Extreme Event Verification for Probabilistic Downscaling

MEGAN C. KIRCHMEIER-YOUNG
Department of Atmospheric and Oceanic Sciences, and Nelson Institute Center for Climatic Research, University of Wisconsin–Madison, Madison, Wisconsin

DAVID J. LORENZ
Nelson Institute Center for Climatic Research, University of Wisconsin–Madison, Madison, Wisconsin

DANIEL J. VIMONT
Department of Atmospheric and Oceanic Sciences, and Nelson Institute Center for Climatic Research, University of Wisconsin–Madison, Madison, Wisconsin

(Manuscript received 18 January 2016, in final form 17 August 2016)

ABSTRACT

Extreme events are important to many studying regional climate impacts but provide a challenge for many “deterministic” downscaling methodologies. The University of Wisconsin Probabilistic Downscaling (UWPD) dataset applies a “probabilistic” approach to downscaling that may be advantageous in a number of situations, including realistic representation of extreme events. The probabilistic approach to downscaling, however, presents some unique challenges for verification, especially when comparing a full probability density function with a single observed value for each day. Furthermore, because of the wide range of specific climatic information needed in climate impacts assessment, any single verification metric will be useful to only a limited set of practitioners. The intent of this study, then, is (i) to identify verification metrics appropriate for probabilistic downscaling of climate data; (ii) to apply, within the UWPD, those metrics to a suite of extreme event statistics that may be of use in climate impacts assessments; and (iii) in applying these metrics, to demonstrate the utility of a probabilistic approach to downscaling climate data, especially for representing extreme events.

1. Introduction

While extreme events are of interest for many climate impacts studies, global coupled general circulation models (GCMs) often mute extremes with their coarse resolution, particularly for precipitation (e.g., Sillmann et al. 2013), and alone are not sufficient for assessments on a regional or local level. Downscaling methodologies aim to solve this scale mismatch issue by extracting high-resolution information from the coarse-resolution simulations. Downscaling methods can be roughly characterized as dynamical or statistical: dynamical downscaling methods utilize a higher-resolution model that is forced by the large-scale GCM (e.g., Giorgi 2006), while statistical downscaling fits an empirical relationship that predicts the local-scale information as a function of various large-scale predictors. There are advantages to either type of downscaling technique; a complete review of downscaling methodologies can be found in Maraun et al. (2010b).

In a focus on statistical downscaling, we draw a further distinction between deterministic and probabilistic statistical downscaling methodologies. Probabilistic methods are defined as those that output a full probability density function (PDF) for each time step, in contrast to deterministic methods that output a unique value for each time step (usually the conditional expectation). With the goal of characterizing the uncertainty at the local scale that cannot be fully explained by the large-scale conditions, the probabilistic method’s daily varying PDFs represent the range of what could have occurred at the local scale on each day, given the associated large-scale
conditions. Despite the wide set of existing verification metrics for downscaled data and for weather forecasting, probabilistic downscaling presents its own set of challenges for verification techniques, as there is only a single observed value for each day available for comparison. Thus, common goodness-of-fit tests would have zero degrees of freedom.

In many ways, the daily varying PDFs from a probabilistic downscaling methodology can be likened to probabilistic forecasts. Numerous metrics for evaluating probabilistic forecasts have been developed (see summaries in Broecker 2012; Gneiting and Katzfuss 2014; Machete 2013; Wilks 2006) and are widely used to assess probabilistic forecasts ranging in complexity from the probability of occurrence of a specific event to full probability distributions to describe the variable of interest. Many of these metrics involve the calculation of a “skill score” relative to a reference forecast, which is often the climatological normal (“climatology”) or the previous value (“persistence”) (Wilks 2006). Some of these forecast verification metrics are directly applicable to probabilistic downscaling output, while others are less useful for evaluating daily varying PDFs of climate variables. While verification via diagram metrics provides useful information regarding performance (e.g., Pinson et al. 2010; Wong et al. 2014), these metrics can be impractical for large domains and broad sets of variables.

In addition to evaluating the daily varying PDFs, probabilistic methodologies should also be evaluated on the ability of random realizations drawn from their PDFs to represent climate statistics. Stochastic realizations are common with weather-generator-type downscaling methods (Maraun et al. 2010b; Harris et al. 2014), and several studies have suggested methods/metrics for their evaluation (e.g., Chen and Brissette 2014; Min et al. 2011; Langousis and Kaleris 2014; Wong et al. 2014). Examples of the evaluation of sets of realizations using metrics designed for deterministic output include averaging the skill across many realizations (Haylock et al. 2006) or choosing a threshold for the number of realizations that need to “pass” a particular test (Bürger et al. 2012). Furthermore, sets of stochastic realizations can also be considered similarly to ensemble forecasts (Maraun et al. 2015) and verification methods from the ensemble evaluation literature can be utilized; see Brown et al. (2010) for a list of metrics. There have been several studies comparing the performance of different downscaling methods/datasets (e.g., Bürger et al. 2012; Gutmann et al. 2014; Godess et al. 2012; Tryhorn and DeGaetano 2011; Sunyer et al. 2012). Furthermore, in attempts to standardize downscaling evaluation, Hayhoe (2010) argued that methods should be tested on their ability to represent the mean values at various time scales, threshold values or extremes, and the persistence of events, while Maraun et al. (2015) developed a user-focused system to facilitate comparisons between downscaling methods in European domains.

The need to consider user interests when evaluating downscaling methodologies has been recognized (Barsugli et al. 2013; Maraun et al. 2015). In particular, many users studying climate and climate impacts are concerned with extreme events. For example, temperature extremes can be a threat to public health (Patz et al. 2005), agriculture (Rosenzweig et al. 2001), and ecosystems (Parmesan et al. 2000), while extreme precipitation events have impacts on storm water management [Wisconsin Initiative on Climate Change Impacts (WICCI) 2011], agriculture (Rosenzweig et al. 2001), ecosystems (Parmesan et al. 2000), and lakeside communities (WICCI 2011). Furthermore, extreme events can have additional indirect influences (e.g., Patz et al. 2005; Curriero et al. 2001) and the timing of these events can also enhance the impacts (e.g., Rosenzweig et al. 2001). While there are many metrics used to define extreme events from daily data, the choice of metric often depends on the situation. Extreme events can be defined as single days that exceed a specified threshold or can focus on longer-duration events (e.g., several consecutive days that all exceed a threshold). Extreme events can also be defined as the top 5% or 10% of days or annual minima and maxima. Additional metrics involve lists of common indices (Zhang et al. 2011) or probability distributions of extreme events, such as the generalized extreme value (GEV) distribution or generalized Pareto distribution (GPD). Using many metrics to describe extreme events and allowing flexibility (e.g., in the choice of a threshold) will contribute to an evaluation that is both useful and relevant.

