

A hierarchical analysis of the impact of methodological decisions on statistical downscaling of daily precipitation and air temperatures

Hierarchical analysis of statistical downscaling skill

S.C. Pryor¹, and J.T. Schoof²

1. Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY 14853, USA
2. Department of Geography and Environmental Resource, Southern Illinois University, Carbondale, IL 62901, USA

***Correspondence to:** S.C. Pryor (sp2279@cornell.edu)

(Submitted 29 July 2018; Reviews received 15 October 2018; Revision submitted: 29 October 2018; Final revision submitted: 2 January 2019)

Funding information

US Department of Energy, Office of Science (DE-SC0016438) and the US National Science Foundation (TG-ATM170024).

Abstract

Despite the widespread application of statistical downscaling tools, uncertainty remains regarding the role of model formulation in determining model skill for daily maximum and minimum temperature (T_{max} and T_{min}), and precipitation occurrence and intensity. Impacts of several key aspects of statistical transfer function form on model skill are evaluated using a framework resistant to model over-specification. We focus on; 1) Model structure: Simple (generalized linear models, GLM) versus complex (artificial neural networks, ANN) models. 2) Predictor selection: Fixed number of predictors chosen *a priori* versus stepwise selection of predictors and inclusion of grid point values versus predictors derived from application of principal components analysis (PCA) to spatial fields. We also examine the influence of domain size on model performance. For precipitation downscaling, we consider the role of the threshold used to characterize a wet day and apply three approaches (Poisson and Gamma distributions in GLM, and ANN) to downscale wet day precipitation amounts. While no downscaling formulation is optimal for all predictands and at ten locations representing diverse U.S. climates, and due to the exclusion of variance inflation all of the downscaling

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: [10.1002/joc.5990](https://doi.org/10.1002/joc.5990)

formulations fail to reproduce the range of observed variability, models with larger suites of prospective predictors generally have higher skill. For temperature downscaling, ANNs generally outperform GLM, with greater improvements for T_{min} than T_{max} . Use of PCA-derived predictors does not systematically improve model skill, but does improve skill for temperature extremes. Model skill for precipitation occurrence generally increases as the wet day threshold increases and models using PCA-derived predictors tend to outperform those based on grid cell predictors. Each model for wet day precipitation intensity overestimates annual total precipitation and underestimates the proportion derived from extreme precipitation events, but ANN-based models and those with larger predictor suites tend to have the smallest bias.

Key words: Statistical downscaling; model; skill; daily; air temperature; precipitation; predictors; CLIMDEX; United States of America; Livneh; ERA-interim.

1. Motivation and objectives

Many techniques have been applied to empirical-statistical downscaling (ESD) of climate variables each of which employ a range of implicit and explicit assumptions (Fowler et al., 2007; Maraun et al., 2018a; Maraun et al., 2015; Wilby et al., 1998). Just as regional climate modelers seek to demonstrate model transferability to enhance confidence that they can capture possible climate change signals (Jacob et al., 2012; Takle et al., 2007), there is a need for robust, reliable and transferable approaches to ESD (Kundzewicz & Stakhiv, 2010), that have sufficient accuracy for use in adaptation planning (Rössler et al., 2018). However, previous research has illustrated a strong regional dependence of ESD skill and optimal ESD approaches (Fowler et al., 2007; Gutmann et al., 2014; Huth, 1999; Maraun et al., 2018a). Further, although there is an extensive literature describing relative strengths and weaknesses of ESD methods for different near-surface climate variables, most integrative analyses have focused on a comparison of individual and independently applied methods (e.g. the

comprehensive inter-comparison of a sample of opportunity comprising over 50 ESD methods at 86 stations across Europe in a perfect predictor experiment (Maraun et al., 2018a)).

Here we present a hierarchical analysis using a simple ESD framework to assess the relative importance of some key methodological decisions in terms of their impact on model skill for four commonly downscaled variables:

1. Daily minimum air temperature (T_{min})
2. Daily maximum air temperature (T_{max})
3. Probability of precipitation (PoP)
4. Amount of precipitation on a wet day

This work does not seek to be inclusive of methodological approaches and instead employs two classes of transfer functions (generalized linear models (GLM) and machine learning approaches that employ artificial neural networks (ANN)). Predictors are either drawn from the single grid cell in which the surface station is located or derive from principal components analysis (PCA) applied to time-evolving spatial fields of the predictors.

