments are in many cases approximately constant across categories, the largest exception being for the maximum value. To further summarize results we examine overall differences between treatments with special emphasis on the effect of varying parameters NPT and TLN. As such, Figure 3 displays skill averaged over all categories, excluding in some cases CAT9 (and of course always excluding CAT5).

To aid in the examination of Figure 3 we utilize significance levels in Table 3 pertaining to tests of the difference in the mean skill levels between treatments. The rationale for and the details of the significance testing procedure are given in Appendix B with only a brief overview here. One of the main complications involves the fact that the mean skill levels in Figures 1 and 3 are derived by averaging the skill over all locations on skill maps containing up to ~15,000 gridpoints (or less depending on missing values). Since results at nearby gridpoints are not independent of one another we face a situation akin to, but more complex than the field significance issue raised in the seminal paper of Livezey and Chen (1983), hereafter referred to as LC83.

In order to account for the dependence between gridpoints we estimate significance via Monte Carlo simulations in which we treat a much smaller number of *blocks* of gridpoints

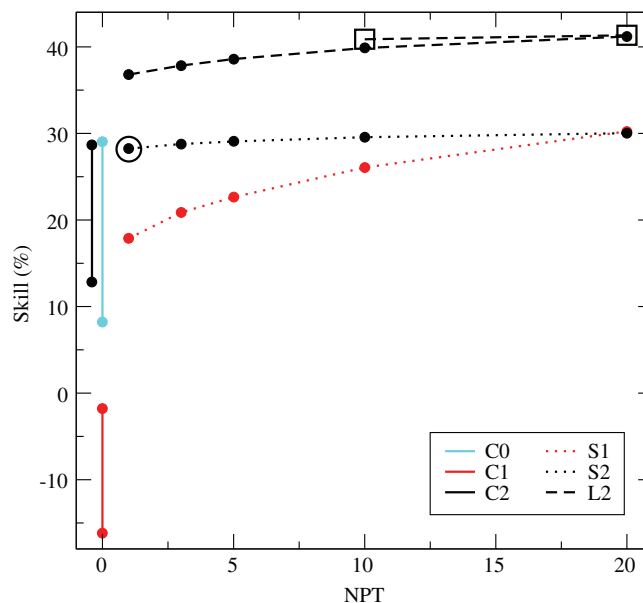


FIGURE 3 Skill (%) averaged over distributional categories 1–4 and 6–9 (except omitting category 9 where noted), weighted by the number of values in each category, as a function of NPT for various downscaling treatments: C0 (solid cyan), C1 (solid red), C2 (solid black), S1_01/S1_03/S1_05/S1_10/S1_20 (dotted red and filled circles), S2_01/S2_03/S2_05/S2_10/S2_20 (dotted black and filled circles), L2_01_01 (open black circle), L2_01_10/L2_03_10/L2_05_10/L2_10_10/L2_20_10 (dashed black and filled circles), and L2_10_20/L2_20_20 (open black squares). For C0/C1/C2 the lower (upper) filled circle includes (excludes) category 9. Note that CDFt (C0/C1/C2), which is not dependent on parameter lastNpts (NPT), is plotted arbitrarily for lastNpts = 0 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Significance level (%) testing the hypothesis that there is no difference in skill (as a weighted average over the specified distributional categories) for the two specified downscaling treatments

