Differential Credibility Assessment for Statistical Downscaling

S. C. PRYOR

Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York

J. T. SCHOOF

School of Earth Systems and Sustainability, Southern Illinois University, Carbondale, Illinois

(Manuscript received 6 December 2019, in final form 11 May 2020)

ABSTRACT

Climate science is increasingly using (i) ensembles of climate projections from multiple models derived using different assumptions and/or scenarios and (ii) process-oriented diagnostics of model fidelity. Efforts to assign differential credibility to projections and/or models are also rapidly advancing. A framework to quantify and depict the credibility of statistically downscaled model output is presented and demonstrated. The approach employs transfer functions in the form of robust and resilient generalized linear models applied to downscale daily minimum and maximum temperature anomalies at 10 locations using predictors drawn from ERA-Interim reanalysis and two global climate models (GCM; GFDL-ESM2M and MPI-ESM-LR). The downscaled time series are used to derive several impact-relevant Climate Extreme (CLIMDEX) temperature indices that are assigned credibility based on 1) the reproduction of relevant large-scale predictors by the GCMs (i.e., fraction of regression beta weights derived from predictors that are well reproduced) and 2) the degree of variance in the observations reproduced in the downscaled series following application of a new variance inflation technique. Credibility of the downscaled predictands varies across locations and between the two GCM and is generally higher for minimum temperature than for maximum temperature. The differential credibility assessment framework demonstrated here is easy to use and flexible. It can be applied as is to inform decision-makers about projection confidence and/or can be extended to include other components of the transfer functions, and/or used to weight members of a statistically downscaled ensemble.

1. Introduction

Downscaling (or right scaling) of climate information is conducted primarily to advance understanding of processes affecting climate that are not resolved well by global models and/or provide decision-makers responsible for adaptation actions with salient and credible information about possible climate change within their specific region/locality of interest (Giorgi 2019; Gutowski et al. 2020; Maraun and Widmann 2018). Empirical-statistical downscaling (ESD) methods are widely applied to develop climate change scenarios from coarsely resolved climate model projections [see synthetic reviews and comparisons in Benestad et al. (2008), Gutiérrez et al. (2019), Hertig et al. (2019), Hewitson et al. (2014), and Maraun and Widmann (2018)].

Two fundamental frameworks are commonly used within ESD: perfect prognosis (PP) and model output

statistics (MOS). In PP approaches transfer functions linking large-scale predictors and near-surface predictands are developed based on historical observations or reanalysis products. These are then applied to output from climate models [coupled atmosphere–ocean general circulation models or Earth system models (ESMs), referred to herein as GCMs]. MOS, on the other hand, relies on development of transfer functions between predictors output from a GCM and historical surface observations. It thus implicitly assumes time synchronization of the predictands (conditions at the ground) and upper-level conditions (predictors) that cannot be achieved with free running coupled models. This assumption is not realized when predictors are drawn from a GCM because they generate a plausible sequence of conditions, rather than date specific realizations. Thus, MOS-based ESD transfer functions derived using GCM output in a timewise sense have been described as being essentially equivalent to bias correction of the mean (Maraun and Widmann 2018).

Corresponding author: S. C. Pryor, sp2279@cornell.edu

DOI: 10.1175/JAMC-D-19-0296.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](https://www.ametsoc.org/PUBSReuseLicenses) (www.ametsoc.org/PUBSReuseLicenses).

PP approaches are also dependent on a range of assumptions (Maraun and Widmann 2018). These include that the predictors (i) are realistically simulated and free from nonstationary (time varying) biases, (ii) are responsible for a large fraction of the variability in the predictands, and (iii) appropriately capture external forcings responsible for climate change. Crucially, use of PP is predicated on the assumption that the GCM reproduces the characteristics of the predictors as manifest in the historical observations or reanalysis used to generate the transfer functions.

An additional challenge to ESD is that any set of predictors account for only part of the variance in the predictand (Maraun and Widmann 2018), resulting in underdispersion in the downscaled series relative to observations. Methods for artificially increasing the variability include randomization (von Storch 1999) and variance inflation (Karl et al. 1990). However, variance inflation increases the mean square error between the observations and adjusted predictands (von Storch 1999), while randomization changes the temporal autocorrelation of the downscaled predictand (Huth et al. 2001).

Robust analyses of climate change projections of impact-relevant near-surface variables requires understanding of model fidelity and causes of spread in an ensemble of projections derived using different model chains, formulations, and assumptions (Chegwidden et al. 2019; Fatichi et al. 2016; Maraun et al. 2015). Providing robust guidance regarding fidelity (or credibility) of a given projection is challenging because skill has often been defined purely on statistical measures of agreement with independent observations in the contemporary climate. “Perfect model” experimental designs address part of this limitation and test the key assumption that ESD performance in the contemporary period is representative of skill in the future climate (Dixon et al. 2016).

The concept of differential credibility assessment (DCA) is predicated on seeking to understand the degree to which the processes that lead to an “outcome” (e.g., a given daily minimum or maximum temperature at a given location) are correctly reproduced by a given model. This represents an improvement over purely statistical assessment of the degree of correspondence between model output and some observed quantity in independent data collected during the contemporary climate in that the credibility is assessed by evaluating if the model correctly represents the drivers of the parameter of interest. Implementation of DCA may address the concern that model performance as measured using statistical skill during a selected period with the historical and contemporary is no guarantee of skill in another time period (Reifen and Toumi 2009). The idea is that if DCA can be used to evaluate and demonstrate

“getting the right answer for the right reason,” the model under question may be more likely to generate meaningful climate projections. DCA is symbiotic with, and leverages, the emergence of process-oriented diagnostics (PODs) (Maloney et al. 2019), and feature-based methods of evaluating model fidelity for extreme events (Sillmann et al. 2017).

Differential credibility is designed to inform users of climate services regarding the confidence with which a specific climate change product can or should be incorporated into knowledge–action systems (Weichselgartner and Arheimer 2019). DCA also offers important information to accelerate model development and improvement (Maloney et al. 2019).

Differential credibility has been previously advanced in the context of dynamical downscaling where higher credibility is assigned to output from model chains that reproduce key climate features and their physical causes [e.g., the southwestern monsoon (Bukovsky et al. 2013), and atmosphere–land coupling responsible for precipitation in the southern Great Plains (Bukovsky et al. 2017)]. Models are subject to PODs to characterize the differential credibility of a suite of model realizations within the contemporary climate. The degree to which the mechanistic processes by which a warming climate may cause changes in the property of interest are well resolved are also assessed (Bukovsky et al. 2013; Bukovsky et al. 2017). Differential credibility can also be assigned to climate projections from different models via expert judgement of model fidelity (Mearns et al. 2017). It has been proposed that differential credibility in the contemporary climate could inform selection of models for inclusion in multimodel ensembles used for climate change projections (Krysanova et al. 2018) or be used to apply differential weightings to model projections within an ensemble (Lorenz et al. 2018).

Uncertainty in climate change projections are sometimes characterized differently by researchers and decision-makers and it can be challenging to effectively communicate certainty and uncertainty in order to develop optimized adaptation strategies (Addor et al. 2015; Barsugli et al. 2013). Here we seek to enable that communication by advancing a methodology to assign differential credibility to climate projections from a PP-based ESD method when it is applied to different GCMs and/or at different locations. The credibility assessment is derived based on two key aspects of ESD method: the degree to which the key predictors in the downscaling are well reproduced by the GCM, and the degree to which enhancing variance in the predictands based on the temporal autocorrelation in those predictors recovers observed levels of variability in daily minimum and daily maximum air temperatures (Tmin and Tmax).

The objectives of this research are as follows:

- 1) evaluate whether GCM-derived predictor variables at the gridcell level possess sufficient fidelity relative to reanalysis data to justify their use in a PP downscaling context,
- 2) advance methods to increase the variance of downscaled predictands without excessive distortion of the temporal autocorrelation, and
- 3) advance methods by which the differential credibility of different downscaled projections can be objectively quantified and readily conveyed to end users.

