SOME PITFALLS IN STATISTICAL DOWNSCALING OF FUTURE CLIMATE

John R. Lanzante, Keith W. Dixon, Mary Jo Nath,
Carolyn E. Whitlock, and Dennis Adams-Smith

Statistical downscaling is generally successful and used widely to refine projections of future climate, though in some circumstances it can lead to highly erroneous results.

There is a growing need for localized climate information as policy makers and stakeholders increasingly recognize the importance of assessing and planning for the impacts of climate change. A primary resource in these endeavors is the output from large-scale global climate models (GCMs). These outputs can be diagnosed directly or used as inputs into impact assessment models. However, the GCM outputs in their raw form are generally not suitable for such purposes.

AFFILIATIONS: Lanzante, Dixon, and Nath—NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey; Whitlock— NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, and Engility Inc., Dover, New Jersey; Adams-Smith—NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, and CPAESS, University Corporation for Atmospheric Research, Boulder, Colorado
CORRESPONDING AUTHOR: John R. Lanzante, john.lanzante@noaa.gov

The abstract for this article can be found in this issue, following the table of contents.
DOI:10.1175/BAMS-D-17-0046.1

In final form 4 November 2017

Downscaling is typically used to transform the raw GCM outputs into a more suitable form. Two broad types of downscaling exist: dynamical downscaling (DD) and statistical downscaling (SD). The former is based on physical models, referred to as regional climate models (RCMs), that share much in common with GCMs, yet they have some distinct advantages for use in studying local impacts. The latter is empirical in nature. One of the chief advantages of SD over DD is that it is much cheaper to implement, requiring far less computational resources. It is not the purpose of this paper to debate the relative merits of SD versus DD; rather, we acknowledge the fact that SD is widely used in the climate change arena.

There are several objectives of SD, including

1) enhanced spatial detail,
2) mitigation of systematic GCM biases, and
3) generation of variables not explicitly rendered by GCMs.

Global climate models used to simulate multidecadal to century time-scale phenomena typically resolve spatial details at scales of ~100–200 km, whereas impact studies may require more localized climate information resolved at scales of ~1–10 km or less. GCM[1] simulations of contemporary climate are often afflicted with systematic biases at the regional level, such as warm or cold, wet or dry, that may vary by geographic location and time of year, as well as by the specific model employed. Furthermore, GCMs may not incorporate or resolve some phenomena of interest to the impacts community, for example, flash floods, wildfires, or tornadoes. Statistical downscaling has the potential to address these three goals.

Although SD methods are quite varied in both complexity and philosophy, they share some fundamental characteristics. There are two basic steps to SD: a training step and an application step. The training step utilizes two datasets: one representative of real-world observations and another from the output of a physical model (e.g., GCM, RCM, or reanalysis). Both sets are contemporary to some past period. Typically, the observations will be of higher spatial resolution than the GCM data. During the training step, some sort of transfer relationship between these two datasets is derived, forming a bridge between observations and GCM.

In the application step, the previously derived transfer function is applied to a set of GCM data corresponding to a different period. The result of the application step is the generation of a dataset of proxy observations, or downscaled results. In the case of climate change,[2] the GCM outputs correspond to some future state for which various forcing agents (e.g., greenhouse gases) have changed. The philosophy behind SD is to recalibrate the raw GCM output for some other (future) climate state, imparting characteristics of the observations during the historical (training) period, and in the process yield information at a higher spatial resolution than available from the GCM. Climate projections that have been refined by SD, informed by observations, are generally regarded to be more suitable for use in many climate impact applications than are the raw outputs of the dynamical models.

Although SD methods vary considerably, in general judicious application of SD has been shown broadly to add value to raw GCM outputs (Fowler et al. 2007; Maraun et al. 2010). For example, Dixon et al. (2016) found that one SD method reduced the raw GCM error in temperature by about 50% during the historical period and 30%–40% during the future. However, lurking beneath the surface of all SD methods is an implicit assumption that the transfer relationships derived from historical data are valid in a future period during which fundamental aspects of the climate system may have changed. In more technical terms, this is referred to as the assumption of *statistical stationarity* (Dixon et al. 2016).

Unfortunately, the only way to strictly test this crucial assumption would be to utilize observations from the future. In lieu of this, researchers often resort to cross validation, which involves withholding part of the historical dataset during training and later using it for independent validation (Efron 1982). Although this exercise is quite useful, nevertheless it is no substitute for actual validation using future observations, because the climate states of the training and application periods in cross validation will not differ nearly as much, or in the same fashion as between the historical past and the more distant future period of interest in the real world.

In an attempt to apply a more rigorous test of the stationarity assumption, some researchers have employed variants of a "perfect model" (PM)[3] experimental design (e.g., Frías et al. 2006; Vrac et al. 2007; Maraun 2012; Dayon et al. 2015; Maraun et al. 2015; Velazquez et al. 2015; Ivanov and Kotlarski 2017). The basic notion is to use climate model output as surrogates for both observations and GCM data. Although many such types of designs are possible, the authors of this paper, members of the Empirical Statistical Downscaling (ESD)[4] team at the Geophysical Fluid Dynamics Laboratory (GFDL), currently employ one such design. Here we will describe only the most essential aspects. The interested

[1] For simplicity, throughout this manuscript we make reference to GCM output as being the target to which SD is applied. It should be understood that it is often appropriate to apply SD to output from the general class of dynamical models to which GCMs belong, along with RCMs and reanalysis systems.

[2] For the class of climate change impact studies considered here, a common approach is to apply SD to refine multidecadal climate model projections, such as those that comprise phase 5 of the Coupled Model Intercomparison Project (CMIP5) archive (Taylor et al. 2012).

[3] While the term *perfect model* is used widely among specialists, to the uninitiated it may be misleading. There is no intent to imply that models are infallible. Rather, it refers to an experimental design in which some of the real-world complexities are eliminated, allowing for a more clarified exploration of causal mechanisms.

[4] More information about the GFDL ESD team can be found at www.gfdl.noaa.gov/esd_eval/.

reader is referred to Dixon et al. (2016) for more details. In this paper all of our results are for daily maximum temperature (Tmax), although we believe that the overarching findings are likely applicable to other variables as well. Results for the future are based on data from a high-climate-sensitivity and high-emissions scenario (scenario C in Dixon et al. 2016) for the period 2086–95; results for the historical period are for 1979–2008.

The philosophy behind the PM design is that a GCM replicates the essential characteristics of the real climate system. In our PM we utilize output from a GCM that has higher spatial resolution (~25 km) than typically found in the current generation of workhorse global models. We treat the raw output from this GCM as "observations" (OBS). We then degrade the raw output, through the use of spatial averaging, to produce lower-resolution data (~200 km) that we then treat as "GCM data" (GCM). Hence, we build into our scheme the inherent mismatch in resolution between the two datasets. Furthermore, because we have simulations from not only the past but also a future climate state that is fundamentally different, we are able to more rigorously test the stationarity assumption.