Nblon × Nblat	Probability (%)	Treatments	Categories
4 × 2, 8 × 4, 13 × 7	0.0, 0.0, 0.0	C0 × C1	1–4, 6–8
4 × 2, 9 × 5, 17 × 9	4.1, 0.0, 0.0	C1 × S1_01	1–4, 6–8
4 × 2, 9 × 5, 16 × 9	54.9, 21.1, 5.6	C1 × S1_01	1–4
4 × 2, 12 × 8, 22 × 12	0.1, 0.0, 0.0	C1 × S1_01	6–8
4 × 2, 7 × 5, 14 × 8	1.3, 0.0, 0.0	C1 × S1_03	1–4, 6–8
4 × 2, 5 × 4, 11 × 7	0.0, 0.0, 0.0	C1 × S1_10	1–4, 6–8
4 × 2, 5 × 3, 8 × 5	1.6, 0.8, 0.0	C2 × S2_01	1–4, 6–9
4 × 2, 6 × 4, 11 × 7	64.5, 62.1, 40.4	C2 × S2_01	1–4, 6–8
4 × 2, 9 × 5, 15 × 9	61.8, 37.1, 17.6	S1_01 × S1_03	1–4, 6–9
4 × 2, 9 × 5, 15 × 9	45.8, 15.4, 2.6	S1_01 × S1_05	1–4, 6–9
4 × 2, 8 × 5, 14 × 9	20.6, 3.7, 0.1	S1_01 × S1_10	1–4, 6–9
4 × 2, 7 × 5, 13 × 8	8.6, 0.5, 0.0	S1_01 × S1_20	1–4, 6–9
4 × 2, 9 × 6, 18 × 10	59.5, 30.8, 10.5	S1_10 × S1_20	1–4, 6–9
4 × 2, 5 × 3, 8 × 5	1.6, 0.8, 0.0	C2 × L2_10_10	1–4, 6–8
4 × 2, 6 × 3, 9 × 5	3.4, 1.0, 0.0	S2_10 × L2_10_10	1–4, 6–9
4 × 2, 5 × 3, 9 × 5	20.5, 17.4, 1.8	S2_10 × L2_10_10	1–4
4 × 2, 7 × 5, 12 × 8	1.5, 0.0, 0.0	S2_10 × L2_10_10	6–9
4 × 2, 9 × 6, 14 × 9	5.6, 0.0, 0.0	S2_10 × L2_10_10	6–8
4 × 2, 5 × 2, 10 × 5	0.0, 0.0, 0.0	S2_10 × L2_10_10	9
4 × 2, 8 × 4, 12 × 7	42.0, 27.1, 9.7	L2_01_10 × L2_10_10	1–4, 6–9
4 × 2, 8 × 4, 11 × 7	25.0, 9.6, 1.6	L2_01_10 × L2_20_10	1–4, 6–9
4 × 2, 7 × 5, 13 × 8	69.0, 58.9, 42.7	L2_10_10 × L2_20_10	1–4, 6–9
4 × 2, 6 × 4, 9 × 6	93.4, 97.4, 99.9	L2_01_01 × L2_01_10	1–4, 6–9

Note. For each comparison there are three probabilities (left–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. 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 Appendix B). Abbreviations for treatments as described in the text and Table 1.

as being independent of one another. For each comparison of treatments in Table 3 there are three significance level estimates based on different numbers of blocks. The first block size (4 × 2; 8 blocks with 4 [2] divisions in the longitudinal [latitudinal] direction), which is the same for all comparisons, is an extremely conservative subjective estimate which we believe is almost certainly far too small (and thus quite likely underestimates the significance). The other two estimates are based on a conservative and more liberal use of Monte Carlo simulation results. By way of three very different estimates spanning a fairly wide range of plausible solutions, we are able to demonstrate the robustness of our results.

We begin examination of Figure 3 with consideration of the results for CDFt (solid curves) with cyan for CBA, red for CIT, and black for overall. There are two points for each curve, the lower based on the weighted average of all eight distributional categories (excluding CAT5), and the higher excluding CAT9. We see that the skill level is considerably higher when we exclude the problematic maximum value. Note also that CBA has considerably higher skill than CIT and this difference is highly significant. This result is indicated in Table 3 by locating the appropriate row that lists treatments C0 (CBA) and C1 (CIT). The categories column lists 1–4 and 6–8, thus CAT9 has been excluded, so the results apply to the higher red dot (C1) and higher cyan dot (C0) in Figure 3. Given that the three values in Table 3 (0.0, a probability in percent rounded to one post-decimal place)

all indicate a high level of significance demonstrates the robustness of this conclusion.

We compare the D07 approach (S1_01), represented by the left-most point on the dotted red curve, with CIT (C1), the upper dot on the solid red curve. From Table 3 we see that even excluding CAT9 the difference between these two treatments is highly significant. Table 3 also shows separate results for the left (CAT1–CAT4) and right (CAT6–CAT8) tails, with the former at best marginally significant, but the latter highly significant. As seen in Figure 3 most of the benefit of S1_01 is in the upper tail.

Next we examine how varying NPT affects SIM. Moving from left to right along the dotted red curve there is a monotonic increase in skill, with a proportionately larger benefit for smaller values of NPT. Examining differences for SIM with various values of NPT in Table 3 we see that only the largest differences in NPT are significant. The difference for NPT 1 versus 10 is likely significant, but for 10 versus 20 the difference is not significant. Based on these results we consider a value of 10 for NPT to be a reasonable choice and yielding an improvement over the D07 approach.