In this study, we focus on a probabilistic downscaled dataset developed at the University of Wisconsin–Madison. This University of Wisconsin Probabilistic Downscaling (UWPD; Lorenz 2015; Notaro et al. 2011) was motivated by the needs of those studying climate and climate impacts as part of WICCI and has since been expanded to cover a larger domain. The UWPD dataset has been used for numerous studies of climate impacts (WICCI 2011; Patz et al. 2014; Schuster et al. 2012; Vavrus and Behnke 2014; Sharma et al. 2011; Notaro et al. 2014) and has proved valuable for the analysis of extreme events (Vavrus et al. 2015). The novelty of this dataset lies in its probabilistic output over a large spatial domain and its ability to accurately represent user-relevant extreme events.

The purpose of this manuscript is threefold: (i) to identify a set of metrics that can be used to evaluate a
probabilistic downscaled dataset, (ii) to apply those methods to a wide range of extreme event characteristics, and in doing so (iii) to demonstrate the utility of a probabilistic downscaling approach for representing extreme event statistics. This paper is directed toward the downscaling community, those both developing and using downscaled datasets, with the goals of (i) demonstrating the advantages of a probabilistic methodology, especially its flexibility and ability to represent extremes; (ii) highlighting a multitude of ways to evaluate a probabilistic downscaled dataset; and (iii) further emphasizing that different users might care about different characteristics of such a dataset. In support of the second goal, we note that a great deal of work has been done in developing verification metrics for probabilistic forecasts; we demonstrate the utility of using these metrics in evaluating probabilistic downscaled data, which is not yet commonplace in downscaling communities.

This paper is organized as follows. Descriptions of the probabilistic downscaling routine and the datasets used for verification are found in section 2. The verification methodology is presented in section 3, and is divided into two categories: tests on the distribution are presented in section 3b and tests on the realizations are presented in section 3c. Section 3c also identifies a set of metrics to describe the characteristics of extreme events in temperature and precipitation. A brief comparison between the probabilistic method and a commonly used deterministic method is included in section 3d. Conclusions are discussed in section 4.

2. Methodology

a. Downscaling routine

The UWPD (Lorenz 2015; Notaro et al. 2014) takes a probabilistic approach via vector generalized linear models (VGLMs; Yee and Stephenson 2007) to estimate daily varying PDFs for station maximum and minimum temperature and daily precipitation. The VGLM is trained using the large-scale environmental conditions from the National Centers for Environmental Prediction–National Center for Atmospheric Research reanalysis product (Kalnay et al. 1996) and daily maximum and minimum temperature and precipitation from the National Oceanic and Atmospheric Administration Cooperative Observer Network (COOP; Robinson 1990). The downscaled data cover the eastern United States (Fig. 1), with historical data for 1950–2009. The VGLM methodology is similar to that presented in Kirchmeier et al. (2014) for daily wind speed and links (daily varying) large-scale environmental conditions to the parameters of an appropriate probability distribution for local-scale variables. Daily maximum (Tmax) and minimum (Tmin) temperature were modeled using a normal distribution with the 2-m estimate of Tmax or Tmin, the 1000-m wind vector, and column-integrated relative humidity as the large-scale predictors. Precipitation was downscaled in a two-step process, modeling the Bernoulli precipitation occurrence and then the precipitation amount with a generalized gamma distribution (Stacy 1962). The collection of large-scale predictors for precipitation is discussed in Notaro et al. (2014). Several other studies have used generalized linear models to describe precipitation amount (Abaurrea and Asín 2005; Fealy and Sweeney 2007; Beuchat et al. 2012), but the UWPD method goes a step further in allowing multiple parameters of the gamma distribution to vary as a function of the large-scale predictors. Focusing on extremes, Maraun et al. (2010a) also used a VGLM to downscale precipitation, but they predicted the parameters of a GEV distribution. The UWPD downscaling model was fit for each station and the resulting distribution parameters (e.g., daily varying mean and standard deviation for temperature) were interpolated to a 0.1° × 0.1° grid. The downsampling methodology is trained on station data (the COOP dataset) and as such, the downscaled PDFs represent a point estimate of local-scale conditions as opposed to the gridbox average output by the GCMs and reanalyses.

The underlying premise for using a probabilistic approach is that local-scale variations cannot be perfectly explained by large-scale variables. Instead, the probabilistic approach quantifies a range of local-scale values that could have occurred (the PDF of the local-scale variable), given the large-scale conditions for that particular day. Thus, the result of the downscaling routine
is a daily varying PDF that is conditioned on the daily varying large-scale environmental conditions. An example of the daily varying PDFs is shown in Fig. 2. The downscaled PDF is displayed for each of four consecutive days (in January 1996) at one station (Madison, Wisconsin), with an \( \times \) indicating the value observed at the station. The local scale cannot be entirely explained by large-scale variables and the goal of the probabilistic methodology is to account for this uncertainty. Because of the probabilistic nature of the results, it is neither expected nor desired that each observed value be located at the peak of the PDF, as the observed value is only one possibility of what could have occurred at the local scale with the given large-scale meteorological conditions.