2. Data

To illustrate our method, we apply it to daily T_{min} and T_{max} at 10 locations across the continental United States (Fig. 1). These data are drawn from the Livneh dataset that is gridded at a spatial resolution of $1/16^\circ$ latitude/longitude (Livneh et al. 2013). These locations are referred to herein by the name of the closest city or by the acronym ARM to indicate the Department of Energy Atmospheric Radiation Measurement Climate Research Facility (Fig. 1). In this illustration of the method ERA-Interim reanalysis (Dee et al. 2011) and two different GCMs, GFDL-ESM2M (Dunne et al. 2012) and MPI-ESM-LR (Giorgetta et al. 2013), are used to provide daily values of seven potential predictors from the grid cell containing each Livneh location. These GCM were selected because they exhibit different equilibrium climate sensitivity (i.e., temperature response to a doubling of carbon dioxide concentration). The equilibrium climate sensitivity is approximately 2.4°C for GFDL-ESM2M (referred to herein as GFDL), and 3.6°C for MPI-ESM-LR (referred to herein as MPI) (Andrews et al. 2012). The predictors are as used and justified in prior research (Pryor and Schoof 2019): geopotential height at 500 hPa (Z500), air temperature at 700 hPa (T700), specific humidity at 700 hPa (Q700), air temperature at 500 hPa (T500), specific humidity at 500 hPa (Q500), zonal (west–east or u component) wind speed at 700 hPa (U700), and meridional (south–north or v component) wind speed at 700 hPa (V700).

We begin by showing an application of our downscaling and DCA method for the contemporary climate (1979–2005) wherein the transfer functions are trained using ERA-Interim and Livneh data from 14 “odd” years (1979, 1981, 1983, ...) and evaluated using independent (test) predictors from 13 “even” years (1980, 1982, 1984, ...) from ERA-interim, and also applied to output from GFDL and MPI. Thus, the training dataset comprises 5110 days of data, whereas the independent testing dataset comprises 4752 days. These datasets

FIG. 1. Locations and names of the Livneh grid cells for which downscaling is performed.

represent the full availability of coincident data but are shorter than should be used for climate projection development. Thus, this analysis is presented solely to demonstrate the method.

A second analysis is presented in which all data from ERA-Interim and Livneh for the contemporary climate (i.e., all 27 years) are used to derive the transfer functions. These transfer functions are then applied to make climate change projections using GCM output for the contemporary climate (1979–2005) and a future climate period (2076–99) from simulations conducted using the RCP8.5 scenario. Under this scenario by the end of the twenty-first century, the equivalent carbon dioxide concentration is 1200 ppm and the radiative forcing is 8.5 W m^{-2} above that in 1850 (Riahi et al. 2011). This analysis method was selected to ensure the longest possible and thus most robust data record from the reanalysis (27 years) is used to derive the transfer functions that are applied to the GCM output to make climate predictions. In most cases the transfer functions were only very modestly changed from the forms derived using the 14-year subset of the reanalysis.

The GCM versions and output are as submitted to the CMIP5 archive. MPI output has an approximate horizontal resolution of 1.87° by 1.88° , whereas that for GFDL is 2.5° by 2.0° . For simplicity, the GCM output was regridded to the ERA-Interim grid (T255 spectral resolution of $\sim 80 \text{ km}$) using bilinear interpolation with inverse distance weighting.

3. Method

a. Transfer functions

Prior to development of the transfer functions all predictors and the predictands (T_{min} and T_{max}) are deseasonalized by creating daily anomalies relative to a 7-day running mean conditioned by day of the year. This methodological decision precludes evaluation of the

degree to which the GCM reproduce the seasonal cycle and removes an important source of relatively low-frequency variability in the predictors and predictands. However, the focus of the current work is to develop projections of impact-relevant daily extremes temperature indices and as described previously, use of daily anomaly values (from the climatological mean) (i) inherently bias corrects the mean, (ii) reduces the temporal autocorrelation in the predictors, (iii) reduces predictor collinearity, and (iv) causes the predictors to more closely approximate a Gaussian distribution (Pryor and Schoof 2019). This last consideration is relevant to the current research where the importance of individual predictors in the generalized linear models (GLM) is assessed using beta weights (Wilks 2011). Further, in testing none of the GCM predictors in their raw form reproduce the probability distribution of that parameter from ERA-Interim, while as shown herein, daily anomalies from climatology are more reasonably reproduced.

The GLM transfer functions are developed using measures designed to be resistant to model overfitting. GLM for both predictands at each of the 10 Livneh locations are trained using all seven possible predictors drawn from ERA-Interim and permitting first-order interaction terms. The GLM are derived using (i) stepwise approaches applied using the Bayesian information criteria (BIC) to select the terms that are maintained (Schwarz 1978), and p values for adding and removing a variable of 0.05 and 0.1, respectively. (ii) L1 regularized regression (least absolute shrinkage and selection operator method) (Ng 2004; Soleh et al. 2015) applied to predictors selected by stepwise regression using tenfold cross validation. The final transfer functions link the ERA-Interim-derived predictors (expressed as anomalies from the climatological mean) to the response parameter at each location; T_{max} or T_{min} anomalies from climatology.

It is important to acknowledge that predictors used in statistical downscaling can be derived using a wide array of approaches. To demonstrate the DCA approach, we adopt a simple approach and use predictor values from the grid cell containing the location at which T_{min} and T_{max} are observed. It is also possible to consider averages over larger domains as well as predictors derived from application of empirical orthogonal functions/principal component analysis to atmospheric fields (Pryor and Schoof 2019) or classified into weather types (Gutiérrez et al. 2013).

b. Differential credibility

Use of a PP downscaling framework implicitly requires that predictors drawn from the GCM exhibit

similar properties to those from the reanalysis. Herein we evaluate this by comparing the distributions of each predictor (following removal of the seasonal cycle and including the cross products) from the GCMs with those from ERA-Interim using the two sample Kolmogorov–Smirnov test (Wilks 2011). Since the distributional forms are unknown, standard tables of critical values of Kolmogorov–Smirnov statistic are not applicable (Crutcher 1975). Therefore, a permutation approach is used to evaluate the resulting Kolmogorov–Smirnov statistic using 100 000 iterations. The two sample Kolmogorov–Smirnov test is most sensitive near the median and less sensitive in the tails of the distribution. This reduces the ability to correctly reject the null hypothesis. To avoid false rejection of the null hypothesis that the two samples are drawn from the same distribution, we adopt a very low threshold for acceptable type-1 error. Thus, the null hypothesis is rejected only if the Kolmogorov–Smirnov statistic lies beyond the 0.1%–99.9% confidence intervals.

Beta weights are normalized values of the regression coefficients and can be used to assess the relative importance of different terms in a regression equation (Wilks 2011). They are computed as the regression coefficient multiplied by the ratio of the standard deviation of the predictor to the standard deviation of the predictand. The beta weights indicate the predicted number of standard deviations that a dependent variable will change per standard deviation change in the predictor with all other variables held constant. The higher the beta weight the more important the associated predictor. The total fraction of the beta weights associated with predictors that are well simulated by the GCM is given by the sum of the beta weights from predictors for which the null hypothesis is not rejected (i.e., that have distributions that do not differ significantly from those in the ERA-Interim reanalysis) relative to the sum of the absolute values of beta weights of all predictors. This fraction is used as a process-level diagnostic of downscaling fidelity.

As discussed above, predictand variance from ESD is typically negatively biased because the predictors do not capture all causes of variability. Here we propose a modified randomization approach to enhance variance in the predictand in a manner that is consistent with the variance and temporal autocorrelation of the GLM predictors. This method is designed to minimize the resulting distortion of the temporal autocorrelation of the downscaled predictand but retains computational efficiency. Specifically,

$$\tilde{y}(t) = \hat{y}(t) + \varepsilon(t) \quad (1)$$

is applied to generate a variance adjusted time series of the downscaled predictand $\hat{y}$, developed from the downscaled time series $\hat{y}$, and

$$\varepsilon(t) = [r_1 \times \text{red}(t) + \text{white}(t)] \times \frac{\sigma^2(\text{Obs})}{\sigma^2(\hat{y})}, \quad (2)$$

where r_1 is a weighted mean of the lag-1 autocorrelation of the predictors weighted by their beta weights, red is a red noise time series, and white is a white noise time series, t refers the time increment, and the final term is the ratio of the variance in the observations of the predictand to that in the downscaled time series. The degree to which the variance of independent predictions and observations is recovered by this procedure provides the second component of the credibility analysis.