If we examine the overall differences in SIM (dotted black) we see that there is very little variation in skill with variations in NPT, and in addition there is little difference between CDFt (upper dot, solid black) and SIM. Thus, the benefit of SIM over CDFt is mostly for OOB conditions. One exception to this is regarding CAT9, which when

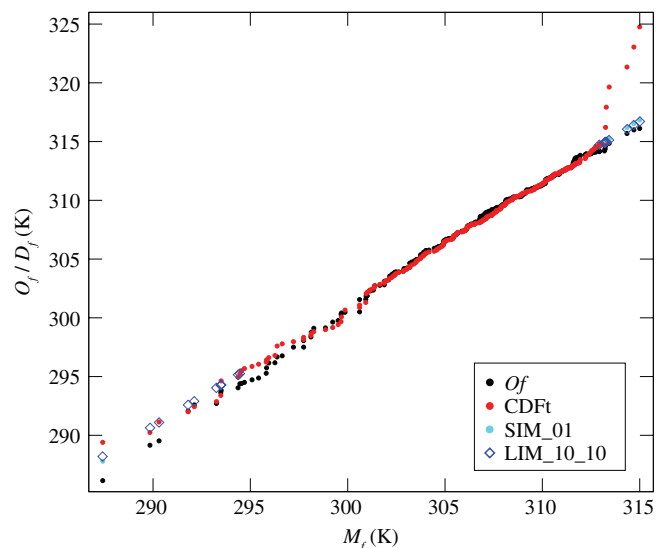


FIGURE 4 Quantile–quantile (q–q) plot for a gridpoint near Douglas, WY (105.2 W, 42.9 N) for ensemble member 2 for September. The abscissa is M_f and the ordinate is O_f (black filled circles), CDFt (red filled circles), SIM_01 (cyan filled circles), and LIM_10_10 (blue open diamonds). For SIM_01 and LIM_10_10 only the points that differ from CDFt are shown. These are 1 (10) in the left tail and 6 (10) in the right tail for the former (latter) [Colour figure can be viewed at wileyonlinelibrary.com]

included results in SIM being much better than CDFt for both CBA (lower dot, solid cyan), CIT (lower dot, solid red), and overall (lower dot, solid black).

Finally we examine the performance of LIM, represented in Figure 3 by the dashed black curve. The black dots represent results for a TLN of 10. As NPT is increased there is a slight increase in skill, although almost all of the differences fail to achieve significance. Only differences in skill for NPT 1 versus 20 are possibly significant. Varying TLN generally has little effect; the open squares show cases for which TLN is 20. Only for a radical reduction in TLN (TLN for open circle is 1) is there a greater impact, although the differences are not even close to being significant. Overall, based on these results a reasonable choice for implementing LIM is with both NPT and TLN having values of 10.

3.4 | Graphical example of tail schemes

To better understand the tail adjustment schemes a graphical representation is given for a select case. The data for this case are provided in Supporting Information. Figure 4 displays a quantile–quantile (q–q) plot based on September values for a gridpoint near Douglas, WY. While the abscissa is the GCM value, the ordinate represents either the observed value (black) or that from one of several tail schemes (CIT–red, SIM–cyan, or LIM–blue). Tail scheme values closer to the observations represent better results. The most outstanding feature is the poor performance of CIT in the right tail as six points lie far above the observed set of points. Both SIM and LIM yield similar, much better results as the “constant correction” paradigm that they employ results in a linear extension of values

out from the points in the interior of the distribution. Results for the left tail are less dramatic as there is only a minor change in values from CIT, although for both SIM and LIM there is a modest improvement for the minimum value.

3.5 | Example involving extreme value analysis

In this section we provide an example of the type that a practitioner might encounter, which demonstrates the ramifications of our findings. The data are similar to those used in the previous section for a gridpoint near Douglas, WY except that we concatenate the data from all three ensemble members yielding a sample of 30 years. We perform an extreme value analysis, an approach which since its introduction to the study of climate change (Kharin and Zwiers, 2000) is now commonly employed.

For our analyses we adopt the block maxima approach in which we first create a subset of data composed of the largest value in each of the 30 available Septembers. We fit the generalized extreme value (GEV) distribution to these 30 values using the “extRemes” package in the CRAN repository (<https://cran.r-project.org/web/packages/extRemes/index.html>). Such analyses are performed separately on data from the observations (OBS), GCM, CDFt, SIM, and LIM.