A key advantage of a probabilistic method is the flexibility it provides to the users of the downscaled data. In addition to analyzing the daily means and standard deviations, the daily varying PDFs can also be used to calculate the probability of experiencing a particular event on each day or to rescale a set of existing observations to a different climate scenario while maintaining the same sequence of events. Furthermore, time series realizations, similarly to the output of weather generator routines (Maraun et al. 2010b), can also be extracted. These realizations produce a single value for each day and represent a possibility of what could have occurred. Because the realizations are extracted by randomly sampling from the PDFs, each realization is just as likely as any other. An example of a set of realizations is shown in Fig. 3. The realizations (gray lines) will not match exactly the observations (black line), but to the extent that local-scale variations are influenced by the large-scale environmental conditions, they should follow the same general pattern as the observed line.

Realizations extracted via independent random sampling of the daily PDFs will possess some autocorrelation because of the persistence of the large scale and its influence on the daily varying distribution parameters. For Tmax and Tmin, the autocorrelation of the independently sampled realizations is too low relative to observations. To remedy this issue, realizations can be extracted in such a way that maintains the appropriate autocorrelation. For this downscaled dataset, the temporal autocorrelation is increased by drawing realizations using normally distributed random numbers that are correlated in time. Inspection of the temperature autocorrelation suggested that the random numbers should be the sum of red noise plus independent white noise and that is the form used here. No such correction is required for the precipitation realizations, as the observed autocorrelation of a time series of precipitation is low.

b. Verification dataset

Subsequent verification metrics will utilize both the daily varying PDFs directly and a set of many realizations, both of which will be evaluated against a collection of National Weather Service (NWS) first-order stations covering the region. First-order stations...
contain the highest quality records and observe a wide set of meteorological variables. A map of the 156 station locations is shown in Fig. 1; a list of the details for each station is presented in Table S1 of the online supplemental material. The stations included are missing less than 10% of their daily data for the period 1950–2009. For each station location, we obtain distribution parameters from the downscaled dataset for the grid box that contains the given station. Recall that the downscaled data are composed of gridded distribution parameters that represent point estimates rather than gridbox averages and hence can be compared with the station data. The distribution parameters are used either directly or to extract realizations from the downscaled data that are then compared with the station observations.

We note that the 156 NWS first-order stations are included in the COOP dataset (~4000 stations) that was used to train the downscaling methodology, and as such, the verification dataset is not formally independent. Clearly, a verification dataset that is independent in all ways from the training dataset would be ideal for a formal evaluation of a specific downscaling methodology, and this is a caveat of the current study. In this manuscript, we discuss the performance of this particular dataset in detail, but the intent is not the evaluation of the dataset itself. Instead, the intent of this study is to describe the application of verification metrics for probabilistic downscaling and to illustrate the utility of a probabilistic approach for representation of extreme events.

3. Results

In this section, we present verification metrics for the probabilistic downscaled data. We begin with a simple example (section 3a) that demonstrates both the character of probabilistic data and the need for multiple verification metrics. The verification metrics for our probabilistic downscaling routine are divided into two categories: tests on the distribution (section 3b) and tests on the realizations (section 3c). Tests on the distribution aim to quantify how well the observed values fit into the daily PDFs, expressed in summary statistics across the time series. Tests on the realizations assess how well the downscaled data capture various climate statistics, summarized in Monte Carlo tests across many realizations.

a. Probabilistic versus deterministic example

Although the focus of this manuscript is on evaluating the representation of extremes, we begin with a simple verification example that highlights both the character of the probabilistic downscaling methodology and the need for numerous metrics for evaluation. As the probabilistic dataset consists of daily varying PDFs, we can either extract realizations from the data (as discussed above) or estimate the most likely value from the PDF (the mean of the PDF; we refer to this as a deterministic downsampling). For precipitation, we evaluate the means and variance of two estimates from the PDFs against observations using a simple Pearson correlation test ($R^2$) and an $F$ test (Fig. 4a). To simulate the deterministic output for a precipitation time series, the mean from each day’s precipitation amount distribution was assigned. Although this results in too many days of very light rainfall, several other deterministic methodologies have the same issue, particularly with a 0-mm threshold (Gutmann et al. 2014). The interpretation of the following results would be the same if a probabilistic component were first used to determine the occurrence of rainfall before assigning the mean (deterministic) amount.

The $R^2$ value describes a pointwise linear relationship between two datasets (e.g., downscaled prediction and observations) and is the square of the Pearson correlation coefficient. Values of $R^2$ range from 0 to 1, with larger values implying a stronger correlation between the two datasets. The $R^2$ value was calculated between the observed values of precipitation and either a single realization of the probabilistic downscaled data (probabilistic) or the mean of the downscaled PDFs (deterministic). The $R^2$ value was calculated for all NWS first-order stations and the resulting spread of $R^2$ values for January and July is plotted in Fig. 4a as a box plot. The edges of the box indicate the 25th and 75th percentile values ($p25$ and $p75$, respectively) across all stations for the given statistic (the $R^2$ value, in this case) and the dashed whiskers extend to the smaller of 1.5 times the length of the interquartile range (IQR; $p75–p25$) or the most extreme value. Additional outliers outside the whiskers are indicated with asterisks. Figure 4a shows that both the deterministic downsampling and the realization from the probabilistic downscaling better match observations during January than July, which is consistent with more synoptic-scale control of (midlatitude) precipitation during boreal winter and small-scale convective precipitation during boreal summer. Furthermore, the deterministic output produces consistently higher $R^2$ values relative to observations and thus produces a better representation of what actually occurred.

Despite the higher $R^2$ metric for the deterministic downsampling, many users are also concerned with realizations possessing appropriate variability and accordingly, an $F$ statistic was also calculated. The $F$ statistic is a ratio of the variance of the downscaled time series to that of the observed time series. As such, a ratio of 1 indicates equal variance, while a ratio less than 1 implies an underestimation of the variance by the downscaled prediction. As seen in Fig. 4a, the
The probabilistic realization possesses a variance that often agrees well with that from observations. In contrast, the deterministic realization grossly underestimates the observed variance, particularly during the summer. The deterministic data are constrained to always choose the (daily varying) mean, and as such, do not capture the local-scale variance that is independent of the large-scale conditions. In this respect, the probabilistic realization performs “better.”