Within the field of SD there is a wide variety of methodological approaches as well as a wide range in the complexity of schemes (e.g., Benestad et al. 2008; Deque 2007; Fowler et al. 2007; Maraun et al. 2010; Pierce et al. 2014, 2015; to name just a few). While there is no unique way of categorizing the different approaches, one convenient way is through

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

Methods from category 1 consist of both linear and nonlinear techniques, both simple and complex. Examples at opposite ends of the spectrum of sophistication include simple linear regression and complex neural nets. Category 2 involves mapping characteristics of statistical distribution functions between observations and models. Category 3 involves spatial mapping methods; some of the most popular are analog methods that seek patterns from the past that match patterns projected to occur in the future. Category 4 is stochastic in nature and produces ensembles of realizations. Some methods may be multistep and employ approaches from more than one class or technique.

Current work by the GFDL ESD team and collaborators is engaged in evaluating a variety of SD methods using an assortment of metrics, the results of which will be reported elsewhere. In this paper we select one representative method to illustrate some of the more interesting findings to date that demonstrate dramatic violation of the stationarity assumption. Much of what we illustrate is independent of the particular SD method. In this paper we have chosen to utilize a method known as quantile mapping (Panofsky and Brier 1968), which is widely used in SD. Inspired by Ho et al. (2012) we refer to it as bias correction quantile mapping (BCQM). It is one of the simplest in a class of methods that involve mapping the characteristics of the cumulative distribution functions between observations and models as a means of establishing transfer relationships. For illustrative purposes we downscale Tmax; with this choice the maladies we find are readily apparent and easily explained. The results for BCQM are consistent with several other popular methods in its class (distributional methods), which are incorporated widely in popular SD methods. For more on this class of methods, the interested reader is referred to Pierce et al. (2015).

Our findings are an outgrowth of cautions raised by Ho et al. (2012). They distinguish between two distinct paradigms used in calibration of climate model output: bias correction versus change factor (or delta). Understanding these concepts proves useful in the interpretation of our results. Bias correction adjusts the model output to fall in line with that of the observations. In its simplest form, bias correction involves subtracting the mean difference (model minus observations) in the historical period from all of the model values in both the historical and future periods. It tacitly assumes that the model discrepancies are time invariant—an assumption that becomes more critical when considering future projections whose climates differ markedly from the historical period used in SD training. Conversely, the change factor strategy assumes that the change from historical to future in the observations is the same as that from the model. Common distributional methods adopt one or both of these strategies, which are manifested as a correction factor that varies by relative position within the statistical distribution, rather than a constant for all values, as in the simplest case. We demonstrate that the concerns raised by Ho et al. (2012) as to the sometimes very different results produced by these two competing strategies were well founded.

In this paper we begin by briefly presenting an overview of the climatological seasonal variation of downscaling errors, and later the corresponding geographical aspects, in order to provide the context for more in-depth analyses. We select locations at which the results are representative of two distinct maladies and link aspects of the maladies' causes to 1) physical climate characteristics, 2) statistical characteristics of the climate variable distributions, and 3) downscaling methodological choices. Later in our discussion we present implications for our findings.

RESULTS. *Seasonal climatology of errors.* Almost all applications of the distributional methods to SD employ one or both of the two approaches based on the paradigms described by Ho et al. (2012).[5] We use BCQM[6] employing *raw* data as inputs (BCQMr) as representative of the bias correction approach. Alternately, we employ BCQM using anomaly inputs, by subtracting the time-smoothed daily climatological means of the GCM from the raw data. After transformation via quantile mapping, we add the means back to the result. This *anomaly* approach (BCQMa) incorporates both the bias correction and change factor paradigms, which is characteristic of a number of distributional methods currently used in practice.

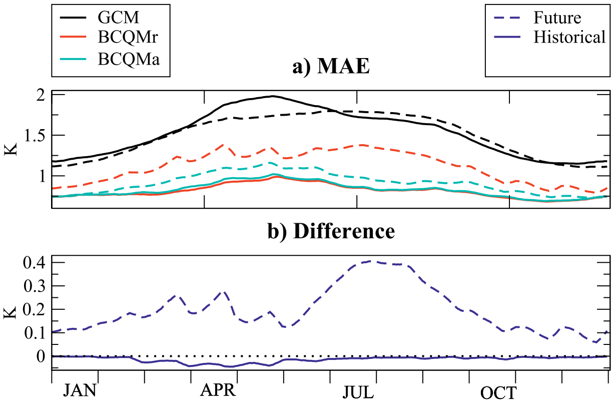


Fig. 1. (a) CONUS average Tmax (K) MAE of BCQM and GCM as a function of day of the year, smoothed via a 15-day running average. Results are given for BCQMa (cyan), BCQMr (red), and GCM (black), for the historical period (1979–2008; solid) and the future period (2088–95; dashed). (b) Difference for BCQM (BCQMr minus BCQMa) for the historical period (solid) and the future period (dashed). All results are based on averages over two 30-yr ensembles for the historical period and three 10-yr ensembles for the future period.

The seasonal variation of the daily mean absolute error (MAE) of downscaling Tmax via BCQM, averaged over the continental United States (CONUS) is shown in Fig. 1a. During the historical period (solid), the two approaches (red for raw and cyan for anomaly) yield nearly the same results. However, during the future period (dashed) the anomaly approach (cyan) is superior. For reference, the MAE based on the raw GCM (i.e., no downscaling) is shown in black (solid for historical, dashed for future). It is noteworthy that in all cases, averaged over the CONUS, downscaling yields a considerable improvement over use of the raw GCM values.

The curves in Fig. 1b show the difference in area-averaged MAE for BCQM between the raw and anomaly approaches. For the historical period (solid), the differences are mostly negligible except slight during the spring. For the future period (dashed), the differences are much greater with a primary maximum in the summer and a secondary maximum in the spring. We will explore distinct maladies that are associated with each of these maxima.

The "coastal effect." An exacerbation of errors along coastal compared to inland regions in downscaling future temperatures was identified by Dixon et al. (2016) and associated with GCM temperatures in the future exceeding the historical maximum. Further investigation revealed a large contribution of this error in the form of bias, the geographic distributions of which during July are displayed in Fig. 2. The biases in the GCM data, which are the inputs into BCQM, are shown in Fig. 2a for the historical period, while those for the future are shown in Fig. 2b. The patterns are similar during both periods with generally negative biases in coastal locations. These (GCM – OBS) biases represent the challenge to any downscaling method. The historical biases remaining after the application of BCQMr are depicted in Fig. 2c and those for the future in Fig. 2d. During the historical period the BCQMr biases are largely removed, whereas during the future not only are the coastal biases present, but

[5] In real-world applications, quantile mapping employing *only* the change factor paradigm is rarely used. The reason is likely that the resulting data sample follows the temporal sequence of the historical observations rather than that of the future model values. Analysts typically find this undesirable. However, a number of methods employing both paradigms produce data following the temporal sequence of future model values. Nevertheless, the delta method, a nonquantile approach employing only the change factor paradigm, has been commonly used, likely because of its utmost simplicity (Fowler et al. 2007).