Many previous studies have developed model frameworks that are—to varying degrees—specific to the study region under consideration. Herein we employ a generic framework, but within which predictors for specific stations are selected objectively in an automated fashion and thus may differ from site to site. Further, although many ESD models are fitted either by calendar month or season (see the review articles referenced above), here we place a premium on developing models that can be applied over the full calendar year to avoid abrupt changes in model form at the transition between seasons/calendar months (see examples

given in (Jeong et al., 2012)), and build a framework that can be applied to data sets in which the seasons are evolving (Dwyer et al., 2012; Pryor & Schoof, 2008; Wallace & Osborn, 2002). We further employ measures designed to resist model over-fitting.

Any set of predictors can inevitably account for only part of the variance in the predictand and ESD transfer functions are typically developed using parameter estimates derived from error minimization relative to observations in a training data set. Hence, ESD downscaled temperature and precipitation tend to be negatively biased in terms of variability and a range of methods for artificially increasing the variability in the downscaled predictands have been proposed (e.g. inflation (Karl et al., 1990) and randomization (von Storch, 1999)). However, application of these approaches is not without penalty. Variance inflation (multiplication of the square root of the ratio of variances between the observations and downscaled time series) implicitly assumes variability in the predictors completely specify the variability in the predictand, and increases the mean square error between the adjusted predictions and the observations. Randomization (addition of white or red noise to account for the unexplained variance) changes the temporal autocorrelation of the downscaled output. Here we do not post-process the downscaling output to enhance the downscaled predictand variability, as may be necessary in climate change adaptation applications, but instead focus on the inherent skill of each ESD model (as in Jeong et al., 2012).

2. Data sets

Analyses presented herein focus on 10 locations drawn from different climate zones across the contiguous United States of America (CONUS), with one (or more) from each of the

regions used within the 2014 U.S. National Climate Assessment (Melillo et al., 2014) (Figure 1). This research is part of the Framework for Assessing Climate's Energy-Water-land Nexus using Targeted Simulations (FACETS) project. The objective of the FACETS project is to develop a hierarchical model evaluation framework for deeper diagnostic evaluation of numerical and statistical simulations of key climate features and to compare the differential performance and credibility of various downscaling approaches. The predictors employed in the ESD component of FACETS are thus drawn from the reanalysis product being used to provide one set of lateral boundary conditions for the limited area (Regional Climate Model, RCM) simulations. Within the project the 'contemporary climate' comprises the period 1979-2005, which is the longest period for which ERA-Interim and CMIP5 generation GCMs are available for the contemporary climate. Within FACETS the future period is an equal number of years at the end of the current century. A gridded data set of predictands (at $1/16^\circ$ resolution, equal to about $6 \times 6 \text{ km}^2$ at 40°N) is used within the project to facilitate evaluation of RCM output from the highest simulation resolutions of 4 km (evaluated relative to individual Livneh "grid cells" (Livneh et al., 2013)) and 12 km (a 2 by 2 grid cell average). A perfect predictors experimental design is used herein to evaluate ESD model skill in the absence of error in the large-scale predictors using two well-known and frequently used data sets:

- The predictands (Tmin, Tmax, precipitation occurrence and amount on a wet day) are drawn from the Livneh data set (Livneh et al., 2013). These data are gridded at a spatial resolution of $1/16^\circ$ latitude/longitude and are derived from daily temperature and precipitation observations at $\sim 20,000$ NOAA Cooperative Observer stations. A meta-

analysis of eight gridded climate datasets found the Livneh data set exhibits the smallest mean bias in precipitation nationwide, good agreement with station observations for Tmax and Tmin, and represents some aspects of temperature extremes (Behnke et al., 2016).