The comparison can be extended to extreme events, which are more relevant for some climate impacts. The observed 99th percentile value (p99) of precipitation amount on wet days for each station is plotted against the p99 from each downscaled precipitation realization in Fig. 4b for January and July. The probabilistic method produces extreme values that agree well with those from observations, while the deterministic method almost uniformly underestimates the extreme values. Note especially that the deterministic method performs better during January than during July. This is also consistent with the synoptic-scale control over precipitation during boreal winter as compared with summer.

Figure 4 shows that the assessment of a downscaling method’s performance is dependent upon the metric

![Graph](image-url)
used. While the deterministic method performed better with the $R^2$ metric, the probabilistic method possessed a more accurate representation of the variance and extreme values. We note that each metric provides different information about how well the particular dataset performs, and hence the choice of verification metric should be attuned to the specific application in which the data are used. Consequently, numerous metrics will provide both a more complete evaluation of a downscaled methodology and more specific information about characteristics of the downsampling for specific applications and specific users. Including metrics that evaluate characteristics of the dataset that are important to its users will provide key information to aid these users in the choice of a downscaled dataset. The following two sections present verification metrics to evaluate many aspects of a probabilistic downscaled dataset, with a focus on extreme events metrics.

b. Tests on the distribution

Tests on the distribution require the comparison of the downscaled daily varying PDFs with a single observed value for each day. Metrics for the evaluation of probability distributions or probabilistic predictions include likelihood, the Brier skill score (BSS), the cumulative ranked probability score, reliability diagrams, and other skill metrics from forecast verification. The goal of probabilistic downscaling of climate variables on a daily time scale is to represent a realistic range of possible local-scale conditions and as such, the PDFs represent additional local-scale variations beyond the direct influence of the large-scale environment. Some metrics for the evaluation of probabilistic forecasts are applicable in this situation and others are less useful. The tests presented in this section compare the daily varying PDFs with the observed values for the whole distribution (e.g., likelihood, probability integral transforms) or for the occurrence of specific events (e.g., BSS). Additional tests on the distribution can be found in Kirchmeier-Young (2015).

Tests on the distribution rely on comparison with a reference prediction or forecast; here, we use a monthly climatology that is calculated as follows. For daily Tmax and Tmin each month’s climatology was determined by fitting a normal distribution to the 60 yr of daily data for that calendar month. For precipitation occurrence, the proportion of days with nonzero rainfall each month was calculated and the precipitation amount was then assessed via fitting a gamma distribution to the amount of rainfall on wet days for each month. The climatology PDFs are assigned to each day in the month, thus producing a time series of PDFs (but not daily varying). Through the use of a climatology reference distribution, the following tests can assess the added value (over climatology) of using information from the large scale via the featured downscaling methodology.

1) LIKELIHOOD

One method to evaluate the fit of a PDF is to calculate the likelihood, which assesses a probability distribution’s parameters for a set of observed values. For a PDF $f$ that describes the distribution of a variable $y$ with parameters $a$ and $b$, the likelihood ($L$) can be calculated across all days ($i$) as follows:

$$L(a, b | y) = \prod f(y_i | a, b)$$

$$\ell = \log(L) = \sum \log[f(y_i | a, b)].$$  (1)

It is common to work with the log likelihood [$\ell$, above or $\log(L)$] to simplify the products into sums. Maximum likelihood methods for fitting a distribution seek to identify the set of parameter values that maximize the likelihood function given a set of observations, thus providing the best fit. Maximizing a likelihood is synonymous with minimizing the negative (log) likelihood.

The likelihood can be used to assess a set of distribution parameters (from the downscaled PDFs), given a time series of observations. The negative log likelihood [$-\log(L)$] of the observed values given each day’s downscaled PDF was calculated for each station. Alone, likelihood values have little meaning, so the $-\log(L)$ given a monthly climatology was also determined. The ratio of these is plotted in Fig. 5 for Tmax (Fig. 5a), Tmin (Fig. 5b), and precipitation occurrence (Fig. 5c) and amount (Fig. 5d). A ratio of less than 1 implies the downscaling has skill over climatology with this metric. The downscaling routine does in fact show skill for almost all stations and months and indicates that the use of large-scale information via the downscaling method better represents what was observed than a simple climatology would.

There is a seasonality to this ratio, with lower skill during the summer months than during the rest of the year. Possible explanations include the local scale exhibiting less of a connection to the large scale during the summer, more skill in climatology during the summer, or less skill in the determination of the large-scale predictors in the reanalysis for this season. The seasonality is most pronounced for precipitation amount, which exhibits very different characteristics during the warm months than during the cool months for most stations. This seasonality is likely because summertime midlatitude precipitation is dominated by convection, a process that is parameterized (not resolved) by the large-scale models and reanalyses, while winter precipitation is more controlled by the large-scale synoptic environment, which is better represented by the reanalyses.
2) PROBABILITY OF EVENT OCCURRENCE

An advantage of a probabilistic downscaling methodology is that each day has a quantifiable probability of experiencing a particular event, which can be evaluated using a test on the distribution. The Brier score (BS; Brier 1950) is a metric to assess the performance of a probabilistic forecast for the occurrence of a particular event, here an extreme event day as determined by the exceedance of a threshold of the climatological 95th percentile value;

$$\text{BS} = \frac{1}{n} \sum (f_i - E_i)^2. \quad (2)$$

Here, $f_i$ is the forecast probability of the event occurring on day $i$ (calculated from the downscaled PDFs) and $E_i$ is a binary indication of whether or not the event occurred (i.e., $E_i = 1$ if the observed value exceeded the extreme threshold on day $i$ and 0 otherwise). The BS is a summary measure across $n$ days. For easier interpretation, a skill score is calculated by comparing the BS values between the forecast of interest and a reference forecast (here, climatology). As such,

$$\text{BSS} = 1 - \frac{\text{BS}_{\text{ds}}}{\text{BS}_{\text{clim}}}, \quad (3)$$

and, if multiplied by 100, the BSS can be interpreted as the percent improvement of the downscaled prediction over the climatology prediction (Wilks 2006).

Box plots representing the spread of BSS values across all stations for each month are shown in Fig. 5. Almost all sites show robust improvement over climatology. This improvement is weaker during the summer months than during the rest of the year and possible explanations for this seasonality are presented in the prior section. Many of the stations where the downscaling performs worse than climatology for Tmax (those with negative BSS) are found in the states along the Gulf Coast.