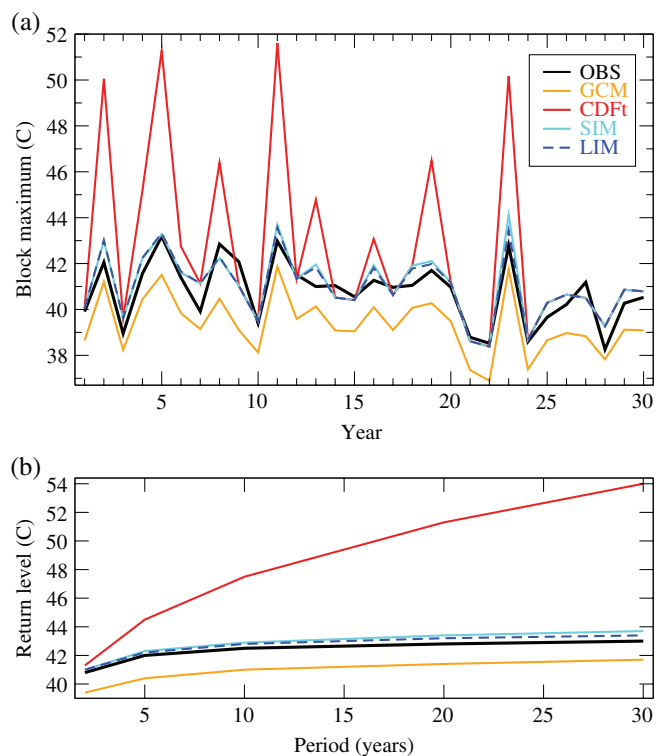


FIGURE 5 Block maxima in C (a) and return levels in C for 2, 5, 10, 20, and 30 years (b) based on extreme value analyses using data for a gridpoint near Douglas, WY (105.2 W, 42.9 N). Data from the three available 10-year ensemble members were concatenated yielding 30 Septembers upon which to base the analyses. Separate analyses were performed using observations (OBS, black), GCM (orange), CDFt (red), SIM (cyan), and LIM (dashed blue). For SIM and LIM NPT = 10 and for LIM TLN = 10 [Colour figure can be viewed at wileyonlinelibrary.com]

The block maxima for each case are plotted in Figure 5a. As expected, the extremes are systematically underestimated by a few percent for the GCM since its data represent a spatial average of the OBS. On the other hand CDFt often overestimates the extremes by as much as about 25%. However, both of our tail schemes (SIM and LIM) produce extremes that are typically close to the OBS with only a very slight tendency for overestimation. The return levels shown in Figure 5b reflect the tendencies reflected by the block maxima. The GCM consistently underestimates return levels for the OBS by a few percent whereas CDFt shows a bias that increases with the period, from negligible at the lowest period (2 years) to about 25% at the highest (30 years). Our tail schemes perform the best having a typical bias of 1% or less.

4 | CONCLUSIONS

We have examined the performance of CDFt in downscaling values from the tails of the distribution. CDFt performs considerably worse in the tails than for the rest of the distribution. The default scheme for handling OOB conditions (CIT) does particularly poorly across distributional categories. Although much better than CIT, the scheme for non-OOB conditions (CBA) is still noticeably worse in the tails than two alternatives. The “constant correction” scheme (SIM), as used extensively in the literature, is a significant improvement over CBA. But we can improve upon it through our modification which employs more values (~10 vs. 1) to compute the correction factor. The best results come from our LIM method, which operates much like SIM, except that it ignores the OOB conditions and is applied to a user selected number of points (~10) in the tail. Given the shortcomings identified in CDFt, and the improvements by way of our tail adjustment schemes, in a follow-up study we intend to apply our methodology to a number of other popular distributional-type downscaling methods.

An attempt was made to diagnose the tail behaviour of CDFt. For a limited number of cases, at a few select grid-points, perturbation analyses were performed in which some of the input data were systematically, incrementally altered over a range of values. Unfortunately, this exercise did not uncover consistent responses and we are not able to shed much light on the possible causes of the aberrant behaviour of CDFt for the tails. However, we note that in some European-based community-wide evaluations (Maraun *et al.*, 2017; Gutierrez *et al.*, 2018) CDFt's poor performance for some selected metrics does stand out as an outlier compared to a large number of other downscaling methods, although it is unclear to what extent that might be related to our findings.

ACKNOWLEDGEMENTS

This study was supported in part by the USGS South Central Climate Adaptation Science Center (G13AC00387). We

thank Tom Knutson and Adrienne Wootten for comments on an earlier draft of this manuscript.