[6] For an explanation of the mechanics of BCQM, the interested reader is referred to Ho et al. (2012, especially their Figs. 1a,b) and to Pierce et al. (2015, especially their Fig. 1a) and the accompanying text.

their sign has reversed! Curiously, while the application of BCQMa yields very similar results to BCQMr during the historical period (cf. Figs. 2c,e), during the future BCQMa (Fig. 2f) it substantially reduces the coastal biases that result from BCQMr (Fig. 2d).

To further investigate potential root causes for this downscaling behavior, we examine histograms of Tmax (Fig. 3) from a grid point in south Florida (near Sixmile Bend). This point was chosen because its future input GCM data have a large negative bias

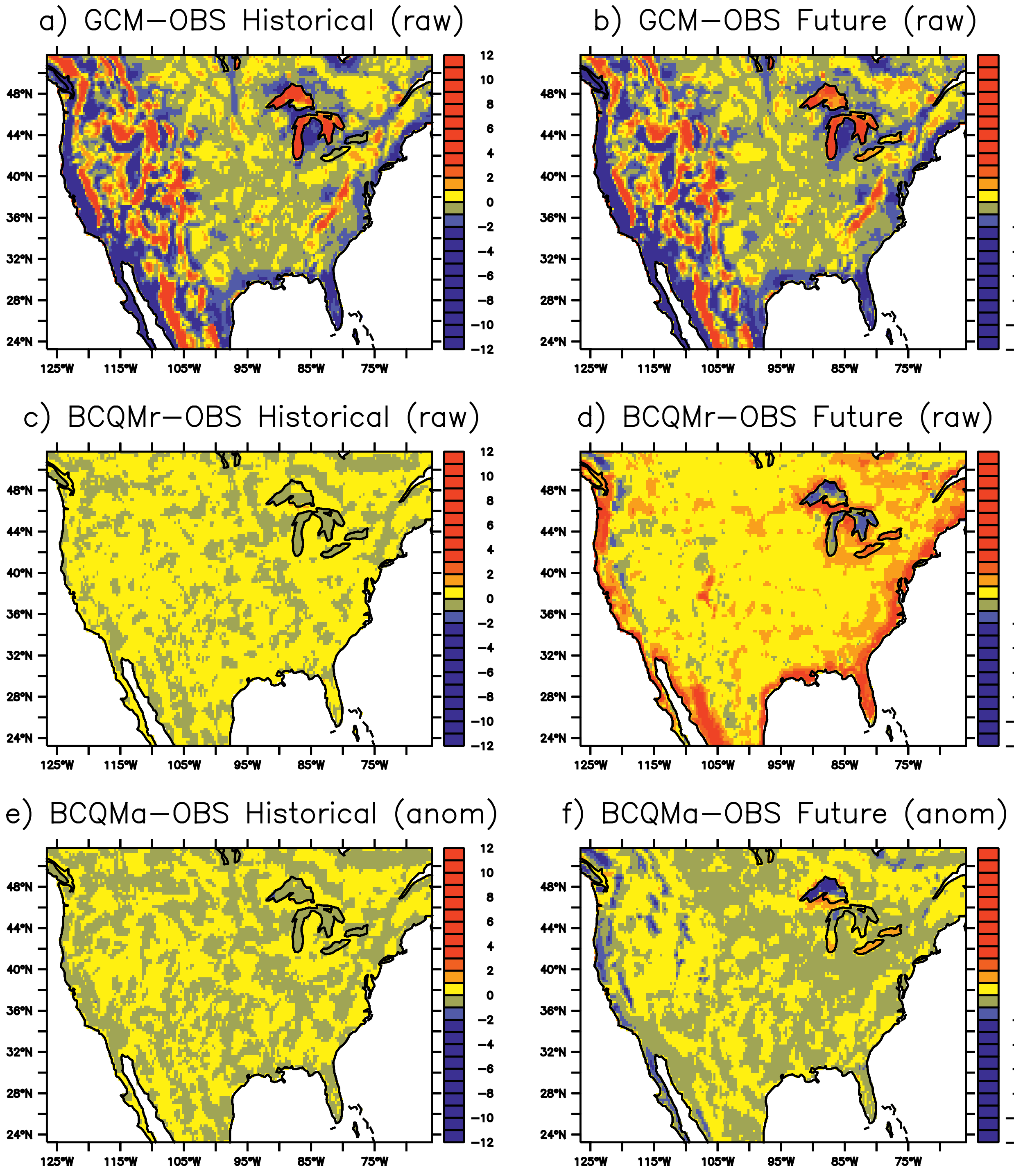

a) GCM–OBS Historical (raw)
b) GCM–OBS Future (raw)
c) BCQMr–OBS Historical (raw)
d) BCQMr–OBS Future (raw)
e) BCQMa–OBS Historical (anom)
f) BCQMa–OBS Future (anom)
48°N
44°N
40°N
36°N
32°N
28°N
24°N
125°W
115°W
105°W
95°W
85°W
75°W
12
10
8
6
4
2
0
−2
−4
−6
−8
−10
−12

Fig. 2. Biases (K) in Jul daily Tmax compared to OBS for the raw GCM inputs and the BCQM outputs for (left) historical and (right) future periods, as defined in Fig. 1. (a) Historical GCM minus OBS, (b) future GCM minus OBS, (c) historical BCQMr minus OBS, (d) future BCQMr minus OBS, (e) historical BCQMa minus OBS, and (f) future BCQMa minus OBS.

(Fig. 2b), while its future output from BCQMr has a large positive bias (Fig. 2d). Based on raw historical data (Fig. 3a), we see a substantial difference in the distributions of Tmax from historical observations (OBSh) and historical GCM (GCMh). While OBSh (black, shaded gray) has a positively skewed distribution, GCMh (red) is distributed more symmetrically and has a noticeably lower variance. Also, the mean of OBSh is several degrees higher than that of GCMh.

The distributional differences between OBSh and GCMh have simple physical explanations. An important factor is that the footprint of the GCM grid box is substantially larger than that of the collocated OBS grid box (see Fig. 2 of Dixon et al. 2016). While the OBS grid box lies entirely over land, the GCM grid box straddles land and ocean. Because of the larger heat capacity of the ocean, the maritime effect both suppresses daily Tmax and reduces variability, imparting negative biases to both the mean and variance of GCM compared to OBS.