For precipitation, extreme events are defined by a threshold from the climatological 95th percentile value for the amount of precipitation on wet days only. Similar to the temperature variables, many months show considerable improvement over a climatological forecast for precipitation, though improvement is again weakest during the summer months. The pronounced seasonality is likely due to the very different characteristics precipitation exhibits during the warm months than during the cool months for most stations. In general, the largest improvements are seen at stations in the northern plains and the Northeast (not shown).
Probabilistic downscaling enables the analysis of extreme precipitation without relying on the actual observation of extreme events by utilizing the probability of event occurrence as discussed above. Extreme precipitation, especially during the summer, can be very localized. A point station will only observe an extreme rainfall event if the storm passes directly over the rain gauge and it is likely that greater rainfall amounts fell somewhere nearby. The downscaled PDFs have the ability to express when extreme precipitation events are likely to occur, given the appropriate large-scale conditions. Smaller improvements over climatology might be seen for precipitation (relative to the temperature variables) because days with a larger probability of an extreme event may often describe situations where the station did not observe an extreme rainfall event but a nearby (unobserved) location did.

3) PROBABILITY INTEGRAL TRANSFORM

A PDF forecast can also be considered as an ensemble forecast with an infinite number of members. One metric to evaluate an ensemble forecast is the probability integral transform (PIT; Gneiting and Katzfuss 2014), which expands upon the rank histogram for finite ensembles. A rank histogram plots the frequency of ranks of the observed value within a prediction ensemble; typically, the best performance is represented by a uniform distribution, where the observations occur at each point in the distribution of ensemble members with equal frequency. A PIT calculates the value of the prediction cumulative distribution function that corresponds to the observed value and plots the frequency with which the observations are associated with each quantile. In general, as with the rank histograms, the ideal form is a uniform distribution of frequencies; concave (convex) histograms allude to an under (over) representation of the variance, while sloping histograms indicate a bias. Hamill (2001) provides a further description of rank histograms and advice for their interpretation, as the simple explanation just given can lead to incorrect conclusions in some situations.

PITs were determined for Tmax and Tmin for a winter month (January) and a summer month (July) and are plotted in Fig. 7. Each quantile bin (colors) contains a bar for each station. The slight upward slope of the Tmin PITs, particularly during summer, indicates a small cold bias of the conditional distribution, since more observations are falling in the upper percentiles. The convex shape of the PITs for Tmax and winter Tmin implies an overestimation of the conditional variance, with observations more often found in the middle percentiles. It should be noted that this metric is testing the conditional bias and variance of the local scale given the large-scale conditions, not the bias or variance of a time series realization (which includes both the large- and local-scale contributions), and thus the results here are not directly comparable with those discussed in section 3a or later in section 3c. The results from this analysis suggest future downscaling efforts should address this potential overestimation of the variance of the local scale given a particular set of large-scale conditions. For both Tmax and Tmin, the prominence of the first bin without a negatively sloping structure through the remaining bins implies an underestimation of the lower tail of the conditional distribution. The identification of this feature is possible through the use of PITs and could be missed with other metrics. Although the dominant shapes discussed above summarize the results across the domain, many individual stations present different patterns.

c. Tests on the realizations

Although some climate statistics, such as the mean and standard deviation, can be analyzed with parametric
parametric tests are less useful for many other important climate statistics (e.g., duration of dry spells or the date of the annual temperature maximum). Instead, we express our tests on the realizations in terms of a standardized anomaly (SA). This standardized anomaly was designed for ease of interpretation, as the implications of simpler metrics such as the bias can be conditional upon the location, season, or magnitude of the variable of interest. For each station, a climate statistic $u$ (mean, 99th percentile value, number of 5-day heat waves, etc.) is determined for the observed time series. Next, a value of $u$ is similarly calculated for each of 1000 realizations of the downscaled PDFs, thus producing 1000 estimates of this climate statistic. To compare the observed $u$ with the distribution of $u$s from the realizations, the SA is calculated as follows:

$$SA = \frac{u_{obs} - \bar{u}_R}{s_{u,R}},$$

(4)

by subtracting the mean $u$ across the realizations ($\bar{u}_R$) from the observed $u_{obs}$ and then dividing by the standard deviation of $u$ across the realizations ($s_{u,R}$).

The spread of SA across all of the stations (Fig. 1) for each statistic $u$ is expressed in box plots (see Fig. 8 for an example). The SA values of smaller magnitude imply the observed statistic fit well into the distribution of $u$ from the realizations. Dotted lines on each plot at $+1.96$ and $-1.96$ estimate a 95% confidence interval (CI) for the distribution of $u$ from the realizations; these limits would be the CI if the set of $u$ from the realizations was normally distributed (note that these lines are added simply for reference as there is no a priori reason to expect a particular statistic to be normally distributed). This metric allows for the performance of our downscaling to be compared between stations, between variables, and between statistics. The SA is defined so that larger positive values indicate an overestimation by the downscaling routine, while more negative values imply a low bias. It should be noted that the SA is standardized by the standard deviation across the realizations. As such, for a statistic such as the mean, which is very precisely estimated from the realizations ($s_{u,R} < 0.1$), large SA values may not be indicative of biases of concerning magnitude. The median value (across stations) of this standard deviation for each of the following analyses is included in Table S2 of the online supplemental material to aid the reader in the interpretation of the bias magnitude.
Figure 8 displays the spread of SA values for the mean and standard deviation of Tmax and Tmin by month. At most stations, Tmax is well represented by the downscaling, as evidenced by the boxes lying mostly within the confidence interval (dotted lines). There is variability between stations, with some showing large positive outliers (mean Tmax overestimated) and others showing large negative outliers (mean Tmax underestimated). There is no robust seasonality to the SA across all stations. The standard deviation of Tmax is often underestimated by the downscaling, particularly during the summer, as the boxes show the largest displacement to the negative side of the confidence interval during these months. The boxes are often wider than the confidence interval, implying large variability across stations. The mean for Tmin is also underestimated and, in general, this underestimation is stronger at stations in the Southeast (not shown). The standard deviation of Tmin is better represented, with only a slight low bias.