Recent studies have advanced evidence of an increase in both the spatial and temporal autocorrelation of near-surface temperatures over some regions in historical data (e.g., [Dillon et al. 2016](#); [Koenig and Liebhold 2016](#)) and in output from GCM for the current century (e.g., [Di Cecco and Gouhier 2018](#)). A key benefit of our approach to variance adjustment is that if the temporal autocorrelation of key predictor(s) evolves under climate change, that will be incorporated into the variance adjustment procedure and the temporal autocorrelation of the predictand will also evolve.

Statistical skill is also evaluated in downscaled Tmin and Tmax from ERA-Interim, GFDL and MPI relative to independent observations from the contemporary period (1979–2005). In this analysis the transfer functions derived using data from odd years are applied to independent data from ERA-Interim and output from GFDL and MPI for even years. These same tests are also applied to predictions of the Livneh observations when transfer functions are conducted using all Livneh and ERA-Interim data from 1979 to 2005 and then applied to output from the GCM. The first is a two-sided t test applied to evaluate the null hypothesis that the downscaled Tmin or Tmax anomalies and the observed Tmin or Tmax anomalies come from distributions with the same mean ([Wilks 2011](#)). The second is an F test for equal variances, applied to test the null hypothesis that the downscaled values of Tmin or Tmax at each station and the observed Tmin or Tmax come from distributions with the same variance ([Wilks 2011](#)). Since these tests are repeated for each downscaled predictand at each location for each source of predictors p values are adjusted to account for the false discovery rate. Results of individual tests are significant at a specified significance α if once all of the p values are ranked (from low to high, $j = 1$ to $j = k$) ([Wilks 2011](#)):

$$p_j \leq \frac{j}{k} \alpha. \quad (3)$$

Use of daily anomalies from climatology reduces the degree to which the probability distributions of the predictands are non-Gaussian but does not completely eliminate it according to testing using a permutation test on skewness. Confidence intervals on skewness are derived using 100 000 draws of random numbers of length equal to the effective size of each sample of predictands (i.e., Tmin or Tmax from each location). In this analysis the effective size n' is computed from ([Wilks 2011](#))

$$n' = n \frac{1 - r_1}{1 + r_1}, \quad (4)$$

where n is the total number of values in the sample and r_1 is the lag-1 autocorrelation.

The samples of daily Tmin anomalies from the Livneh observations have a skewness beyond the 0.1–99.9th-percentile values from the random number samples at Birmingham, Alabama; Phoenix, Arizona; Pittsburgh, Pennsylvania; and Yosemite, California. Daily Tmax anomalies have a skewness beyond the 0.1–99.9 confidence intervals in samples from Pittsburgh and Sioux City, Iowa. The F test is very sensitive to deviations from normality ([Markowski and Markowski 1990](#)), so a nonparametric test of equality of variance for two samples is also applied. In this analysis the two samples are combined and ranked and the sum of the squared ranks in the two samples is used as the test statistic ([Conover 1999](#)).

c. Depicting differential credibility

The key objective of this research is to derive a method to characterize the differential credibility of results from GLM-based downscaling. An additional objective is to develop tools to concisely convey the DCA to decision-makers so they can assess relative confidence in projections across (i) space (i.e., between sites), (ii) different climate metrics, and (iii) different GCMs that are being downscaled. Here we depict differential credibility using a simple traffic light system wherein red indicates low credibility in the climate change projections of a given parameter, yellow indicates moderate credibility, and green indicates a relatively high degree of credibility. There are two components to the DCA: (i) the degree to which important predictors in the transfer functions are well reproduced by the GCM relative to ERA-Interim in the contemporary climate and (ii) the degree to which the variance of the downscaled series matches the variance of the observed series when the transfer functions are applied to independent data. Thus, projections of Tmin at a given location are assigned high

credibility (and color coded green-green) if (i) they are derived using a transfer function conditioned using predictors from ERA-Interim and applied to output from a specific GCM if predictors associated with >75% of beta weights in the contemporary climate are well reproduced by the GCM and (ii) the variance of the downscaled predictands (Tmax or Tmin anomalies) is within $\pm 15\%$ of the observed value in the contemporary climate. The thresholds used to allocate high, medium, or low differential credibility based on these two components are shown in Fig. 2 and used here to illustrate the DCA approach. These thresholds are subjective and could be tailored to the specific risk tolerance of the end user.

d. CLIMDEX metrics

For the climate projections, transfer functions derived using the ERA-Interim reanalysis at each of the 10 locations and both predictands are applied to predictors from GFDL and MPI to develop projections for the contemporary climate and for the end of the current century. The downscaled time series of daily Tmin and Tmax anomalies are added to the mean climatologies from the Livneh data for the contemporary climate (1979–2005) and used to compute six Climate Extreme (CLIMDEX) indices for the contemporary and future climate periods:

- 1) number of tropical nights (i.e., $T_{\min} > 20^{\circ}\text{C}$);
- 2) number of frost days (i.e., $T_{\min} < 0^{\circ}\text{C}$);
- 3) cold-spell duration index (CSD), which is computed by first estimating the 10th-percentile (p_{10}) value for that calendar date d over a 5-day window centered on that day of the year (i.e., from $d - 2$ to $d + 2$) (i.e., $T_{\min_{dt}}$) in the contemporary climate:

$$T_{\min_{dt}} = p_{10}(T_{\min_{d-2:d+2}}), \quad (5)$$

then tallying the number of times for which $T_{\min} < T_{\min_{dt}}$ on a given day of the year d and for the following five or more consecutive days, and then summing the number of days per period for which, in intervals of at least six consecutive days, $T_{\min} < T_{\min_{dt}}$;

- 4) number of summer days (i.e., $T_{\max} > 25^{\circ}\text{C}$);
- 5) number of icing days (i.e., $T_{\max} < 0^{\circ}\text{C}$); and
- 6) warm-spell duration index (WSD), which is computed by first estimating the 90th percentile (p_{90}) value for that calendar date d over a 5-day window centered on that day of the year (i.e., from $d - 2$ to $d + 2$) (i.e., $T_{\max_{dt}}$) in the contemporary climate:

$$T_{\max_{dt}} = p_{90}(T_{\max_{d-2:d+2}}), \quad (6)$$

then tallying the number of times for which $T_{\max} > T_{\max_{dt}}$ on a given day of the year d and for the following five or more consecutive days,

Credibility	Color coding	Beta weights	Variance
High	GREEN	> 75%	$\pm 15\%$
Moderate	YELLOW	50-75%	$\pm 15\text{-}25\%$
Low	RED	<50%	>25%

FIG. 2. Summary of the thresholds for high, moderate, and low credibility as used in the DCA along with the color coding used in depicting differential credibility.

and then summing the number of days per period for which, in intervals of at least six consecutive days, $T_{\max} > T_{\max_{dt}}$ (results of this calculation are reported as WSD per year).

These metrics are chosen because they exhibit a high degree of variability in frequency of occurrence across the 10 locations in the contemporary climate (Fig. 3) and also because the accuracy with which they can be reproduced by downscaling relies on the fidelity with which different aspects of the time series of Tmin and Tmax are reproduced. For example, the number of summer days and icing days require the variability and range of Tmax to be well described, while WSD requires both the variability in the magnitudes of Tmax be realistic and the temporal autocorrelation of Tmax. Naturally, CSD and WSD represent a very stringent challenge to any downscaling method.

4. Results

a. Illustration of the approach for the contemporary climate

To illustrate the downscaling approach and DCA, transfer functions linking the downscaling predictors to the two different predictands (Tmin and Tmax anomalies) are derived using output from ERA-Interim and the Livneh data for the odd years during the contemporary climate (1979, 1981, ..., 2005). These GLMs are then applied to predictors from the even years to generate projections that can be evaluated relative to Livneh data from the even years. Transfer functions for Tmin and Tmax at each location exhibit a range of complexity and all exhibit at least one compound predictor (i.e., an interaction term, the cross product of two of the predictors).

There is considerable spatial variability in the ability of GFDL and MPI to reproduce the probability distributions of the important predictors in the transfer functions for Tmin and Tmax. According to the Kolmogorov–Smirnov statistic, output from the two GCMs for the odd years correctly reproduces from 0 to over 90% of the total beta weights from all important predictors selected based on transfer development using ERA-Interim.

FIG. 3. Annual average (mean) frequency of occurrence of (a) tropical nights (i.e., $T_{\min} > 20^{\circ}\text{C}$), (b) frost days (i.e., $T_{\min} < 0^{\circ}\text{C}$), (c) CSD, (d) summer days (i.e., $T_{\max} > 25^{\circ}\text{C}$), (e) icing days (i.e., $T_{\max} < 0^{\circ}\text{C}$), and (f) WSD computed from Livneh observations during 1979–2005.