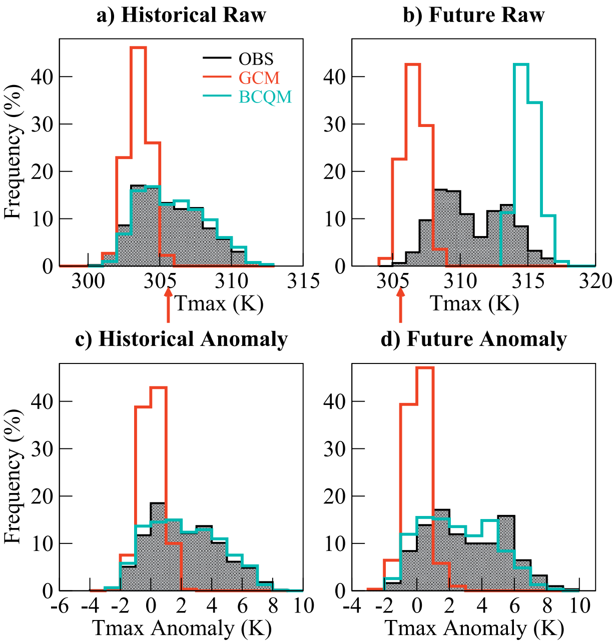


Fig. 3. Histograms of Jul Tmax (K) based on OBS (black, shaded gray), GCM (red), and BCQM downscaled (cyan) for a grid point near Sixmile Bend (26.6°N, 80.5°W) using 1° bins. (a) Historical raw, (b) future raw, (c) historical anomaly, and (d) future anomaly data. Red arrow in (a) and (b) denotes the position of the GCM maximum during the historical period. Note that irregularities in the smooth daily climatology used to generate anomalies could account for differences in the appearance of histograms between raw and anomaly versions of the data, along with random variations resulting from the use of discrete bins.

The stark differences between the distributions of Tmax for OBSh and GCMh seen in Fig. 3a present a considerable challenge to any downscaling method. However, BCQMr (cyan in Fig. 3a) produces a distribution that closely mimics that of OBSh. In the future (Fig. 1b) the characteristics of OBS (OBSf) and GCM (GCMf) are quite similar to that for the historical period except for a shift to the right of both distributions of ~5 K. The most striking difference is the extremely poor performance of BCQMr; instead of mimicking the shape of OBSf, it essentially retains the shape of GCMf and shifts it too far to the right. For BCQMa (Figs. 3c,d) during the historical period, its excellent performance is almost identical to that of BCQMr. However, during the future, in sharp contrast to BCQMr, BCQMa also performs quite well.

The explanation for the considerable difference in performance between BCQMr and BCQMa in the future lies in a small but important methodological detail. Although distributional downscaling methods differ in their algorithms, some of them have “out of bounds” conditions for which the basic algorithm cannot be applied. BCQM is unable to downscale any future GCM value that falls outside the range of the GCM values used for training the method in the historical period. When this occurs in our tests, a simple and commonly used tail scheme is applied instead (Deque 2007). Note the red arrow at the bottom of Figs. 3a and 3b which denotes the historical maximum of GCM values. As seen in Fig. 3b, most of the GCMf values exceed this value and thus are subject to the tail scheme.

To better understand how distributional methods operate, we examine quantile–quantile (Q–Q) plots as shown in Fig. 4. This figure does not display any downscaled results but instead explores the relationship between the OBS and GCM distributions, which may be used in training a distributional method. The abscissa (GCM) and ordinate (OBS) values are sorted separately and then plotted as (x, y) points, such that both x and y have the same rank in their respective sets. The line $y = x$ (green) is included as a reference. Points that fall on this line indicate zero bias so that downscaling has no “work” to perform. Conceptually, the difference pattern between the historical curve and the green line depicts what a downscaling method “sees” as being the adjustment needed in order to bring a GCM value into agreement with observations. The black curve represents the training relationship derived during the historical period that is used to downscale future values. The orange curve represents the actual future relationship (i.e., what ought to be used).

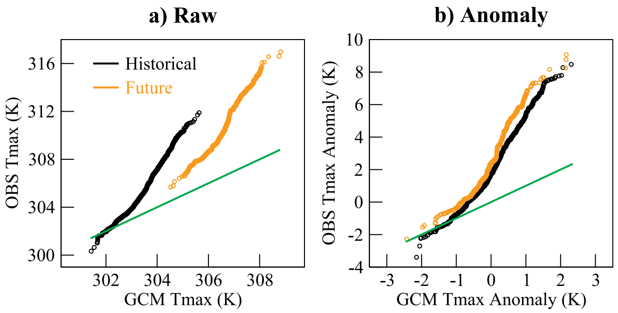


Fig. 4. Sixmile Bend Jul Tmax Q–Q plots for data in (a) raw and (b) anomaly form. Abscissa corresponds to GCM and the ordinate to OBS Tmax values (K). Values for the historical (black points) period and the future period (orange points) are given. Points lying on the y = *x* (green) line require no correction via downscaling.

For the case of raw values (Fig. 4a), during the historical (training) period the black curve shows that for low values the GCM exhibits little bias (i.e., lies near the green line). As Tmax increases, the (negative) bias of the GCM values increases monotonically. The relationship between GCM and OBS is similar in the future (orange) except that the curve is shifted to the right. The problem for BCQMr is twofold: 1) where historical and future values overlap, the nature of the relationship is different (i.e., the black curve is farther above the green line than is the orange curve), and 2) the historical transfer function (black) covers only the lower portion of future (orange) values. To overcome problem 2 and extend beyond the historical maximum, a constant offset (Deque 2007) is used (i.e., the distance of the last "good point" above the $y = x$ line is applied to all values greater than the historical maximum). This explains why in Fig. 3b the cyan histogram is almost a rightward-shifted version of the red one; a constant correction is used for most of the distribution. While examining Fig. 4b it is easy to see why the performance of BCQMa is much better than BCQMr. For the anomaly case (Fig. 4b), the future (orange) curve is shifted to the left and now fairly closely aligns with the historical (black) curve. Hence, the relationship available for training in the historical period is close to what is needed for application in the future.

To understand why the simple tail scheme was adopted, we must take a historical perspective. These types of SD methods were typically developed some time ago for applications seeking to refine data from the past. The implicit reasoning behind them is that one would not expect values from two samples (training and application) from the same period to have many out-of-bounds values compared to each other. However, when applied to multidecadal climate change projections, there is potential for situations as we have seen here in which this simple scheme goes awry.