The downscaling model involves a linear prediction of the mean value that should be inherently unbiased. A repeat of the above analysis using distribution parameters before interpolation to the grid (directly downscaled to the station) eliminates the bias in the means of Tmax and Tmin (not shown). Thus, the process of interpolating the distribution parameters to a grid (with the influence of many other neighboring stations) results in the slight biases in the mean seen in Fig. 8. Although not eliminated, the biases in the standard deviation for both Tmax and Tmin are reduced with the pre-interpolation parameters. Furthermore, the length of the boxes is smaller pre-interpolation, implying less variation across stations. Although slight biases are introduced when comparing directly with these stations, it is possible the inclusion of the influence of additional nearby stations produces a set of PDFs that are more representative of a region than a point estimate.

A collection of metrics to describe extremes in temperature and precipitation were identified. This section describes each metric, its motivation and uses, and the corresponding SA values to assess how it is represented by the featured probabilistic downscaling routine.

1) PERCENTILES

A popular metric for characterizing the tails of a distribution is the calculation of percentile values. The $p$th percentile value represents the threshold that $(1 - p)\%$ of the data exceeds. For example, 5% of a time series will exceed the value defined by the 95th percentile. We utilize a common notation for percentile values by placing a small $p$ in front of the corresponding number (e.g., $p95$ for the 95th percentile). Percentile values are typically calculated empirically and thus do not rely...
upon any assumptions of a parametric distribution for the data, requiring only a time series of sufficient length. Even if the data are divided into individual months, the 60 yr of daily data provided in our downscaled dataset are more than sufficient for most metrics.

The spread of SA values for percentiles in the tails of the annual distribution of Tmax and Tmin is shown in Fig. 9. The downscaling routine tends to slightly overestimate the lower-tail percentiles of Tmax and underestimate the upper-tail percentiles, which is consistent with an underestimation of the standard deviation that was shown in Fig. 8. There is slight regional clustering (not shown), but regional patterns vary between percentiles. The boxes for both Tmax and Tmin are often much taller than the CI, which indicates large variability across stations. Tmin shows a low bias for all but the lowest percentiles. This finding is consistent with a general underestimation of Tmin and its standard deviation as seen from Fig. 8. For many percentiles, the strongest underestimations occur in the same regions where the mean value of Tmin is also underestimated.

The percentiles for a precipitation distribution are calculated for the precipitation amounts on wet days. Because of random sampling, it is likely that each realization will have a slightly different sequence of wet–dry days. We note that if all days in the time series were included, the resulting percentile values could be heavily influenced by the proportion of wet days. Since the proportion metric was tested separately (not shown), only wet days are used to assess the distribution of precipitation amount. Furthermore, as the distribution of precipitation amount is bounded by zero on the lower tail, only the upper-tail percentile values are important to the assessment of precipitation extremes. The analysis of dry-day precipitation extremes (e.g., drought) requires a different metric.

Box plots for the spread of SA across all stations for percentiles from the annual distribution of precipitation amount are shown in Fig. 9c. There is good agreement between the downscaling and the observations for the upper-tail percentiles, especially for p95 and p99. While the bias for p80 is small, the spread is greater than that for the other percentiles displayed. Most of the stations with underestimated values of p80 are found in the Southeast and Atlantic coast regions, while the stations seeing a stronger overestimation are generally located in the northern plains (not shown).

2) SINGLE-DAY EVENTS

Individual days of extreme temperatures are important to public health and agriculture. For example, extreme heat days can be detrimental to the health of high-risk populations, such as the elderly, or to those with outdoor occupations and can also cause respiratory stresses through increased pollution (Patz et al. 2014). While a common metric is to count the number of days exceeding a specific threshold, meaningful thresholds can vary greatly by region or application. As an alternate method, the annual maximum and minimum values of Tmax and Tmin were averaged across the 60 yr in the time series. Furthermore, the Julian day (or day of year number) of each value was included to evaluate how well the probabilistic downscaling represents the seasonality of extreme temperatures. For the annual

Fig. 9. Spread of SA across all stations for percentile values of (a) Tmax, (b) Tmin, and (c) precipitation amount. Dotted lines at +1.96 and −1.96 estimate the 95% CI from the realizations.
minimums, the Julian day 1 was shifted to 1 July to ensure that the discontinuity between day 1 and day 365 occurs in the summer when minimums are unlikely to occur. The results are presented in Fig. 10.

For Tmax, the mean annual maximum is slightly overestimated in the downscaled realizations, but the date at which this value occurs is only slightly too late in the year. In contrast, the downscaled realizations agree better with the observations regarding the mean annual minimum and the timing of these events. For Tmin, the downscaled realizations show good agreement in the values of the annual maxima and minima as well as the timing of these events. There is a large spread for the value of the mean annual minimum temperature, with southern stations seeing smaller than observed values and many of the more northern stations seeing larger than observed values (not shown).

Large rainfall events from a single storm (here assessed as single-day events) are important to users in storm water management, agriculture, public health, lake management, and coastal communities. To assess the frequency of single-day extreme events, the number of events in the seasonal time series of different magnitudes was determined. Events were placed in a single bin, determined by the largest threshold exceeded. For example, a day recording 3.5 in. (1 in. = 2.54 cm) of rain was placed only in the 3-in. bin and not also in the 1- and 2-in. bins. The spread of SA values for the number of events in each bin are shown in Fig. 11.

There is good agreement for all seasons and for all precipitation bins. The spread is typically largest for the smaller bins where the number of events is greatest. The 5- and 6-in. bins see good agreement but also very few events, with many sites not observing any events of these magnitudes. Furthermore, the larger rainfall events are more common during the summer months and during spring and/or fall in some regions. It is important to note that summertime precipitation is dominated by convection, a process that is parameterized (not resolved) by the reanalysis.

3) DURATION EVENTS

Statistics describing the duration of heat-wave events are important to many fields, including energy, agriculture, and public health. There is no standard definition of a heat wave, as it can vary by application (e.g., Smith et al. 2013). Here, a heat-wave day is defined as one with a Tmax greater than the observed climatological 90th percentile value for that site. Throughout the 60-yr record, the number of distinct events of various numbers of consecutive heat-wave days were determined. Similar results were seen with a threshold defined by p95 instead.