ORCID

John R. Lanzante  <https://orcid.org/0000-0002-1736-7170>

REFERENCES

- Bierbaum, R., Lee, A., Smith, J., Blair, M., Carter, L.M., Chapin, F.S., Fleming, P., Ruffo, S., McNeely, S., Stults, M., Verduzco, L., Seyller, E., Melillo, J.M., Richmond, T. and Yohe, G.W. (2014) Adaptation. In: *Climate Change Impacts in the United States: The Third National Climate Assessment*. Washington, DC: U.S. Global Change Research Program. <https://doi.org/10.7930/J07H1GGT>.
- Bretherton, C., Widmann, M., Dymnikov, V., Wallace, J. and Blade, I. (1999) The effective number of spatial degrees of freedom of a time-varying field. *Journal of Climate*, 12(7), 1990–2009. [https://doi.org/10.1175/1520-0442\(1999\)012<1990:TENOSD>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<1990:TENOSD>2.0.CO;2).
- 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>.
- Colette, A., Vautard, R. and Vrac, M. (2012) Regional climate downscaling with prior statistical correction of the global climate forcing. *Geophysical Research Letters*, 39, L13707. <https://doi.org/10.1029/2012GL052258>.
- 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>.
- Famien, A.M., Janicot, S., Ochoa, A.D., Vrac, M., Defrance, D., Sultan, B. and Noel, T. (2017) A bias-corrected CMIP5 dataset for Africa using CDF-t method. A contribution to agricultural impact studies. *Earth System Dynamics Discussions*, 9(1), 313–338. <https://doi.org/10.5194/esd-2017-111>.
- Flaounas, E., Drobinski, P., Vrac, M., Bastin, S., Lebeau-pin-Brossier, C., Stefanon, M., Borga, M. and Calvet, J.-C. (2013) Precipitation and temperature space-time variability and extremes in the Mediterranean region: evaluation of dynamical and statistical downscaling methods. *Climate Dynamics*, 40(11–12), 2687–2705. <https://doi.org/10.1002/2015JD023977>.
- 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>.
- Fraedrich, K., Ziehmann, C. and Sielmann, F. (1995) Estimates of spatial degrees of freedom. *Journal of Climate*, 8(2), 361–369. [https://doi.org/10.1175/1520-0442\(1995\)008%3C0361:EOSDOF%3E2.0.CO;2](https://doi.org/10.1175/1520-0442(1995)008%3C0361:EOSDOF%3E2.0.CO;2).
- 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., Manzanar, R., Iturbide, M., Vrac, M., Dubrovsky, M., Ribalaygua, J., Portoles, J., Raty, O., Raisanen, J., Hingray, B., Raynaud, D., Casado, M.J., Ramos, P., Zerener, T., Turco, M., Bosshard, T., Stepanek, P., Bartholy, J., Pongracz, R., Keller, D.E., Fischer, A.M., Cardoso, R.M., Soares, P.M.M., Czernecki, 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*. <https://doi.org/10.1002/joc.5462>.
- Gutmann, E., Pruitt, T., Clark, M.P., Brekke, L., Arnold, J.R., Raff, D.A. and Rasmussen, R.M. (2014) An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resources Research*, 50(9), 7167–7186. <https://doi.org/10.1002/2014WR015559>.
- Hall, P. (1985) Resampling a coverage pattern. *Stochastic Processes and their Applications*, 20(2), 231–246. [https://doi.org/10.1016/0304-4149\(85\)90212-1](https://doi.org/10.1016/0304-4149(85)90212-1).
- Harnack, R. and Lanzante, J. (1984) Specification of seasonal mean 700 mb heights over North America by North Pacific and North Atlantic sea surface