As one example, the transfer function for $T_{\min}$ anomalies at Fort Logan, Montana, derived using

data from odd years in the contemporary climate has the form

$$T_{\min} = b_1 T_{700} + b_2 Q_{700} + b_3 Q_{500} + b_4 U_{700} + b_5 V_{700} + b_6 (T_{700} \times Q_{700}) + b_7 (T_{700} \times T_{500}) + b_8 (T_{700} \times U_{700}) + b_9 (Q_{700} \times Q_{500}) + b_{10} (Q_{700} \times U_{700}) + b_{11} (T_{500} \times Q_{500}) + b_{12} (T_{500} \times V_{700}) + c, \quad (7)$$

where X_{Y00} are the daily predictor values for the grid cell containing the surface observations ($T_{\min}$) (e.g., T_{700} and Q_{700} denote the daily values of air temperature and specific humidity at 700 hPa). The beta weights derived from the 12 regression coefficients (b_n , where $n = 1, \dots, 12$) are 0.53, 0.28, 0.01, 0.23, -0.09 , -0.057 , -0.14 , 0.08, -0.047 , -0.085 , 0.09 and 0.11, respectively. Thus, the total sum of the absolute beta weights is 1.749, and 78% (1.360 of the 1.749 total) of the beta weights are associated with predictors for which the Kolmogorov–Smirnov test hypothesis could not be rejected. Conversely, 0.389 of the 1.749 total absolute beta weights are associated with predictors for which output from the GFDL model fails the Kolmogorov–Smirnov test. The high fraction of beta weights that are associated with variables that have similar probability distributions in output from GFDL and the reanalysis used to create the transfer functions results leads to a rating of high credibility for this DCA component.

When the transfer functions are applied to independent data from the ERA-Interim reanalysis and/or the

two GCMs for even years during the contemporary climate, the output exhibits substantial underestimation of variance (Fig. 4). This underestimation of variance is decreased by application of an error term comprising a mix of red and white noise (Fig. 4). This approach leads to excess variability at sites with low variance: Seattle, Washington; Yosemite, and Phoenix for $T_{\min}$ and Orlando, Florida, and Seattle for $T_{\max}$. However, it causes variance in the downscaled $T_{\min}$ and $T_{\max}$ anomalies to more closely match that in independent observations at all locations for predictors drawn from all three sources (ERA-Interim, GFDL, and MPI).

The variability enhancement does not substantially distort the autocorrelation in downscaled $T_{\min}$ or $T_{\max}$ anomalies at any of the 10 locations (Fig. 5). Indeed, the autocorrelation for lags ≥ 2 days is generally improved particularly for downscaling of $T_{\max}$ using predictors from the GFDL model (Fig. 5). For example, the time series of downscaled daily $T_{\max}$ anomalies during even years from both MPI and GFDL exhibit excess temporal autocorrelation at lags of >2 days relative at independent

FIG. 4. Variance of (a) Tmin and (b) Tmax anomalies from climatology (1979–2005) in independent observations and downscaled using ERA-Interim (E), GFDL (G), or MPI (M) output for the 10 Livneh locations. Results derived without the randomization approach to enhancing variance are shown by the asterisks. Results after application of the randomization approach where additional variability is added using a combination of red and white noise, where the red noise autocorrelation is defined by the lag-1 autocorrelation of the predictors weighted by the beta weights in the GLM, are shown by the filled circles (labeled var-E for ERA-Interim, var-G for GFDL, and var-M for MPI). Observed values are indicated by the solid black squares and are labeled by the location name. These results are derived using transfer functions conditioned using ERA-Interim and Livneh data from odd years in the contemporary climate and applied to independent data from the even years.

observations from Birmingham (Fig. 5) and only about half as much variance as the observations (Fig. 4). After the procedure to enhance variability is applied both the variance (Fig. 4) and temporal autocorrelation (Fig. 5) more closely match the observations.

Thus, use of the procedure outlined in Eqs. (1) and (2) to enhance the variance of the downscaled time series has the advantage that the time series of downscaled Tmin and Tmax exhibit a level of variance that more

closely approximates that in the observations (Fig. 4) and exhibits better agreement with the temporal autocorrelation in the observations (Fig. 5). These qualities are important for developing projections of CLIMDEX variables focused on the occurrence and persistence of temperature extremes. However, it is not without penalty. Addition of scaled red and white noise in an amount conditioned on the weighted predictor variance to the downscaled time series inevitably increases the root-mean-square error (RMSE) [and the mean absolute error (MAE)] between the downscaled time series of Tmin and Tmax made using predictors from ERA-Interim and independent observations from the even years (Fig. 6a). The increase in RMSE is slightly larger than that if the variance enhancement employs white noise scaled as the ratio of the variance in the observations to that in the predictands (Fig. 6a).

In general, when the transfer functions derived using predictors from ERA-Interim are applied to independent data from ERA-Interim (i.e., the transfer functions are conditioned based on odd years, and applied to data from even years) or output from GFDL and MPI, the resulting first and second moments of the probability distributions for both predictands (Tmin and Tmax) at all sites are well described (Fig. 6b). Application of a *t* test to test the hypothesis that the downscaled Tmin projections and observations of Tmin derive from Gaussian distributions with equal means and equal (but known) variances results in failure to reject the null hypothesis at a significance level of 1% for both parameters at all sites from all three applications of the transfer functions (to ERA-Interim, GFDL, and MPI). An *F* test applied to test if these data derive from Gaussian distributions with equal variance, is rejected at the 99.9% confidence level for Tmin at Seattle for output from ERA-I, GFDL and MPI. For Tmax the *t*-test statistic results in failure to reject the hypothesis of equal means in all cases. The *F* test for equality of variance results in rejection of the null hypothesis for ARM, Boulder, Colorado; Orlando, and Sioux City in downscaling of Tmax from GFDL. The nonparametric test for equality of variance rejects the null hypothesis only at Boulder and Seattle. Thus, based on this purely statistical assessment of skill downscaling of MPI is equally valid across space and across the two predictands (excluding Seattle for Tmin), while downscaling from GFDL is less skillful for Tmax at Boulder. However, analyses of the probability distribution of predictors indicate that less than 10% of the predictors weighted by their beta weights are correctly reproduced by MPI for the grid cell containing Boulder. Equally, GFDL correctly reproduces predictors that are responsible for only about one-third of the beta weights in the transfer function for Tmin at Phoenix, but this is

FIG. 5. Temporal autocorrelation: (top) Tmin and (bottom) Tmax when the transfer functions developed using ERA-Interim output from the odd years in 1979–2005 are applied to independent data from even years. The colored lines show the temporal autocorrelation for downscaling results from ERA-Interim (red), GFDL (blue), and MPI (green). The dashed lines show results without variance enhancement, and the solid lines show results after variance enhancement.

FIG. 6. (a) RMSE in downscaled T_{min} and T_{max} at the 10 Livneh locations. Five RMSE values are reported for each variable and location: 1 indicates the RMSE for downscaling using transfer functions derived using data from odd years and applied to ERA-Interim (ERA) output from even years without variance inflation, 2 indicates the RMSE for downscaling using transfer functions derived using data from odd years and applied to ERA output from even years with variance enhancement as described using Eqs. (1) and (2), 3 indicates the RMSE for downscaling using transfer functions derived using data from odd years and applied to ERA output from even years with variance enhancement using only white noise multiplier by the variance of the observations divided by the variance in the downscaled predictands, 4 indicates the RMSE for downscaling using transfer functions derived using data from all 27 years of ERA and applied to that same data without variance enhancement, and 5 indicates the RMSE for downscaling using transfer functions derived using data from all 27 years of ERA and applied to that same data with variance enhancement as described using Eqs. (1) and (2). (b) Mean ($\langle X \rangle$) and standard deviation (σ) of T_{min} and T_{max} in independent data from the contemporary climate (i.e., even years in 1979–2005) (Obs) and as downscaled from ERA, GFDL, and MPI GCMs for each of the 10 Livneh locations. The downscaled output is slightly offset to enhance legibility.

not immediately evident in the comparison of mean and variance of T_{min} in the contemporary climate. Internal climate variability may be an important cause of variability in the degree to which single realizations with individual GCMs reproduce the marginal probability distributions of the predictors (Deser et al. 2012).