- The predictors are daily mean values computed from 6-hourly values of upper-air variables at a spatial resolution of approximately 80 km (T255 spectral resolution) drawn from the ERA-Interim reanalysis data set (Dee et al., 2011):
 - 1) Geopotential height at 500hPa (Z500).
 - 2) Air temperature at 700hPa (T700).
 - 3) Specific humidity at 700hPa (Q700).
 - 4) Air temperature at 500hPa (T500).
 - 5) Specific humidity at 500hPa (Q500).
 - 6) West-east (u-component) wind speed at 700hPa (U700).
 - 7) South-north (v-component) wind speed at 700hPa (V700).

One of the first key decisions in ESD is selection of the predictors (Maraun et al., 2018a). This can be done objectively (e.g. using backward stepwise regression and/or partial correlation coefficients (Hessami et al., 2008)) or subjectively as is recommended in use of the Statistical Down-Scaling Model (SDSM) (a hybrid regression based weather generator approach (Wilby et al., 2002)). Further key decisions include the number of upper air variables necessary to represent important drivers of variability in the predictand, and whether to use upper air variables directly (from a specific grid cell or over a domain (Amendola et al., 2017; Jeong et al., 2012)) or linear combinations of variables derived from

principal component analysis (PCA) or a synoptic classification (Schoof & Pryor, 2001; Vrac et al., 2007; Yang et al., 2010). All of the predictor variables considered here have been used in prior downscaling analyses of precipitation and/or near-surface temperature over the contiguous USA (Pryor et al., 2017; Schoof et al., 2010; Schoof et al., 2009). The justification for including them is as follows: We focus on upper-air variables because ESD is frequently applied to relatively coarse resolution output from global models that may distort near-surface fields (McKendry et al., 2006; Reichert et al., 1999). Most past research has indicated that predictors describing the upper-level circulation are necessary, but insufficient, to capture the variability of near-surface air temperature and precipitation. For example, research over northern Europe found large-scale temperature fields (often at 850 hPa) were the best predictors for near-surface air temperature (Huth, 1999), while a combination of atmospheric circulation indices and tropospheric humidity were best for local precipitation (e.g. (Hanssen-Bauer et al., 2005), and both Tmax and Tmin over Canada (Jeong et al., 2012)). Thus, our minimum set of predictors is Z500, T700 and Q700 (i.e. 1) to 3) in the list above). The additional variables considered herein (T500, Q500, U700 and V700) were selected because lower-tropospheric moisture and thermal advection (e.g. $U700 \times T700$) is known to be of great importance to near-surface temperature extremes over the eastern USA (Pryor et al., 2017), and has been used to detect and diagnose the southwestern monsoon (Cavazos et al., 2002). Mid-tropospheric T and/or Q (i.e. T500 and Q500) were selected by stepwise optimization for either precipitation occurrence/sequencing or wet-day amount in one or more regions of the contiguous US, even when T700 and Q700 were included in the potential predictor pool (Schoof et al., 2010).

We evaluate two aspects of predictor selection. First, whether expanding the predictor suite from 1)-3) from the list above, to 1)-7) and permitting first-order interactions substantially increases model skill. Prior research has suggested that while ESD using single grid point predictors exhibit skill, the optimal grid cell is sometimes displaced from the location from which the predictand(s) are drawn (Brinkmann, 2002) and that a meteorological classification may offer advantages over predictors drawn from an individual grid cell (Sauter & Venema, 2011). Thus, we also evaluate the impact on skill of using grid-cell predictors versus indices that incorporate larger-scale spatial information. S-mode principal components analysis (PCA) with Varimax (orthogonal) rotation (Richman, 1986) is applied to daily spatial fields of all seven predictors transformed to Z-scores drawn from an $n \times n$ grid of ERA-interim output centered on the station of interest. Past research has illustrated the importance of domain size in dictating the relationship between circulation patterns and surface climate, and provided evidence that highest skill is achieved for domain sizes approximately equal to the majority of synoptic scale systems (i.e. 1300-1800 km) (Beck et al., 2016). Thus, the number of grid cells used is varied over $n = 13:2:25$ (where $n=19$ equates to an area $\sim 1500 \times 1500$ km centered on the station of interest). Based in part on inspection of scree plots (White et al., 1991), the first X unrotated principal components that explain $> 1\%$ of the variance are retained and rotated. PC scores are computed for each component on each day. This approach is inclusive (i.e. PC scores are computed for all days) and is not used to define distinct weather types, but to develop continuous indices that describe the similarity of each day to major nodes of variability (Schoof & Pryor, 2001), thus avoiding inherent problems of within class variability (Brinkmann, 2000).