- temperature. *Monthly Weather Review*, 112(8), 1626–1633. [https://doi.org/10.1175/1520-0493\(1984\)112%3C1626:SOSMMH%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(1984)112%3C1626:SOSMMH%3E2.0.CO;2).
- Harnack, R., Harnack, J. and Lanzante, J. (1986) Seasonal temperature predictions using a jackknife approach with an intraseasonal variability index. *Monthly Weather Review*, 114(10), 1950–1954. [https://doi.org/10.1175/1520-0493\(1986\)114%3C1950:STPUAJ%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(1986)114%3C1950:STPUAJ%3E2.0.CO;2).
- 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\(1996\)09%3C1350:AOMCSM%3E2.0.CO;2](https://doi.org/10.1175/1520-0442(1996)09%3C1350:AOMCSM%3E2.0.CO;2).
- Kallache, M., Vrac, M., Naveau, P. and Michelangeli, P.A. (2011) Nonstationary probabilistic downscaling of extreme precipitation. *Journal of Geophysical Research*, 116(D5), D05113. <https://doi.org/10.1029/2010JD014892>.
- Kharin, V.V. and Zwiers, F.W. (2000) Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere–ocean GCM. *Journal of Climate*, 13(21), 3760–3788. [https://doi.org/10.1175/1520-0442\(2000\)013<3760:CITEIA>2.0.CO;2](https://doi.org/10.1175/1520-0442(2000)013<3760:CITEIA>2.0.CO;2).
- 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%3C1197::AID-JOC89%3E3.0.CO;2-L](https://doi.org/10.1002/(SICI)1097-0088(199611)16:11%3C1197::AID-JOC89%3E3.0.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>.
- Lavaysse, C., Vrac, M., Drobinski, M., Lengaigne, M. and Vischel, T. (2012) Statistical downscaling of the French Mediterranean climate: assessment for present and projection in an anthropogenic scenario. *Natural Hazards and Earth System Science*, 12, 651–670. <https://doi.org/10.5194/nhess-12-651-2012>.
- 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%3C0046:SFSaid%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(1983)111%3C0046:SFSaid%3E2.0.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. and 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. <https://doi.org/10.1029/2009RG000314>.
- Maraun, D., Huth, R., Gutiérrez, J., San Martín, D., Dubrovsky, M., Fischer, A., Hertig, E., Soares, P.M.M., Bartholy, J., Pongrácz, R., Widmann, M., Casado, M.J., Ramos, P. and Bedia, J. (2017) The VALUE perfect predictor experiment: evaluation of temporal variability. *International Journal of Climatology*. <https://doi.org/10.1002/joc.5222>.
- 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>.
- Nature Editorial. (2018) Governments must take heed of latest IPCC assessment. *Nature*, 562(7726), 163. <https://doi.org/10.1038/d41586-018-06952-7>.
- NCAR. (2017) *NCAR Command Language (Version 6.4.0) [Software]*. Boulder, CO: UCAR/NCAR/CISL/TDD. <https://doi.org/10.5065/D6WD3XH5>.
- Oettli, P., Sultan, B., Baron, C. and Vrac, M. (2011) Are regional climate models relevant for crop yield prediction in West Africa? *Environmental Research Letters*, 6, 014008. <https://doi.org/10.1088/1748-9326/6/1/014008>.
- Perlwitz, J. and Graf, H. (2001) The variability of the horizontal circulation in the troposphere and stratosphere—a comparison. *Theoretical and Applied Climatology*, 69(3–4), 149–161. <https://doi.org/10.1007/s007040170021>.
- 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>.
- 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](https://doi.org/10.1175/1520-0469(1983)040%3C1884:NMRITU%3E2.0.CO;2).
- Thiebaut, H. and Zwiers, F. (1984) The interpretation and estimation of effective sample size. *Journal of Climate and Applied Meteorology*, 23(5), 800–811. [https://doi.org/10.1175/1520-0450\(1984\)023%3C0800:TIAEOE%3E2.0.CO;2](https://doi.org/10.1175/1520-0450(1984)023%3C0800:TIAEOE%3E2.0.CO;2).
- Tisseuil, C., Vrac, M., Grenouillet, G., Gevrey, M., Oberdorff, T., Wade, A. and Lek, S. (2012) Strengthening the link between hydro-climatic downscaling and species distribution modelling: climate change impacts on freshwater biodiversity. *Science of the Total Environment*, 424, 193–201. <https://doi.org/10.1016/j.scitotenv.2012.02.035>.
- Vautard, R., Noël, T., Li, L., Vrac, M., Martin, E., Dandin, P. and Jousseaume, S. (2013) Climate variability and trends in downscaled high-resolution simulations and projections over metropolitan France. *Climate Dynamics*, 41(5–6), 1419–1437. <https://doi.org/10.1007/s00382-012-1621-8>.
- Vigaud, N., Vrac, M. and Caballero, Y. (2013) Probabilistic downscaling of GCM scenarios over southern India. *International Journal of Climatology*, 33(5), 1248–1263. <https://doi.org/10.1002/joc.3509.1>.
- Vrac, M., Drobinski, P., Merlo, A., Herrmann, M., Lavaysse, C., Li, L. and Somot, S. (2012) Dynamical and statistical downscaling of the French Mediterranean climate: uncertainty assessment. *Natural Hazards and Earth System Sciences*, 12(9), 2769–2784. <https://doi.org/10.5194/nhess-12-2769-2012>.
- Vrac, M., Noel, T. and Vautard, R. (2016) Bias correction of precipitation through singularity stochastic removal: because occurrences matter. *Journal of Geophysical Research*, 121(10), 5237–5258. <https://doi.org/10.1002/2015JD024511>.
- Wang, X. and Shen, S. (1999) Estimation of spatial degrees of freedom of a climate field. *Journal of Climate*, 12(5), 1280–1291. [https://doi.org/10.1175/1520-0442\(1999\)012%3C1280:EOSDOF%3E2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012%3C1280:EOSDOF%3E2.0.CO;2).
- Wilks, D.S. (2006) *Statistical Methods in the Atmospheric Sciences*, 2nd edition. San Diego, CA: Academic Press.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Lanzante JR, Nath MJ, Whitlock CE, Dixon KW, Adams-Smith D. Evaluation and improvement of tail behaviour in the cumulative distribution function transform downscaling method. *Int J Climatol*. 2019;39:2449–2460. <https://doi.org/10.1002/joc.5964>