The appearance of downscaling skill without representation of the underlying predictor probability distributions emphasizes the need for, and value of, a process-level credibility analysis.

b. Projections of T_{min} , T_{max} , and the CLIMDEX metrics

To illustrate use of the DCA in the context of climate projections, GLM transfer functions are derived for both

predictands (T_{min} and T_{max} daily anomalies from the 7-day running mean climatology during 1979–2005) using the entire ERA-Interim and Livneh datasets. The RMSE of the transfer functions for this training period with and without addition of variance enhancement are shown in Fig. 6a. These transfer functions [with variance enhancement as in Eqs. (1) and (2)] are then applied to output from GFDL and MPI for 1979–2005 and 2075–99 to develop time series of daily anomalies of T_{min} or T_{max} at each location. The T_{min} and T_{max} climatologies from the contemporary climate are then added to derive time series of T_{min} and T_{max} and the CLIMDEX indices for the contemporary climate and the end of the current century.

Figure 7 summarizes the downscaling transfer functions and differential credibility assessment traffic lights

Tmin	Predictors	Beta weights		Variance		DCA	
		GFDL	MPI	GFDL	MPI	GFDL	MPI
ARM	1,2,3,5,6,1×7,2×7,3×6,3×7	0.83	0.56	0.84	0.79		
Birmingham	1,2,3,5,6,7,1×2,1×5,1×7,2×7,3×5,3×6,3×7,5×6,6×7	0.56	0.18	0.88*	0.85*		
Boulder	2,3,4,5,6,7,2×6,3×4,3×5,3×6,4×5,5×6	0.51	0	0.85	0.79		
Ft. Logan	2,3,4,5,6,7,2×4,2×6,3×5,3×6,3×7,4×5,4×7	0.76	0.77	0.77*	0.82*		
Orlando	1,2,3,4,5,6,7,1×4,1×5,1×6,2×5,2×6,3×7,4×5	0.62	0.32	0.98	0.96		
Phoenix	2,3,4,6,7,2×3,2×4,2×7,3×4,3×6	0.3	0.47	1.26	1.28		
Pittsburgh	2,3,4,5,7,2×4,2×5,3×4,3×5,3×7,5×7	0.02	0.62	0.77**	0.75**		
Seattle	2,3,4,5,6,7,2×4,3×4,3×5,3×6,4×5,5×7,6×7	0.58	0.75	1.13**	1.17**		
Sioux City	2,3,4,5,6,3×6	0.79	0.85	0.72*	0.7		
Yosemite	1,2,3,4,6,7,1×2,1×6,1×7,2×4,2×7,3×6,3×7	0.9	0.44	1.19*	1.21*		
Tmax	Predictors	Beta weights		Variance		DCA	
		GFDL	MPI	GFDL	MPI	GFDL	MPI
ARM	1,2,3,4,5,7,1×4,1×7,2×3,2×4,2×7,3×7,4×5,4×7	0.8	0.47	0.73**	0.71**		
Birmingham	1,2,3,4,5,6,7,1×2,1×4,1×5,1×6,1×7,2×3,2×5,3×5,3×7,4×5,5×7	0.45	0.06	0.82**	0.79**		
Boulder	1,2,3,4,5,6,7,1×6,2×3,2×6,2×7,3×6,4×6	0.74	0	0.85*	0.78**		
Ft. Logan	1,2,4,5,6,7,1×2,1×4,2×5,2×7,6×7	0.55	0.9	0.8	0.78		
Orlando	1,2,3,5,6,7,2×5,3×5,3×7,5×7	0.38	0.31	1.2**	1.05		
Phoenix	1,2,3,4,5,6,7,1×5,1×6,1×7,2×3,3×5,4×5,4×6,4×7,5×6,5×7	0.27	0.37	1.02	1.03		
Pittsburgh	1,2,3,4,5,6,7,1×3,1×4,1×5,1×7,2×4,2×5,2×7,3×4,3×7	0.02	0.39	0.76**	0.76**		
Seattle	1,2,4,5,6,7,1×2,1×4,1×5,1×7,2×7,4×6,5×6,6×7	0.2	0.94	1.01	0.92		
Sioux City	1,2,4,5,6,7,2×5,2×7,5×7	0.38	0.43	0.63**	0.62**		
Yosemite	2,3,4,5,7,2×3,2×4,2×7,3×4,3×7	0.93	0.54	0.85	0.85		

FIG. 7. Predictors included in the transfer functions of Tmin and Tmax derived using all observations from Livneh and ERA-Interim during 1979–2005. Also shown is the fraction of the beta weights for which the probability distribution of the predictors (anomalies from climatology) for transfer functions for Tmin and Tmax is similar from each of the two GCMs and ERA-Interim during the contemporary climate (1979–2005). The third–sixth columns show the fraction of the beta-weight normalized predictors that are correctly reproduced by GFDL and MPI in the contemporary climate and the amount of variance in Tmin or Tmax reproduced by the downscaling model when applied to GFDL or MPI output. The asterisks indicate rejection of the null hypothesis for the *F* test at the 99% (single asterisk) or 99.9% (two asterisks) confidence level. The columns under the heading DCA show the color-coded results to aid visualization of the DCA. The background color shows the assessment of the beta weights—red: <50%, yellow: 50%–75%, and green: >75% of beta weights are associated with predictors that are well reproduced by the GCMs. The foreground hatching (sometimes the same color as the background) shows the degree to which the predictands (Tmin anomalies from climatology) exhibit the correct amount of variance in the contemporary climate—red: variance is over/underpredicted by >25%, yellow: variance is within 15%–25%, and green: variance is within $\pm 15\%$ of the observed value in the contemporary climate. Predictors: 1 = geopotential height at 500 hPa (Z500), 2 = air temperature at 700 hPa (T700), 3 = specific humidity at 700 hPa (Q700), 4 = air temperature at 500 hPa (T500), 5 = specific humidity at 500 hPa (Q500), 6 = west–east (*u* component) wind speed at 700 hPa (U700), and 7 = south–north (*v* component) wind speed at 700 hPa (V700).

in tabular form to aid interpretation and assess confidence in the projections of these properties. Figures 8–10 illustrate different ways of displaying the climate projections in the context of the DCA. Figure 8 displays projections of the six CLIMDEX indices in tabular form, along with the DCA traffic lights and the observed frequency computed from the Livneh data in the contemporary climate. Figures 9 and 10 show maps of the annual frequency of tropical nights and frost days (from Tmin) and summer days and icing days (from Tmax) at the 10 Livneh locations for the contemporary climate and the projected change in frequency for the end of the twenty-first century. The DCA summarized in Fig. 8 is transparent and affords traceability. For example, at Boulder, the GFDL model reproduces 51% of the beta-weight weighted predictors, and the downscaling approach recovers 85% of the variability in Tmin. Conversely, the MPI model does not reproduce the probability distribution of any of the predictors and also

the downscaling reproduces a lower fraction of the variance in Tmin. Thus, the projections of all Tmin-related CLIMDEX metrics derived using predictors from GFDL should be seen as more credible than those from MPI. For Tmin at ARM the downscaling credibility is higher for GFDL than MPI largely because GFDL better represents the variability in specific humidity at 700 hPa, which is the most important predictor (i.e., has the highest beta weight) in the GLM. If the primary variable of interest is changes in the mean Tmax, then the projections downscaled from MPI for Fort Logan are considerably more credible than those from GFDL, because MPI correctly reproduces the probability distribution of 90% of the beta-weighted normalized predictors, while GFDL has an equivalent score of only 55%. Thus, the projections for greatly increased frequency of WSD from MPI at Fort Logan must be viewed as more credible than the more modest increases indicated by downscaling of GFDL.