After repeating the same analysis for 1000 downscaled realizations, a 95% nonparametric confidence interval was constructed for the number of events. An example of this CI for Madison (MSN) in comparison with the observed event counts is shown in Fig. 12a. Although the observed and downscaled number of events are often similar at other locations, the threshold that determines an event day (p90) varies greatly by region. The downscaled predictions slightly overestimate the events of shorter durations (3 and 4 days), but the observed number of events for longer durations falls well within the CI from the realizations. The spread of the standardized anomalies across all sites is presented in the right panel of Fig. 12a. Overall, the downscaling accurately represents the number of events of each duration for most sites. The exception is an overestimation of the number of three-day events; stations in the northern plains see the best agreement for the three-day events (not shown), while most of the domain experiences an overestimation.

Extreme precipitation events can also be defined in terms of durations. Extended periods of wet days or extended periods of dry days are especially important to those in water management, city planning, and agriculture. While continuous days of rainfall often lead to flooding that results in damages to infrastructure, homes, and crops, it can also cause contamination of the surface or groundwater, which facilitates waterborne diseases (Curriero et al. 2001). Additionally, if a single-day heavy rainfall event occurs after periods of drought, the water tends to run off before being absorbed into the soil, which also results in flooding and contamination (Rosenzweig et al. 2001). The temporal autocorrelation
of precipitation events is generally low, and as a result, no additional autocorrelation is included beyond that contributed to the distribution parameters from the large-scale predictors. However, spatial autocorrelation is important for precipitation events. Although it is possible to consider this relationship when extracting realizations, we do not include the spatial autocorrelation when drawing precipitation realizations for this analysis since each station’s statistics are assessed separately.

Similar to the temperature analysis (Fig. 12a), the number of dry- and wet-spell events of various durations was determined for both the observations and for each of 1000 downscaled realizations. For the observations, a wet day was defined as any day with a precipitation amount greater than 0 mm. Wet or dry days for the realizations were determined based on each day’s probability of precipitation. Plots of the observed number of events and the 95% confidence intervals from the realizations are shown in Figs. 12b and 12c for the MSN example. In general, there is strong agreement between the downscaled data and the observations for most wet and dry durations, with the exception of an underestimation of the number of shorter-duration wet-spell events. The spread of SA values for all stations is shown in the right panel of Figs. 12b and 12c. There is very good agreement for the dry-spell durations, with only a few stations exhibiting SA values outside the estimated confidence interval. The outlying stations locations vary between durations. For the wet-spell durations, there is good agreement for most durations. However, the spread is much greater than for dry days, especially for the shorter-duration wet spells. For the 3- and 4-day events, most of the underestimated stations are found in the northern plains, while the fewer overestimated stations are located in the Southeast (not shown).

4) EXTREME VALUE DISTRIBUTION

The GEV distribution is the limit distribution of block maxima and is (along with its counterpart, the generalized Pareto distribution) a common way to calculate return levels for extreme events (e.g., designing a storm water system that can handle up to the 100-yr event). It is categorized into three types based on the sign of the shape parameter. The type I, or Gumbel, sees shape parameters of 0, while positive shape parameters indicate a type II, or Fréchet, and negative shape parameters indicate a type III, or Weibull. Type II distributions are characterized by a heavier upper tail and a bounded lower tail, while type III distributions exhibit a bounded upper tail. In contrast, the GEV location parameter simply describes a shift of the distribution in magnitude.

A Monte Carlo test with the GEV distribution was used to assess the representation of the extremes in the downscaled data. For each month separately, yearly maxima from the observations were fit to the GEV distribution, noting the values of each of the three distribution parameters (shape, scale, and location). The process was repeated for each of 1000 realizations, resulting in 1000 estimates of each GEV parameter for each station and month. The spread of SA values for each parameter is shown in Fig. 13. The GEV parameters were fit using L-moments (Hosking et al. 1985) through a software package from Brodtkorb et al. (2000). The goodness-of-fit of the GEV distribution for

![Fig. 11. Spread of SA across all stations for the number of daily precipitation events of various magnitudes (1, 2, 3, 4, 5, and 6 in.) by season for (a) December–February, (b) March–May, (c) June–August, and (d) September–November. Dotted lines at +1.96 and −1.96 estimate the 95% CI from the realizations.](image-url)
FIG. 12. (left) Example for MSN with the observed number of events for each duration in black and the 95% CI across 1000 realizations in gray, and (right) the spread of SA across all stations for the duration of (a) heat-wave events, (b) dry spells, and (c) wet spells.

(a) heat wave

(b) dry spell

(c) wet spell

Unauthenticated | Downloaded 12/30/22 12:31 AM UTC
the observations was assessed visually, and nearly all (>98%) station/month combinations showed very good agreement between an empirical distribution from the observations and a theoretical distribution using the fit parameters.

An advantage of using a probability distribution to describe the extreme events is that characteristics can be analyzed without relying upon a large enough number of such events being observed. This is especially important for precipitation, as extreme precipitation events are more localized. Figure 13 displays the SA results for the GEV fit for precipitation amount, which denotes very good agreement between the observations and the downscaled data for all parameters. An agreement on the shape parameter indicates that both the downscaled and observed data show similar tail behavior in the distribution of their precipitation extremes. Since the downscaling routine accurately represents each of the GEV distribution parameters, the subsequent calculation of return levels should also agree well. The GEV parameters for Tmax and Tmin are also fairly well represented (not shown), with the exception of an overestimation of the location parameter for Tmax during summer and fall.

d. Comparison with a deterministic method

In this section, the performance of a realization from the UWPD downscaled PDFs is compared with a widely used deterministic dataset. We use the bias correction constructed analog (BCCA; Maurer et al. 2010) data, chosen for its daily output and widespread availability (an important contribution to national efforts at climate impact assessment; Melillo et al. 2014), even if it shows less skill than other deterministic downscaling methods (Gutmann et al. 2014). The BCCA method is used for gridded data on a large spatial domain and employs the quantile-mapping bias correction method of Maurer and Hidalgo (2008) before the constructed analog methodology is applied. BCCA data are not available for historical observations based on reanalysis and UWPD data are not yet available for the CMIP5 models. As such, a common set of CMIP3 models was used as the large-scale forcing for both BCCA and UWPD. As the first step in BCCA (and in UWPD) is a bias correction, the evaluation results show only small differences between models. As such, downscaled data from a single model (GFDL CM2.0; Delworth et al. 2006; Gnanadesikan et al. 2006) will be discussed in this section.