The "mountain snow effect." In Fig. 1b we saw a secondary maximum of future downscaling error during the spring. For a representation of the related spatial aspects, we examine in Fig. 5 the ratio of May MAE[7] to annually averaged MAE for BCQMr. The extent to which this quantity is larger than 1 indicates the degree to which May errors are larger than those from the year as a whole. The most prominent aspects

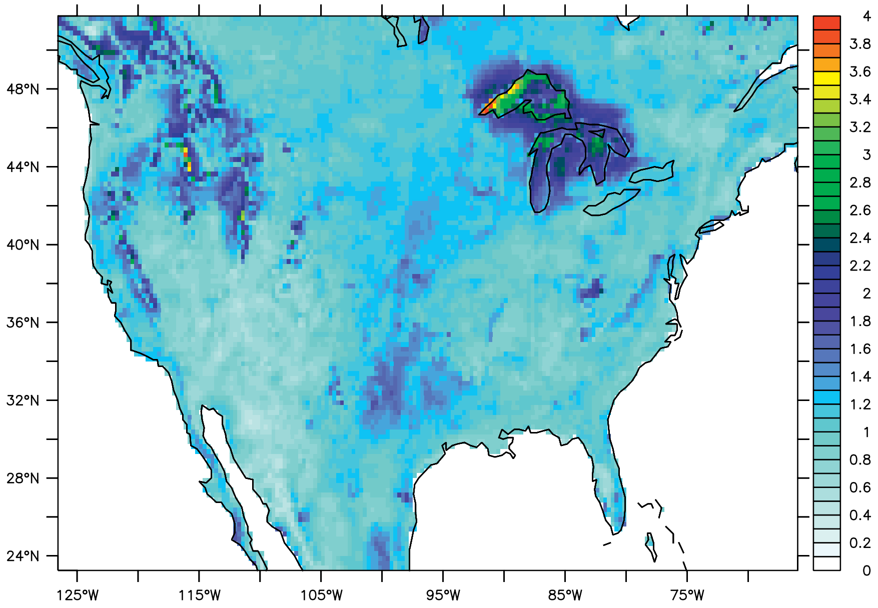


Fig. 5. Ratio (May/annual) of BCQMr MAE in Tmax for 2086–95.

[7] The MAE is computed by averaging, over the selected period (May or an entire year), the absolute value of the error for each day of each year.

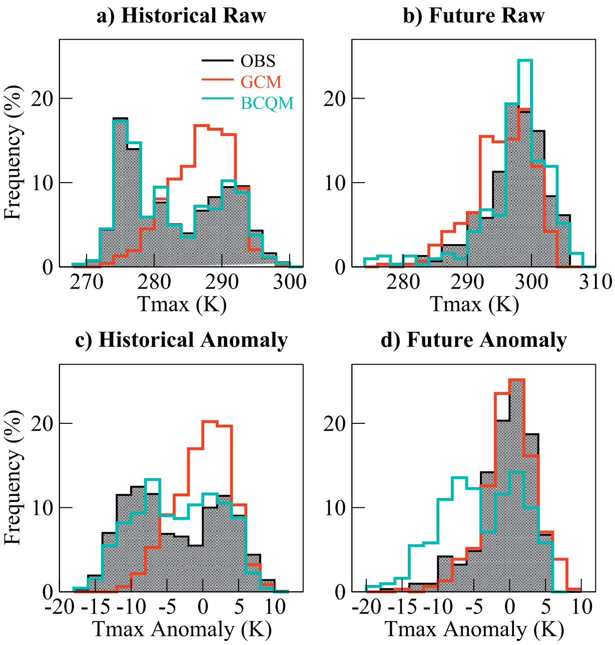


Fig. 6. As in Fig. 3, but for a grid point near Carbondale (39.4°N, 107.2°W) during May.

are the maxima over the mountainous regions of the west and the Great Lakes.[8] For further study we select a relatively high altitude point in an area having strong elevation gradients, located near Carbondale, Colorado.

Histograms for Carbondale, analogous to the ones seen earlier, are shown in Fig. 6. Having seen how the anomaly approach substantially mitigated the problems associated with the coastal effect, we begin with the expectation that BCQMa will be superior to BCQMr. For the historical anomaly case (Fig. 6c), the distributions are bimodal for the OBS but unimodal for the GCM. BCQMa is able to handle this challenge reasonably well. For the future (Fig. 6d) both OBS and GCM have very similar unimodal distributions. However, BCQMa quite erroneously produces bimodal results! By contrast, the BCQMr results are reasonably good for both the historical (Fig. 6a) and future (Fig. 6b) periods despite the substantial differences in distributional properties between the historical training period and the future application period.

The *x*–*y* plot in Fig. 7a helps illustrate the challenge that any statistical downscaling method faces at Carbondale in May. Historical points (days) for which there is no snow cover are depicted in black, while those with snow cover are depicted in red; future points, all snow free, are orange. The *y* = *x* reference line is green. During the historical training period, the snow-free points follow the reference line quite well, whereas the points with snow cover depart in a manner indicative of OBS colder than GCM, often by a substantial amount (~10 K). The future points, although exhibiting greater scatter, seem reasonably compatible with the snow-free historical points and the reference line. The challenge that any downscaling method faces is that there are two dominant relationships. Except for more sophisticated approaches, most downscaling methods will develop a training relationship in the historical period that is a compromise between the snow-covered and snow-free cases. The *x*–*y* plot in Fig. 7b is based on anomaly data. The crucial difference is that because of the normalization, the future points are shifted back toward the region where the divergence between snow-covered and snow-free clusters in the historical period is largest.

It is worth noting in Fig. 7a that at the lowest temperatures, the two historical clusters are not that far apart. As the temperature increases, the clusters increasingly diverge until some point (~294 K) at which the snow-free relationship becomes dominant. This transition occurs abruptly. From a physical standpoint, we expect that the high

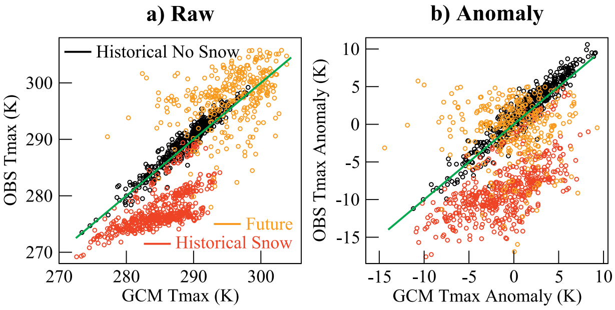


Fig. 7. Tmax (K) in May at Carbondale, with GCM on the abscissa and OBS on the ordinate, for historical snow-free (black), historical snow-covered (red), and future snow-free (orange) days. Based on (a) raw and (b) anomaly data. The *y* = *x* reference line (green) is indicated. A snow day is defined as one for which the water equivalent on the ground for the OBS grid point is at least 0.1 in. (0.254 cm); assuming a 10-to-1 snow-to-water ratio, this corresponds to 1 in. (2.54 cm) of snow cover.

[8] The maxima in and around the Great Lakes are not investigated here. In the GCM that has been employed (Dixon et al. 2016), the simplified treatment of large lakes does not reproduce a realistic seasonal cycle of ice cover (Milly et al. 2014), rendering these features somewhat suspect.