Tmin	DCA		Tropical nights			Frost Days			CSD		
	GFDL	MPI	Livneh	GFDL: Current/Future	MPI: Current/Future	Livneh	GFDL: Current/Future	MPI: Current/Future	Livneh	GFDL: Current/Future	MPI: Current/Future
ARM			61	52/72	57/78	92	93/82	91/80	3	2/<1	<1/1
Birmingham			68	58/100	60/102	56	50/37	47/38	2	2/<1	4/1
Boulder			<1	<1/1	<1/2	164	164/143	165/137	2	4/1	2/<1
Ft. Logan			<1	<1/<1	<1/<1	195	199/182	199/178	3	2/2	2/<1
Orlando			144	121/170	169/175	3	1/<1	1/1	1	13/1	8/1
Phoenix			138	133/147	133/150	1	3/3	2/1	3	7/3	5/2
Pittsburgh			13	12/30	13/31	110	114/99	110/98	2	<1/<1	<1/<1
Seattle			<1	<1/<1	<1/<1	27	28/27	30/25	4	4/1	2/1
Sioux City			16	17/29	21/37	151	151/142	150/140	3	<1/<1	1/<1
Yosemite			<1	<1/<1	<1/<1	183	180/165	180/163	6	8/4	3/2
Tmax	DCA		Summer days			Icing days			WSD		
	GFDL	MPI	Livneh	GFDL: Current/Future	MPI: Current/Future	Livneh	GFDL: Current/Future	MPI: Current/Future	Livneh	GFDL: Current/Future	MPI: Current/Future
ARM			154	148/171	146/170	12	5/4	4/3	4	2/10	2/10
Birmingham			175	174/200	168/197	2	<1/<1	<1/<1	6	5/28	6/27
Boulder			92	90/123	89/123	20	18/13	17/11	2	4/18	4/21
Ft. Logan			62	60/80	60/85	42	45/36	47/34	4	2/7	2/14
Orlando			276	260/289	260/288	<1	<1/<1	<1/<1	7	2/58	10/48
Phoenix			245	253/266	251/273	<1	<1/<1	<1/<1	7	8/28	8/40
Pittsburgh			103	98/122	95/124	26	23/18	23/20	4	1/8	1/13
Seattle			42	41/51	39/52	1	1/1	1/1	<1	3/6	4/9
Sioux City			111	104/125	101/126	50	50/41	49/38	2	1/4	1/6
Yosemite			70	67/90	70/91	15	15/12	16/9	8	7/20	6/27

FIG. 8. The frequency of the CLIMDEX metrics in the contemporary climate as derived from the Livneh data and projections for the current and future climates downscaled from GFDL and MPI. Also shown is the DCA for downscaling of Tmin and Tmax. The background color shows the assessment of the beta weights—red: <50%, yellow: 50%–75%, and green: >75% of beta weights are associated with predictors that are well reproduced by the GCMs. The foreground hatching shows the degree to which the predictands (Tmin anomalies from climatology) exhibit the correct amount of variance in the contemporary climate—red: variance is over/underpredicted by >25%, yellow: variance is within 15%–25%, and green: variance is within $\pm 15\%$ of the observed value in the contemporary climate.

Downscaling of both GCM represents the spatial variability in the annual frequency of the CLIMDEX indices (i.e., variation across the 10 locations) and also performs relatively well in terms of agreement at specific locations in the contemporary climate (Fig. 8). The Livneh observations imply a range of annual frequency of tropical nights from <1 to 144, with a median of 14.5, whereas the downscaling of GFDL and MPI for the contemporary climate yields values from <1 to 133 (median = 14.5) and from <1 to 169 (median = 17), respectively. For frost days the observed frequencies range from 1 to 195 days per year (median = 101), whereas downscaled values from GFDL are from 1 to 199 (median = 104) and from MPI are from 1 to 199 (median = 101). CSD is a more challenging metric to downscale. Values from the GLM exhibit the correct range of values across the locations but lower reliability at each individual location. Estimates of CSD in the contemporary climate are from 1 to 6 days per year in the Livneh dataset but range from <1 to 13 using predictors from GFDL and from <1 to 8 using predictors from MPI (Fig. 8).

For the CLIMDEX metrics associated with Tmax, the annual frequency of summer days ranges from 42 to 276 (median = 107) in the Livneh data, and frequencies derived using the downscaled Tmax time series are 41–260 (median = 101) for GFDL and 39–260 (median = 98) from downscaling of MPI. Comparable statistics for icing days indicate a median of 17.5 (range of 1–50) in the observations, median of 10 (<1–50) from GFDL and a median of 10 (<1–49) in downscaling of MPI. The CLIMDEX metric that is downscaled with lowest

confidence is WSD. WSD across the 10 locations ranges from <1 per year to 8 in the observations, whereas values from downscaling of GFDL range from 1 to 8 and from downscaling of MPI range from 1 to 10 per year in the contemporary climate (Fig. 8).

Both the GFDL and MPI model versions downscaled herein were included in an evaluation of the CMIP5 multimodel ensemble for its depiction of regional climate over contiguous United States (CONUS) during 1979–2005 (Sheffield et al. 2013). That analysis found GFDL-ESM2M exhibits a positive bias in terms of number of summer days (average bias of $+33 \text{ yr}^{-1}$) relative to the Hadley Centre Global Historical Climatology Network-Daily (HadGHCND) observations while MPI was negatively biased (spatially averaged bias of -30 yr^{-1}). GFDL exhibits no bias in number of frost days, and MPI exhibits a negative bias (spatial average of -12.5 yr^{-1}). Assuming the 10 Livneh locations are a random sample distributed across CONUS comparison of those results with data from Fig. 8 suggest the statistical downscaling conducted here greatly reduces these biases. Analyses of the CMIP5 ensemble also found spatial variability in the frequency of Tmax > 32°C over the south-central United States (where ARM is located) is better reproduced by GFDL than MPI (Sheffield et al. 2013). This is consistent with the DCA presented here that suggests the drivers of Tmax variability are better captured by GFDL (Fig. 8).

Projected differences in four of the CLIMDEX metrics (future minus current climate) based on downscaling of GFDL and MPI shown in Figs. 9 and 10 are also color coded according to the DCA. At the locations where tropical nights currently occur more than once per year

FIG. 9. Results of T_{min} downscaling as manifest in the frequency of (a),(b) tropical nights and (c),(d) frost days for the contemporary climate from GFDL or MPI as labeled. Also shown is (e)–(h) the difference in frequency per year in the future (2075–99) minus the current (1979–2005) of these CLIMDEX indices. The size of the symbols scales linearly with frequency of occurrence in the contemporary period [in (a)–(d)] or the difference [future minus current; in (e)–(h)] in each metric. The maximum, median, and minimum values across the 10 locations are shown in the legend. The colors denote the differential credibility analysis. The inner circle shows the DCA based on the degree to which the beta-weight weighted GCM-derived predictors correctly characterize the probability distribution of ERA-Interim in the contemporary climate—red: <50% of beta weights are well described, yellow: 50%–75% of beta weights are well described, and green: >75% of beta weights are correctly reproduced. The outer ring denotes the degree to which the predictands (T_{min} anomalies from climatology) exhibit the correct amount of variance in the contemporary climate—red: variance is over/underpredicted by >25%, yellow: variance within 15%–25%, and green: variance is within $\pm 15\%$ of the observed value in the contemporary climate.

on average, they are projected to increase by the end of the century under RCP8.5 in downscaling of both GFDL and MPI (Figs. 8 and 9). The fractional increase in frequency is of greatest magnitude for Birmingham, Orlando, Pittsburgh, and Sioux City for downscaling of both GCM (Fig. 8). Conversely, the frequency of frost days is projected to decline at all locations (Figs. 8 and 9). Cold spell duration is projected to substantially decrease at Phoenix in downscaling of both GFDL and MPI (Figs. 8 and 9). However, Phoenix is a challenging location for downscaling T_{min} from both GCMs, so these projections must be viewed as having low credibility due to the poor

representation of key predictors by both models and the excessive variance (Fig. 8). This may reflect the low variability in T_{min} (Fig. 4), and/or the inability of the GLM as currently formulated to capture important drives of T_{min} such as soil moisture and land–atmosphere coupling (Zhang et al. 2008). This model deficiency is clearly indicated by the DCA and applies to projections for this location of all three CLIMDEX metrics that are derived from T_{min} .