BCCA provides a single deterministic dataset, and this will be compared with one realization from the UWPD. Only 154 stations are used for the verification analysis, since the BCCA grid does not have data for two coastal stations. CMIP3 provides daily data from the period of 1961–2000, and this date range was used for both BCCA and UWPD. Because of the shorter time period and the different large-scale dataset, the results for UWPD will not necessarily be the same as those presented previously, even with debiasing.

Figure 14 displays the difference between the value of a precipitation statistic from the downscaled data and that from observations. Both UWPD (gray) and BCCA (purple) agree well with the total precipitation. However, the BCCA dataset greatly overestimates the proportion of wet days. For the results presented in Fig. 14, a threshold of 0 mm was used to determine wet/dry days. Raising this threshold (for just the BCCA) decreases the discrepancy with the observed proportion of wet days, particularly during the cool season months. Even with a threshold of 1 mm, the summertime
The proportion of wet days is still overestimated by the BCCA, by about 0.1; consequently, the wintertime proportion of wet days is now underestimated. As the choice of a wet-day threshold is arbitrary and would need to vary by season, the threshold is left at 0 mm. The deterministic BCCA dataset underestimates the upper-tail percentiles of the precipitation amount distribution (Fig. 14c) and shows increased spread of the metric spatially, relative to the UWPD dataset, which shows no overall bias in the percentiles of precipitation amount. For seasonal events exceeding various precipitation thresholds (Fig. 14d), the BCCA dataset often shows an underestimation of the number of events. The small bias in 3-in. events, for both BCCA and UPWD, is partly due to the rarity of these events, particularly in cooler seasons.

4. Discussion and conclusions
The intent of this study was to synthesize a set of metrics with which to evaluate a probabilistic downscaled dataset, to apply those metrics to a range of extreme event statistics, and to demonstrate the utility of a probabilistic downscaling approach for the representation of extreme events. The novelty of this probabilistic dataset is that it retains both the connection between the large-scale environmental conditions and local-scale variation, as well as a quantification of the local-scale variations that are not explained by the large-scale environment. While this characteristic of probabilistic downscaling allows considerable flexibility, it also introduces challenges in choosing verification techniques. This paper aims to demonstrate that probabilistic downscaling has numerous advantages, including the flexibility of its datasets and the ability to represent extremes, while highlighting an array of verification metrics that can be applied to a probabilistic dataset, particularly for nonstandard variables.

Here, we present a set of metrics and a methodology for evaluating the representation of extreme events in a probabilistic downscaled dataset. A set of climate statistics was defined, encompassing many characteristics...
of extreme events, particularly those important to the users of the downscaled data. These include probability of occurrence, percentile values, measures of single-day events, measures of duration events, and extreme value distribution parameters. Extreme events were evaluated using 1) the PDFs directly (as in section 3b) or 2) realizations of the daily varying PDFs (as in section 3c). Tests on the distribution were inspired by verification metrics from probabilistic forecasting and include a ratio of likelihoods and probability integral transforms to analyze the entire distribution and the BSS to assess the probability of occurrence of extreme events.

For the tests on the realizations, a standardized anomaly (SA) methodology was developed for a standardized comparison between the observations and the downscaled data. A benefit of the SA methodology is the ability to compare results across statistics, across stations, and across time periods. The SA plots presented provide information about how well the observed value of each metric compares to the distribution of values of that metric from a set of realizations of the downscaled PDFs. Combined with the probabilistic downscaling, the SA methodology is general enough to be used for any extremes metric. Illustrative examples were presented, but the tests discussed could easily be adapted for specific user needs. For example, the duration of heat wave analysis could be repeated with a different definition of a heat-wave day or the thresholds for single-day events could be adjusted for a specific application. Overall, the probabilistic downscaling performs well for a wide variety of extreme event statistics.

Many of the evaluation metrics discussed (e.g., likelihood, BSS, and PITs) were taken from methodologies for verification of probabilistic forecasts. While additional evaluation metrics are discussed in Kirchmeier-Young (2015), other forecast verification metrics were not included because of their impracticality (e.g., diagram-based metrics that, while potentially useful, would be impractical to apply across the large number of stations used in this study) or their limited application to probabilistic climate downscaling. Still, given the wide variety of users of downscaled climate data, we emphasize that any (forecast) verification method may be useful for a specific application.

Extreme events are important to many users of downscaled climate data. An understanding of how well a particular downscaled dataset represents these extreme events will aid users in determining how best to employ the downscaled data for their own purposes. As the users of downscaled data represent a wide range of interests and a broad set of needs, numerous metrics are needed for evaluation. It was shown that conclusions about the “skill” of a downscaling methodology can be conditional on the evaluation metric used. As such, we emphasize the importance of not only using many metrics for verification, but also choosing metrics that assess characteristics important to the users of the dataset. This study demonstrates the utility of a probabilistic downscaling methodology for representing a wide range of extreme event statistics; such a probabilistic approach is not yet commonplace for climate impacts assessments.

Acknowledgments. We thank three anonymous reviewers and the editor, Andrew Ellis, whose comments lead to a substantially improved paper. This work was accomplished through support from the Nelson Institute Center for Climatic Research Climate, People, and the Environment Program (CPEP) at the University of Wisconsin–Madison. For the development of the downscaled dataset, David Lorenz was funded by the Upper Midwest and Great Lakes Landscape Conservation Cooperative (LCC; Grant YMA00000512) and the U.S. Environmental Protection Agency’s Great Lakes Restoration Initiative through a contract with the Michigan Department of Natural Resources (2015-0201).

REFERENCES