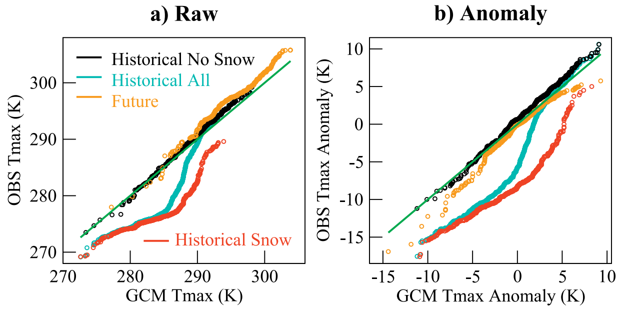


Fig. 8. Carbondale May Tmax Q–Q plots for data in (a) raw and (b) anomaly form. Abscissa corresponds to GCM and the ordinate to OBS Tmax values (K). Historical snow-free days (black), historical snow-covered days (red), and future snow-free days (orange). Cyan curve is based on all historical days; thus, it is based on the union of points from the black and red curves. Points lying on the *y* = *x* (green) line require no correction via downscaling. Definition of a snow-covered day is as in Fig. 7.

albedo of the snow affects the energy balance such that snow-covered days have lower temperature; however, latent heating effects associated with phase transition may play a role as well. The larger footprint of the GCM grid point captures lower-elevation regions that are snow free, so the GCM has a warm bias compared to OBS (3.58-K average in May historical).

Previous studies employing high-resolution models have shown that snow albedo feedback will likely amplify warming in mountain areas (e.g., Salathé et al. 2008; Walton et al. 2017) and that typical GCMs do not have sufficient resolution to properly simulate snow cover in the complex topography of the mountainous western United States. However, the model used in this study, which has a higher resolution than most contemporaneous GCMs used in century time-scale studies, is able to sufficiently capture the relevant effects.

Figure 8 shows Q–Q plots for Carbondale in May, which are similar in nature to the earlier ones for Sixmile Bend. The color scheme is the same as the prior figure except that there is an additional curve (cyan) based on the merger of the historical snow and snow-free clusters. For both panels the overall historical (cyan) curve represents a compromise between the snow-covered (red) and snow-free (black) curves. The compromise curve is similar to the snow-covered curve at lower temperatures and the snow-free curve at higher temperatures, with a sharp transition at the high end.

For the anomaly case (Fig. 8b), most of the future values (orange) depart significantly from the training relationship (cyan), which is heavily influenced by the all-snow cases (red). Only a relatively small fraction of the future points, at the higher end, have a relationship compatible with that for the training; this results in substantial error because, as a result of climate change warming, in the future period being analyzed here there is no snow cover at Carbondale in May.

The BCQMr approach (Fig. 8a) is much better suited because the majority of future points (orange) are downscaled using a relationship (cyan) heavily influenced by the snow-free conditions (black). Nevertheless, it would seem that the most common SD methods are not optimally suited to handle such situations—perhaps a conditional approach, in which the transfer relationship depends on external factors (such as the presence or absence of snow cover), would lead to further improvement.

DISCUSSION. In summary, we have seen two distinct cases in which incorporation of the “change factor” paradigm either substantially improves or degrades the downscaling results in the future. Furthermore, we have demonstrated that it would be difficult—if not impossible—to divine the existence of this paradox using historical data alone. Only through the use of a perfect model experimental design were we able to both identify and explain this oddity. These results are exemplified in Table 1, which gives the skill (Wilks 2006) for the two locations highlighted earlier. This illustrates how the two distinct approaches (BCQMr vs BCQMa) yield similar results in the historical period but very different results for the future. In one case BCQMr is much better than BCQMa, and vice versa.

The practical implications of these findings may be considerable. Our work pertains to the distributional class of downscaling methods. However,

Table 1. Downscaling skill (dimensionless) based on the root-mean-square error (rmse) for Tmax at Sixmile Bend in Jul and Carbondale in May, for BCQMr and BCQMa. Historical skills are based on cross validation.

	Sixmile Bend (Jul)		Carbondale (May)	
	BCQMr	BCQMa	BCQMr	BCQMa
Historical	42.8	39.9	29.0	21.8
Future	5.1	42.1	71.2	24.7

many—if not most—of the more complex downscaling methods incorporate a distributional component. The methods potentially affected by our findings include, but are not limited to, asynchronous regional regression model (ARRM; Stoner et al. 2013), bias-corrected spatial disaggregation (BCSD; Wood et al. 2004), BCSD daily (Thrasher et al. 2012), bias-corrected constructed analog (BCCA; Maurer et al. 2010), multivariate adaptive constructed analogs (MACA; Abatzoglou and Brown 2012), and localized constructed analogs (LOCA; Pierce et al. 2014). Collectively these methods account for the lion's share of U.S. downscaled data available on public servers. However, without actual application of these methods in our PM test bed, it is not known to what extent our findings affect these methods.[9]

For several reasons the impacts of the pitfalls we found could have a disproportionately large practical effect. A relatively large fraction of the U.S. population lives in coastal regions (NOAA 2013). In coastal regions of the Gulf and the Southeast, where warm-season temperature variance is small, the magnitudes of the errors seen for Sixmile Bend could translate into relatively large errors in overstating implied impacts across some sectors, such as health, energy, and ecosystems (Fig. 3b). On the other hand, at Carbondale the errors imply a large underestimate of warming. In a region that is already water stressed, with concern for increasing stress in the future (Melillo et al. 2014), this might imply an underestimate of important hydrological changes. Although our GCM output is from the 2090s, these errors will be introduced gradually long before the 2090s. Indeed, Ashfaq et al. (2010) found that biases in dynamical model output can significantly affect implied hydrological responses in both historical and future periods.

It is important to keep in mind that by no means do our findings invalidate the utility of existing downscaled data. In fact, we find that in most cases downscaling adds considerable value. However, in certain well-defined situations a wide variety of downscaling methods have the potential to yield highly erroneous results. We have been able to define two such circumstances. For two reasons it is unknown to what extent other such circumstances exist: 1) we have yet to fully mine our PM archive, and 2) our initial PM design is quite simple—some of the quirks we have uncovered, as well as others that may be lurking beneath the surface, may manifest themselves in ways yet to be discovered in more realistic PM designs as well as in the real world.

The fact that our PM incorporates only the effect of differing spatial resolution between OBS and GCM enabled us to identify and explain certain pitfalls in SD. A more complex design may have made it more difficult to do so. However, in real-world applications of SD the mismatch in distributional and relational properties between OBS and GCM may lead to other pitfalls that are not so easily characterized or generalized as the coastal effect and the mountain snow effect, may have arbitrary spatial patterns and seasonal behavior, and may be entirely dependent upon a particular combination of model and observational datasets. Future designs of PM could be geared to this more general situation by pairing two different GCMs or by pairing one GCM and one RCM in lieu of the coarsening that was employed here. The GFDL ESD team, in conjunction with outside collaborators, is in the initial stages of such an effort involving the use of several RCM–GCM pairings produced under the auspices of the North American Coordinated Regional Climate Downscaling Experiment (CORDEX) program (https://na-cordex.org/).