In general, downscaling of T_{max} and associated CLIMDEX metrics is associated with lower credibility than T_{min} (Figs. 8–10). Thus, projections of T_{max} and

FIG. 10. Results of $T_{\max}$ downscaling as manifest in the frequency of (a),(b) summer days and (c),(d) icing days for the contemporary climate from GFDL or MPI as labeled. Also shown is (e)–(h) the difference in frequency per year in the future (2075–99) minus the current (1979–2005) of these CLIMDEX indices. The size of the symbols scales linearly with frequency of occurrence in the contemporary period [in (a)–(d)] or the difference [future minus current; in (e)–(h)] in each metric. The maximum, median, and minimum values across the 10 locations are shown in the legend. The colors denote the differential credibility analysis. The inner circle shows the DCA based on the degree to which the beta-weight weighted GCM-derived predictors correctly characterize the probability distribution of ERA-Interim in the contemporary climate—red: <50% of beta weights are well described, yellow: 50%–75% of beta weights are well described, and green: >75% of beta weights are correctly reproduced. The outer ring denotes the degree to which the predictands ($T_{\max}$ anomalies from climatology) exhibit the correct amount of variance in the contemporary climate—red: variance is over/underpredicted by >25%, yellow: variance is within 15%–25%, and green: variance is within $\pm 15\%$ of the observed value in the contemporary climate.

related CLIMDEX metrics should generally be viewed with greater caution, although the projected changes are broadly consistent with a priori expectation of an increased frequency of summer days, a decline in icing days, and increased WSD (Figs. 8 and 10).

5. Concluding remarks

To aid societal preparedness for climate change it is critical to develop tools to describe projection credibility in different regions and metrics derived using different GCMs. This research is an initial step toward application

of DCA to statistical downscaling using GLM. It provides input to assessments of the credibility with which we can extrapolate contemporary credibility or realism into scenarios of future climate conditions developed using statistical transfer functions conditioned using data from the contemporary climate. While there is no definitive basis for assigning credibility to such projections process-level DCA offers one approach to selecting which are the best candidates for extrapolation.

Our specific focus is on projections of daily extreme temperatures and indices of persistent temperature extremes. The desire to illustrate the DCA in a clear and

succinct manner dictates the precise nature of the transfer functions employed herein. To illustrate how DCA can be undertaken, it is demonstrate here using ESD performed with generalized linear models resistant to overfitting. The transfer functions also employ a new method to enhance predictand variance without distortion of the temporal autocorrelation. This new approach leads to improved representation of variability in T_{min} and T_{max} and the temporal persistence (Figs. 4–6) that are critical to representation of some CLIMDEX metrics. Further this approach allows the temporal autocorrelation of the predictands to evolve under climate change consistent with changes in the temporal autocorrelation of the predictors. However, the approach is not without penalty and leads to increased RMSE between the predictand time series and independent observations (Fig. 6).

Our DCA application focuses on daily distributions as they relate directly to CLIMDEX temperature extremes. Future work could employ different forms of the transfer functions, different methods to derive the predictors and/or predictands, employ different predictors including those designed to capture low-frequency information (e.g., teleconnections) and/or employ GCM ensembles.

This initial DCA focusses on two key aspects of the transfer function structure: reproduction of the predictors and predictand variance. We demonstrate the approach by deriving transfer functions using Livneh observations and reanalysis output and then applying them to output from GCMs. Daily T_{min} and T_{max} is downscaled for 10 locations distributed across the continental USA for the contemporary and possible future climate at the end of the twenty-first century under the RCP8.5 scenario. Differential credibility is objectively assigned to the resulting T_{min} and T_{max} projections and is shown to offer insights that extend beyond those derived solely using standard statistical hypothesis testing of independent data from the contemporary climate.

It is hoped that this research will provide inspiration to those seeking to convey credibility from ESD and that this type of approach can offer guidance to researchers engaged in climate impact studies. In this first attempt to integrate DCA into ESD we have used a simple downscaling approach applied to gridcell-derived predictors and have illustrate a method to convey differential credibility that employs a simple traffic light system. Naturally, the choice of how many classes to use in conveying the DCA results is subjective. Three classes were chosen for ease of visualization and end-user interpretability. The classes could be further discretized or different thresholds applied depending on, for example, the specific risk tolerance of the decision-makers.

Multimodel ensembles are useful tools to explore uncertainty in climate projections. Much has been written about the potential advantages of breaking the model democracy (i.e., apply uneven weightings) in climate model ensembles [such as phase 6 of the Coupled Model Intercomparison Project (CMIP6)] (Collins 2017; Knutti et al. 2017; Lorenz et al. 2018). In addition to providing information relevant to end users of downscaled climate information, the DCA described here may also provide a first step toward constructing more robust ensembles of climate projections from statistical downscaling of multiple GCMs. It could be used to apply varying weights and/or excluding models with relatively low credibility. It could also be extended to include different ESD methodological approaches and/or to incorporate consideration of other assumptions inherent in use of statistical downscaling (e.g., elements such as temporal transferability; Dayon et al. 2015).

Acknowledgments. This work was support by the U.S. Department of Energy (DoE) (DE-SC0016438 and DE-SC0016605). The research used computing resources from the National Science Foundation: Extreme Science and Engineering Discovery Environment (XSEDE) (allocation award to SCP is TG-ATM170024). We acknowledge Seth McGinnis and Rachel McCrary of NCAR for providing the Livneh, ERA-Interim, GFDL, and MPI datasets analyzed herein. The authors acknowledge the thoughtful comments and suggestions of the reviewers.

Data availability statement. Output from ERA-Interim is available online (<http://apps.ecmwf.int/datasets/>). The Livneh dataset is available online (<https://www.esrl.noaa.gov/psd/data/gridded/data.livneh.html>). Data from the CMIP5 GCM suite, including the two models used herein, are available online (<https://esgf-node.llnl.gov/projects/cmip5/>).

REFERENCES

- Addor, N., T. Ewen, L. Johnson, A. Çöltekin, C. Derungs, and V. Muccione, 2015: From products to processes: Academic events to foster interdisciplinary and iterative dialogue in a changing climate. *Earth's Future*, **3**, 289–297, <https://doi.org/10.1002/2015EF000303>.
- Andrews, T., J. M. Gregory, M. J. Webb, and K. E. Taylor, 2012: Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere–ocean climate models. *Geophys. Res. Lett.*, **39**, L09712, <https://doi.org/10.1029/2012GL051607>.
- Barsugli, J. J., and Coauthors, 2013: The practitioner's dilemma: How to assess the credibility of downscaled climate projections. *Eos, Trans. Amer. Geophys. Union*, **94**, 424–425, <https://doi.org/10.1002/2013EO460005>.
- Benestad, R. E., I. Hanssen-Bauer, and D. Chen, 2008: *Empirical-Statistical Downscaling*. World Scientific, 215 pp.
- Bukovsky, M. S., D. J. Gochis, and L. O. Mearns, 2013: Towards assessing NARCCAP regional climate model credibility for