APPENDIX A: PERFECT MODEL DATA FILES

The data used as input (O_h , M_h , and M_f) as well as that used in validation (O_f) are the Perfect Model data created by Dixon *et al.* (2016). The files (a single ensemble member for historical and three members for the future) are self-documented in the NetCDF format and can be accessed as follows:

O_h

ftp://ftp.gfdl.noaa.gov/perm/Oar.Gfdl.Esd/PMdata/GFDL-HIRAM-C360/amp/r1i1p1/tasmax/tasmax_day_GFDL-HIRAM-C360_ampi_r1i1p1_US48_19790101-20081231.nc

M_h

ftp://ftp.gfdl.noaa.gov/perm/Oar.Gfdl.Esd/PMdata/GFDL-HIRAM-C360-COARSENEDED/ampi_r1i1p1/tasmax/tasmax_day_GFDL-HIRAM-C360-COARSENEDED_ampi_r1i1p1_US48_19790101-20081231.nc

O_f

ftp://ftp.gfdl.noaa.gov/perm/Oar.Gfdl.Esd/PMdata/GFDL-HIRAM-C360/sst2090/r1i1p2/tasmax/tasmax_day_GFDL-HIRAM-C360_sst2090_r1i1p2_US48_20860101-20951231.nc

ftp://ftp.gfdl.noaa.gov/perm/Oar.Gfdl.Esd/PMdata/GFDL-HIRAM-C360/sst2090/r2i1p2/tasmax/tasmax_day_GFDL-HIRAM-C360_sst2090_r2i1p2_US48_20860101-20951231.nc

ftp://ftp.gfdl.noaa.gov/perm/Oar.Gfdl.Esd/PMdata/GFDL-HIRAM-C360/sst2090/r3i1p2/tasmax/tasmax_day_GFDL-HIRAM-C360_sst2090_r3i1p2_US48_20860101-20951231.nc

M_f

ftp://ftp.gfdl.noaa.gov/perm/Oar.Gfdl.Esd/PMdata/GFDL-HIRAM-C360-COARSENED/sst2090/r1i1p2/tasmax/tasmax_day_GFDL-HIRAM-C360-COARSENED_sst2090_r1i1p2_US48_20860101-20951231.nc

ftp://ftp.gfdl.noaa.gov/perm/Oar.Gfdl.Esd/PMdata/GFDL-HIRAM-C360-COARSENED/sst2090/r2i1p2/tasmax/tasmax_day_GFDL-HIRAM-C360-COARSENED_sst2090_r2i1p2_US48_20860101-20951231.nc

ftp://ftp.gfdl.noaa.gov/perm/Oar.Gfdl.Esd/PMdata/GFDL-HIRAM-C360-COARSENED/sst2090/r3i1p2/tasmax/tasmax_day_GFDL-HIRAM-C360-COARSENED_sst2090_r3i1p2_US48_20860101-20951231.nc

APPENDIX B: SIGNIFICANCE ASSESSMENT PROCEDURE

B1 | OVERVIEW OF THE SIGNIFICANCE ASSESSMENT PROCEDURE

Our procedure consists of two steps, the first of which is to estimate an effective block size embedded within our grid which we use in the second step to estimate significance. For each assessment we have a pair of skill maps corresponding to two different treatments. We wish to estimate the significance of the difference in mean skill over the two treatments (i.e., maps).