While some might argue that improvements to physical models will reduce the need for SD and/or reduce the impact or likelihood of pitfalls, we would argue that this is not necessarily the case for several reasons, including the following:

- Greater sophistication and increased spatial resolution does not eliminate all dynamical model biases. For example, it has been shown that downscaling a GCM can produce better results than downscaling an RCM (Eden et al. 2014) and that downscaling an RCM can produce better results than using the raw RCM output (Turco et al. 2017).
- Some applications require climate information at a finer spatial resolution than GCMs used to produce projections on the multidecadal time scales will be able to produce for the foreseeable future because of computational limitations: increases in the sophistication of climate models go hand in hand with increases in computational resources.

[9] Based on their own PM design, Walton et al. (2017, their Fig. 10) demonstrate that both BCSD and BCCA perform worse than the use of raw GCM output in downscaling temperature in the Sierra Nevada in the late twenty-first century. This is consistent with the notion that the most popular SD methods fail as result of the mountain snow effect. However, the hybrid dynamical–statistical approach used by Walton et al. (2017) performs well in this regard.

- Ironicly, as the resolution of models increases, the resolution of some publically available data may decrease in a compensatory fashion. Because of increased data volume associated with higher-resolution models and demand for higher temporal frequency of products, data may be archived at lower than native resolution (i.e., subsampling data in both space and time). The result may be data that do not always match the needs of the myriad potential climate impact applications, for example, in regions of sharp horizontal gradients associated with discontinuities in underlying surface types.
- Many state-of-the-art GCMs utilize different grid configurations for different components, such as atmospheric and oceanic model components. This results in atmospheric grid cells overlying portions of both ocean and land cells, effectively blurring sharp horizontal gradients for some climate variables along coastlines.

Furthermore, to borrow a historical lesson, we note that the forerunners of SD—that is, model output statistics (MOS)—were first applied operationally to numerical weather prediction (NWP) in the early 1970s (Benestad et al. 2008). Since then, although the skill of NWP has increased dramatically (Harper et al. 2007), MOS products have proliferated and are used far more extensively than ever before (www.weather.gov/mdl/mos_home). This argues for an expansion of efforts to scrutinize SD methods. Some coordinated efforts for the European sector have recently been launched (Maraun et al. 2015). The ESD team at GFDL and its collaborators will continue their contribution to efforts of this nature, most immediately by reporting elsewhere in greater detail more extensive evaluations that led to the discovery of the pitfalls reported in this paper. We are also just beginning evaluation of daily precipitation. Given the more complex nature of the frequency distribution of precipitation, it would not be surprising to find additional pitfalls. Finally, we are in the initial stages of developing a more complex PM design.

The research presented here provides examples of how the occurrence of statistical downscaling pitfalls can vary geographically, with time of year, climate conditions, and across SD techniques. Other work (not shown) reveals that the performance of the SD method also can vary markedly with the amount of model-simulated climate change, across different climate variables, and be influenced by preprocessing and training step methodological choices that typically do not garner much attention. Viewed from the perspective of climate impact studies, the question of whether a particular pitfall may be a serious concern depends on the details of a study's climate data needs and sensitivities—a factor that can preclude simple one-size-fits-all guidance. However, increased awareness that pitfalls and nonstationarities of the type exposed in these perfect model experiments exist can help promote the better-informed use of statistically downscaled climate projections.

What kind of pitfalls have yet to be discovered? We can only speculate. As an example, inspired by Hall (2014), consider the region affected by snow induced by the Great Lakes. But cold-season precipitation there is influenced by an additional mechanism: that from large-scale cyclonic systems. Suppose a GCM or RCM performed well in simulating precipitation from large-scale systems but not as well with that induced by lake effects. In the historical period, SD might be based on a compromise relationship, as seen for Carbondale, but yield acceptable results. In the future, as a result of decreasing baroclinicity, the importance of large-scale precipitation might decline. But with an increase in lake temperatures and a longer ice-free season, the relative importance of lake-effect snows might increase. The combination of these two changes might result in much poorer SD results in the future, something not readily apparent from historical diagnostics alone.

ACKNOWLEDGMENTS. This work was supported in part by the USGS South Central Climate Science Center (G13AC00387). We thank Isaac Held and Tom Delworth for their comments on an earlier draft of this manuscript as well as Daniel Vimont and an anonymous reviewer for *BAMS*. This work benefited from discussions with Katherine Hayhoe and Anne Stoner.

REFERENCES

Abatzoglou, J. T., and T. J. Brown, 2012: A comparison of statistical downscaling methods suited for wildfire applications. *Int. J. Climatol.*, **32**, 772–780, https://doi.org/10.1002/joc.2312.

Ashfaq, M., L. C. Bowling, K. Cherkauer, J. S. Pal, and N. S. Diffenbaugh, 2010: Influence of climate model biases and daily-scale temperature and precipitation events on hydrological impacts assessment: A case study of the United States. *J. Geophys. Res.*, **115**, D14116, https://doi.org/10.1029/2009JD012965.

Benestad, R. E., D. Chen, and I. Hanssen-Bauer, 2008: *Empirical-Statistical Downscaling.* World Scientific Publishing Co., 228 pp., https://doi.org/10.1142/6908.

Dayon, G., J. Boé, and E. Martin, 2015: Transferability in the future climate of a statistical downscaling method for precipitation in France. *J. Geophys. Res. Atmos.*, **120**, 1023–1043, https://doi.org/10.1002/2014JD022236.

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 Planet. Change*, **57**, 16–26, https://doi.org/10.1016/j.gloplacha.2006.11.030.

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

Eden, J. M., M. Widmann, D. Maraun, and M. Vrac, 2014: Comparison of GCM- and RCM-simulated precipitation following stochastic postprocessing. *J. Geophys. Res. Atmos.*, **119**, 11 040–11 053, https://doi.org/10.1002/2014JD021732.

Efron, B., 1982: *The Jackknife, the Bootstrap, and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 38, SIAM, 92 pp., https://doi.org/10.1137/1.9781611970319.

Fowler, H., S. Blenkinsop, and C. Tebaldi, 2007: Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.*, **27**, 1547–1578, https://doi.org/10.1002/joc.1556.

Frías, M. D., E. Zorita, J. Fernández, and C. Rodríguez-Puebla, 2006: Testing statistical downscaling methods in simulated climates. *Geophys. Res. Lett.*, **33**, L19807, https://doi.org/10.1029/2006GL027453.

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

Harper, K., L. W. Uccellini, L. Morone, E. Kalnay, and K. Carey, 2007: 50th anniversary of operational numerical weather prediction. *Bull. Amer. Meteor. Soc.*, **88**, 639–650, https://doi.org/10.1175/BAMS-88-5-639.

Ho, C.-K., D. B. Stephenson, M. Collins, C. A. T. Ferro, and S. J. Brown, 2012: Calibration strategies: A source of additional uncertainty in climate change projections. *Bull. Amer. Meteor. Soc.*, **93**, 21–26, https://doi.org/10.1175/2011BAMS3110.1.

Ivanov, M. A., and S. Kotlarski, 2017: Assessing distribution-based climate model bias correction methods over an alpine domain: Added value and limitations. *Int. J. Climatol.*, **37**, 2633–2653, https://doi.org/10.1002/joc.4870.