- the North American monsoon: Current climate simulations. *J. Climate*, **26**, 8802–8826, <https://doi.org/10.1175/JCLI-D-12-00538.1>.
- , R. R. McCrary, A. Seth, and L. O. Mearns, 2017: A mechanistically credible, poleward shift in warm-season precipitation projected for the U.S. southern Great Plains? *J. Climate*, **30**, 8275–8298, <https://doi.org/10.1175/JCLI-D-16-0316.1>.
- Cheggwidden, O. S., and Coauthors, 2019: How do modeling decisions affect the spread among hydrologic climate change projections? Exploring a large ensemble of simulations across a diversity of hydroclimates. *Earth's Future*, **7**, 623–637, <https://doi.org/10.1029/2018EF001047>.
- Collins, M., 2017: Still weighting to break the model democracy. *Geophys. Res. Lett.*, **44**, 3328–3329, <https://doi.org/10.1002/2017GL073370>.
- Conover, W., 1999: *Practical Nonparametric Statistics*. John Wiley and Sons, 583 pp.
- Crutcher, H. L., 1975: A note on the possible misuse of the Kolmogorov–Smirnov test. *J. Appl. Meteor.*, **14**, 1600–1603, [https://doi.org/10.1175/1520-0450\(1975\)014<1600:ANOTPM>2.0.CO;2](https://doi.org/10.1175/1520-0450(1975)014<1600:ANOTPM>2.0.CO;2).
- 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>.
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, <https://doi.org/10.1002/qj.828>.
- Deser, C., R. Knutti, S. Solomon, and A. S. Phillips, 2012: Communication of the role of natural variability in future North American climate. *Nat. Climate Change*, **2**, 775–779, <https://doi.org/10.1038/nclimate1562>.
- Di Cecco, G. J., and T. C. Gouhier, 2018: Increased spatial and temporal autocorrelation of temperature under climate change. *Sci. Rep.*, **8**, 14850, <https://doi.org/10.1038/s41598-018-33217-0>.
- Dillon, M. E., H. A. Woods, G. Wang, S. B. Fey, D. A. Vasseur, R. S. Telemeco, K. Marshall, and S. Pincebourde, 2016: Life in the frequency domain: The biological impacts of changes in climate variability at multiple time scales. *Integr. Comp. Biol.*, **56**, 14–30, <https://doi.org/10.1093/icb/icw024>.
- Dixon, K. W., J. R. Lanzante, M. J. Nath, K. Hayhoe, A. Stoner, A. Radhakrishnan, V. Balaji, and C. F. Gaitán, 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>.
- Dunne, J. P., and Coauthors, 2012: GFDL's ESM2 global coupled climate–carbon Earth system models. Part I: Physical formulation and baseline simulation characteristics. *J. Climate*, **25**, 6646–6665, <https://doi.org/10.1175/JCLI-D-11-00560.1>.
- Fatichi, S., and Coauthors, 2016: Uncertainty partition challenges the predictability of vital details of climate change. *Earth's Future*, **4**, 240–251, <https://doi.org/10.1002/2015EF000336>.
- Giorgetta, M. A., and Coauthors, 2013: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *J. Adv. Model. Earth Syst.*, **5**, 572–597, <https://doi.org/10.1002/jame.20038>.
- Giorgi, F., 2019: Thirty years of regional climate modeling: Where are we and where are we going next? *J. Geophys. Res. Atmos.*, **124**, 5696–5723, <https://doi.org/10.1029/2018JD030094>.
- Gutiérrez, J. M., D. San-Martín, S. Brands, R. Manzananas, and S. Herrera, 2013: Reassessing statistical downscaling techniques for their robust application under climate change conditions. *J. Climate*, **26**, 171–188, <https://doi.org/10.1175/JCLI-D-11-00687.1>.
- , and Coauthors, 2019: An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment. *Int. J. Climatol.*, **39**, 3750–3785, <https://doi.org/10.1002/joc.5462>.
- Gutowski, W. J., Jr., and Coauthors, 2020: The ongoing need for high-resolution regional climate models: Process understanding and stakeholder information. *Bull. Amer. Meteor. Soc.*, **101**, E664–E683, <https://doi.org/10.1175/BAMS-D-19-0113.1>.
- Hertig, E., and Coauthors, 2019: Comparison of statistical downscaling methods with respect to extreme events over Europe: Validation results from the perfect predictor experiment of the COST Action VALUE. *Int. J. Climatol.*, **39**, 3846–3867, <https://doi.org/10.1002/joc.5469>.
- Hewitson, B., J. Daron, R. Crane, M. Zermoglio, and C. Jack, 2014: Interrogating empirical-statistical downscaling. *Climatic Change*, **122**, 539–554, <https://doi.org/10.1007/s10584-013-1021-z>.
- Huth, R., J. Kysely, and M. Dubrovsky, 2001: Time structure of observed, GCM-simulated, downscaled and stochastically generated daily temperature series. *J. Climate*, **14**, 4047–4061, [https://doi.org/10.1175/1520-0442\(2001\)014<4047:TSGGS>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<4047:TSGGS>2.0.CO;2).
- Karl, T., W. Wang, M. Schlesinger, R. Knight, and D. Portman, 1990: A method of relating general circulation model simulated climate to the observed local climate. Part I: Seasonal statistics. *J. Climate*, **3**, 1053–1079, [https://doi.org/10.1175/1520-0442\(1990\)003<1053:AMORGC>2.0.CO;2](https://doi.org/10.1175/1520-0442(1990)003<1053:AMORGC>2.0.CO;2).
- Knutti, R., J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring, 2017: A climate model projection weighting scheme accounting for performance and interdependence. *Geophys. Res. Lett.*, **44**, 1909–1918, <https://doi.org/10.1002/2016GL072012>.
- Koenig, W. D., and A. M. Liebhold, 2016: Temporally increasing spatial synchrony of North American temperature and bird populations. *Nat. Climate Change*, **6**, 614–617, <https://doi.org/10.1038/nclimate2933>.
- Krysanova, V., C. Donnelly, A. Gelfan, D. Gerten, B. Arheimer, F. Hattermann, and Z. W. Kundzewicz, 2018: How the performance of hydrological models relates to credibility of projections under climate change. *Hydrol. Sci. J.*, **63**, 696–720, <https://doi.org/10.1080/02626667.2018.1446214>.
- Livneh, B., E. A. Rosenberg, C. Lin, B. Nijssen, V. Mishra, K. M. Andreadis, E. P. Maurer, and D. P. Lettenmaier, 2013: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions. *J. Climate*, **26**, 9384–9392, <https://doi.org/10.1175/JCLI-D-12-00508.1>.
- Lorenz, R., N. Herger, J. Sedláček, V. Eyring, E. M. Fischer, and R. Knutti, 2018: Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. *J. Geophys. Res. Atmos.*, **123**, 4509–4526, <https://doi.org/10.1029/2017JD027992>.
- Maloney, E. D., and Coauthors, 2019: Process-oriented evaluation of climate and weather forecasting models. *Bull. Amer. Meteor. Soc.*, **100**, 1665–1686, <https://doi.org/10.1175/BAMS-D-18-0042.1>.
- Maraun, D., and M. Widmann, 2018: *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press, 360 pp.
- , 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>.

- Markowski, C. A., and E. P. Markowski, 1990: Conditions for the effectiveness of a preliminary test of variance. *Amer. Stat.*, **44**, 322–326, <https://doi.org/10.1080/00031305.1990.10475752>.
- Mearns, L. O., M. S. Bukovsky, and V. J. Schweizer, 2017: Potential value of expert elicitation for determining differential credibility of regional climate change simulations: An exercise with the NARCCAP co-PIs for the southwest monsoon region of North America. *Bull. Amer. Meteor. Soc.*, **98**, 29–35, <https://doi.org/10.1175/BAMS-D-15-00019.1>.
- Ng, A. Y., 2004: Feature selection, L1 versus L2 regularization, and rotational invariance. *ICML'04: Proc. 21st Int. Conf. on Machine Learning*, Banff, AB, Canada, International Machine Learning Society, 354, <https://icml.cc/impls/conferences/2004/proceedings/papers/354.pdf>.
- Pryor, S. C., and J. T. Schoof, 2019: A hierarchical analysis of the impact of methodological decisions on statistical downscaling of daily precipitation and air temperatures. *Int. J. Climatol.*, **39**, 2880–2900, <https://doi.org/10.1002/joc.5990>.
- Reifen, C., and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill? *Geophys. Res. Lett.*, **36**, L13704, <https://doi.org/10.1029/2009GL038082>.
- Riahi, K., and Coauthors, 2011: RCP 8.5—A scenario of comparatively high greenhouse gas emissions. *Climatic Change*, **109**, 33–57, <https://doi.org/10.1007/s10584-011-0149-y>.
- Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464, <https://doi.org/10.1214/aos/1176344136>.
- Sheffield, J., and Coauthors, 2013: North American climate in CMIP5 experiments. Part I: Evaluation of historical simulations of continental and regional climatology. *J. Climate*, **26**, 9209–9245, <https://doi.org/10.1175/JCLI-D-12-00592.1>.
- Sillmann, J., and Coauthors, 2017: Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Wea. Climate Extremes*, **18**, 65–74, <https://doi.org/10.1016/j.wace.2017.10.003>.
- Soleh, A. M., A. H. Wigena, A. Djuraidah, and A. Saefuddin, 2015: Statistical downscaling to predict monthly rainfall using linear regression with L1 regularization (LASSO). *Appl. Math. Sci.*, **9**, 5361–5369, <https://doi.org/10.12988/AMS.2015.56434>.
- von Storch, H., 1999: On the use of “inflation” in statistical downscaling. *J. Climate*, **12**, 3505–3506, [https://doi.org/10.1175/1520-0442\(1999\)012<3505:OTUOH>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<3505:OTUOH>2.0.CO;2).
- Weichselgartner, J., and B. Arheimer, 2019: Evolving climate services into knowledge–action systems. *Wea. Climate Soc.*, **11**, 385–399, <https://doi.org/10.1175/WCAS-D-18-0087.1>.
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. Vol. 100. Academic Press, 704 pp.
- Zhang, J., W. C. Wang, and L. R. Leung, 2008: Contribution of land–atmosphere coupling to summer climate variability over the contiguous United States. *J. Geophys. Res.*, **113**, D22109, <https://doi.org/10.1029/2008JD010136>.