Following Livezey and Chen (1983), hereafter referred to as LC83, the collection of skill values on a given map are not independent metrics. As a simplification we can seek an effective number of gridpoints in the same vein that Thiébaux and Zwiers (1984) sought an effective (temporal) sample size. Our blocking procedure is also inspired by the block bootstrap introduced by Kunsch (1989). That approach resamples blocks of consecutive values in time series in order to retain the essential temporal coherence. We resample blocks of adjacent gridpoints, as inspired by Hall (1985), in order to retain the essential spatial coherence. In LC83 they equated the distribution of results from Monte Carlo trials to the binomial distribution in order to infer an effective

sample size. Analogously we infer an effective sample size by equating the distribution of our results with that from the theoretical distribution of the Spearman correlation.

We convert an effective spatial sample size to an effective set of blocks (latitude \times longitude spacing) by rounding to the nearest integer longitude and latitude increments that best retain the aspect ratio of the original full grid (194 \times longitudes \times 114 latitudes). The actual assignment of gridpoints to blocks is done in such a fashion so as to have nearly the same number of gridpoints in each block.

Finally, given the approximate nature of our procedure we subjectively estimated a minimum possible spatial sample size through a survey of the literature. We surveyed Harnack and Lanzante (1984), Harnack *et al.* (1986), Fraedrich *et al.* (1995), Huang *et al.* (1996), Bretherton *et al.* (1999), Wang and Shen (1999), and Perlwitz and Graf (2001). Because none of their grids were the same as ours we used scaling arguments via ratios of grid sizes. In addition, none of these works dealt with extremes and some used monthly or seasonal rather than daily data. Spatial scales of time-averaged data will be greater than daily data. Furthermore, spatial scales associated with our limited sample of extreme values taken from the tails will be smaller than the scales associated with data from the full distribution as extremes may be more influenced by unique, rarer or more isolated local factors. In using our judgement when given choices, we erred on the side of smaller estimates. We were very conservative in our estimate (4 \times longitudes \times 2 latitudes) and consider this to be almost certainly too few blocks.

B2 | DETERMINATION OF EFFECTIVE BLOCK SIZE

The starting point is a pair of skill maps corresponding to two treatments. Since maps may have missing values we first fill these (but not ocean points that have been masked out) via bivariate interpolation (IDSFFT) obtained as Fortran code in the NCAR graphics package contained in the NCAR Command Language (NCAR, 2017).

Each trial utilizes three uniform random numbers varying between 0 and 1. Two are used to generate random translations/shifts (i.e., a fraction of the total grid dimension) in the longitudinal and latitudinal dimensions. As needed we wrap values around the edges and only shift between interior land points (i.e., ignore ocean points). The other random number is used to assign a random algebraic sign. In summary, in this step we translate the original maps semi-rigidly, preserving the essential patterns and attendant spatial scales, yet randomizing the phase (i.e., position).

The next step is to compute the Spearman pattern correlation (r) between the original pair of maps and the shifted pair. These correlations are based on the same set of valid (i.e., non-missing) gridpoints as in the original pair of maps. We perform 1,000 trials and accumulate the correlations as well as their Fisher- z transforms (F):

$$F = 0.5 \times \ln[(1+r)/(1-r)]. \quad (\text{B1})$$

We compute (a) the variance of the correlations and (b) the average of the values that place 2.5% of the distribution in the lower and upper tails. For (a) we invert the formula for the standard error of the Spearman correlation (σ):

$$\sigma = 0.6325/\sqrt{(n-1)}, \quad (\text{B2})$$

and for (b) we invert the formula for the z -score used in assessing the Fisher's z of the Spearman correlation:

$$z = F \times \sqrt{[(n-3)/1.06]} \quad (\text{B3})$$

(https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient). The inversion processes yields two sample sizes (n) for which method (a) tends to yield smaller estimates than that of (b).

B3 | PERMUTATION TEST FOR SIGNIFICANCE

For each of the three block sizes (two quantitative and one subjective) we can derive estimates of significance

employing a Monte Carlo procedure based on a pool permutation procedure (Preisendorfer and Barnett, 1983). For each trial we construct pseudo-maps corresponding to each of the two treatments. These are constructed by randomly assigning each block of gridpoint values to one of the maps. This assumes that there is no difference between the two maps (the null hypothesis) so that the values are interchangeable. By reassigning blocks rather than individual gridpoints we attempt to retain the essential coherence between gridpoints. Since not all blocks have the same number of gridpoints it may be necessary to split a block in order for the number of gridpoints on each pseudo-map to equal those from the original maps.

After all blocks have been reassigned the mean skill over each is computed and the absolute value of the difference in skill is saved. After 1,000 trials the absolute value of the difference in mean skill between the two original treatment maps is referenced in the distribution from the trials to determine a significance level.