Maraun, D., 2012: Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. *Geophys. Res. Lett.*, **39**, L06706, https://doi.org/10.1029/2012GL051210.

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

——, and Coauthors, 2015: VALUE: A framework to validate downscaling approaches for climate change studies. *Earth's Future*, **3**, 1–14, https://doi.org/10.1002/2014EF000259.

Maurer, E. P., H. G. Hidalgo, and T. Das, 2010: The utility of daily large-scale climate data in the assessment of climate change impacts on daily streamflow in California. *Hydrol. Earth Syst. Sci.*, **14**, 1125–1138, https://doi.org/10.5194/hess-14-1125-2010.

Melillo, J. M., T. C. Richmond, and G. W. Yohe, Eds., 2014: Climate change impacts in the United States: The Third National Climate Assessment. U.S. Global Change Research Program, 841 pp., https://doi.org/10.7930/J0Z31WJ2.

Milly, P. C. D., and Coauthors, 2014: An enhanced model of land water and energy for global hydrologic and earth-system studies. *J. Hydrometeor.*, **15**, 1739–1761, https://doi.org/10.1175/JHM-D-13-0162.1.

NOAA, 2013: National coastal population report: Population trends from 1970 to 2020. NOAA's State of the Coast, National Oceanic and Atmospheric Administration, 19 pp., http://oceanservice.noaa.gov/facts/coastal-population-report.pdf.

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

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

——, ——, E. P. Maurer, J. T. Abatzoglou, and K. C. Hegewisch, 2015: Improved bias correction techniques for hydrological simulations of climate change. *J. Hydrometeor.*, **16**, 2421–2442, https://doi.org/10.1175/JHM-D-14-0236.1.

Salathé, E. P., Jr., R. Steed, C. F. Mass, and P. H. Zahn, 2008: A high-resolution climate model for the U.S. Pacific Northwest: Mesoscale feedbacks and local responses to climate change. *J. Climate*, **21**, 5708–5726, https://doi.org/10.1175/2008JCLI2090.1.

Stoner, A. M. K., K. Hayhoe, X. Yang, and D. J. Wuebbles, 2013: An asynchronous regional regression model for statistical downscaling of daily climate variables. *Int.*

J. Climatol., **33**, 2473–2494, https://doi.org/10.1002/joc.3603.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Thrasher, B., E. P. Maurer, C. McKellar, and P. Duffy, 2012: Bias correcting climate model simulated daily temperature extremes with quantile mapping. *Hydrol. Earth Syst. Sci.*, **16**, 3309–3314, https://doi.org/10.5194/hess-16-3309-2012.

Turco, M., C. Llasat, S. Herrera, and J. M. Gutiérrez, 2017: Bias correction and downscaling of future RCM precipitation projections using a MOS-Analog technique. *J. Geophys. Res. Atmos.*, **122**, 2631–2648, https://doi.org/10.1002/2016JD025724.

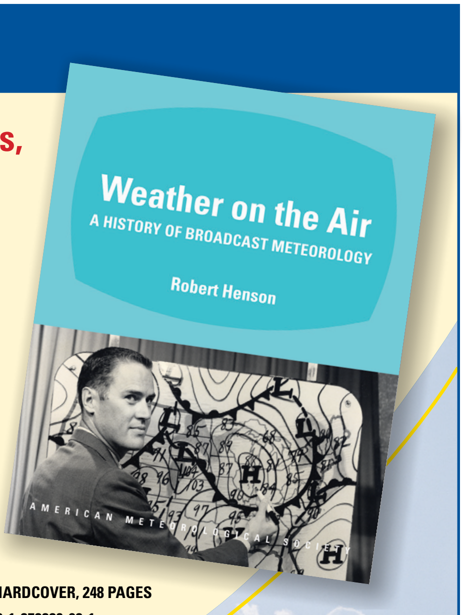
Velazquez, J. A., M. Troin, D. Caya, and F. Brissette, 2015: Evaluating the time-invariance hypothesis of climate model bias correction: Implications for hydrological impact studies. *J. Hydrometeor.*, **16**, 2013–2026, https://doi.org/10.1175/JHM-D-14-0159.1.

Vrac, M., M. Stein, K. Hayhoe, and X.-Z. Liang, 2007: A general method for validating statistical downscaling methods under future climate change. *Geophys. Res. Lett.*, **34**, L18701, https://doi.org/10.1029/2007GL030295.

Walton, D. B., A. Hall, N. Berg, M. Schwartz, and F. Sun, 2017: Incorporating snow albedo feedback into downscaled temperature and snow cover projections for California's Sierra Nevada. *J. Climate*, **30**, 1417–1438, https://doi.org/10.1175/JCLI-D-16-0168.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 676 pp.

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


Weather on the Air
A HISTORY OF BROADCAST METEOROLOGY
Robert Henson
AMERICAN METEOROLOGICAL SOCIETY

AMS titles now available as eBooks at springer.com

AMS BOOKS

RESEARCH APPLICATIONS HISTORY

www.ametsoc.org/amsbookstore

Springer

Scan to see
AMS eBook titles
at springer.com

Springer

HOME | MY SPRINGER | SUBJECTS | SERVICES | IMPRINTS & PUBLISHERS | ABOUT US

» New User
LOGIN

Search...
GO
Advanced Search

» American Meteorological Society

Home > Societies & Publishing Partners > American Meteorological Society

AMS BOOKS
RESEARCH APPLICATIONS HISTORY

American Meteorological Society

Promoting the development and dissemination of information and education on the atmospheric and related sciences.

American Meteorological Society eBooks

AMS Books grew out of a monograph publishing program related to the peer-reviewed journals published by the AMS and now encompasses a series of meteorological and historical monographs that reaches back five decades (formerly out-of-print volumes of which have been brought back on SpringerLink as eBooks) as well as a growing list of academic, policy, and general-interest books related to the atmospheric sciences.

Sort listing by: ○ Relevance ● Publication date ○ Title ○ Author ○ Copyright Year

Export to CSV

AMERICAN METEORLOGICAL SOCIETY EBOOKS

» Show all 11 results

Eloquent Science

A Practical Guide to Becoming a Better Writer, Speaker, and Atmospheric Scientist

Schultz, David
2009

Price from $45.00

Available Formats:

AMERICAN METEOROLOGICAL SOCIETY

» About the American Meteorological Societ...
» Order print copies of AMS Books
» Booksellers seeking print copies of AMS ...
» For AMS Journals Online

RESOURCES

Stay Informed
Receive notification of new releases from AMS Books

Submit a Book Proposal to the AMS
Download our book proposal guidelines

Become an AMS Member
Learn about the benefits of membership

CONNECT WITH THE AMERICAN METEOROLOGICAL SOCIETY

Like AMS on Facebook
Follow AMS on Twitter

AMERICAN METEOROLOGICAL SOCIETY