The time period used is 1979-2005. The FACETS project protocol is that ESD techniques are trained using data from 14 “odd” years (1979, 1981, 1983, ...) and evaluated using independent (test) data from 13 “even” years (1980, 1982, 1984, ...). Thus, the training data set comprises 5110 days of data, while the testing data set comprises 4752 days.

3. Temperature downscaling

Particularly in the mid-latitudes, near-surface air temperature variability comprises systematic variations (dominated by diurnal and seasonal variability), long-term trends and “residual” temporal variations. For downscaling of T_{min} and T_{max} , the predictors and predictands are subject to seasonal detrending prior to constructing any ESD models. Removing the nonstationarity associated with the seasonal cycle permits a more direct assessment of predictor-predictand relationships and also addresses some key concerns with ESD using GLM (Estrada et al., 2013). Model over-fitting is a persistent challenge within ESD approaches (Soleh et al., 2015), and multiple co-linearity of predictors can make least-squares estimates of regression coefficients unstable. Use of daily anomaly values (anomalies from the climatological mean) reduces the temporal autocorrelation; e.g. at ARM the lag-1 temporal autocorrelations of the Z500, T700 and Q700 grid cell predictors are 0.93, 0.87 and 0.75, respectively, which drop to 0.73, 0.64 and 0.44 for the daily anomaly values. Use of daily anomalies also decreases predictor co-linearity; e.g. for ARM Pearson cross-correlations (r) of the Z500, T700 and Q700 grid cell predictors are 0.91, 0.64 and 0.64, respectively, which drop to 0.75, 0.06 and 0.17 for the daily anomaly values. Finally, it causes the predictors to more closely approximate a Gaussian distribution which is useful if

one wishes to evaluate the importance of individual predictors by computing beta weights for the coefficients (the ratio of the standard deviations of the predictor and predictand multiplied by the regression coefficient) (Wilks, 2011).

3.1. Model form

To reduce the risk of overfitting, GLM built using all predictors (i.e. all seven grid cell predictors and/or all PC scores) and permitting first-order interaction terms are trained by first selecting predictors using 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. Regression with L1 regularization (using the lasso (least absolute shrinkage and selection operator) method in MATLABTM) (Ng, 2004; Soleh et al., 2015) is applied to predictors selected by stepwise regression using 10-fold cross validation to derive the final model that is applied to independent predictors from the test period (i.e. even years).

Machine learning (including ANN) is increasingly being employed within the ESD community (Lakshmanan et al., 2015; Pryor et al., 2017). The primary potential advantage over GLM approaches is that ANN can represent highly non-linear relationships. Herein ANN are constructed separately for each site (within MATLABTM) using the Levenberg–Marquardt back-propagation algorithm and three hidden layers (where the number of hidden layers dictates the possible complexity of the model form). To avoid overtraining of the ANN models, 80% of the training data set (i.e. data from ‘odd years’) are randomly drawn to build the ANN, while the remaining 20% is used for network validation. ANN developed here are

relatively complex deep learning algorithms (three hidden layers) and so require a relatively large sample for the network training. Thus, we develop 100 network realizations for each parameter and station based on 100 random draws from the training data sets to ensure the ‘first’ network does not exhibit anomalous validation behavior and to provide insights into the stability of the network performance as represented by the dispersion of the mean square error (MSE) generated when the 100 network forms are applied to the test data sets (i.e. independent data from the even years).

There are six models for each temperature predictand (daily anomalies of T_{min} and T_{max}) for each location. Model 1 uses GLM and three predictors (grid cell Z500, T700 and Q700). Model 2 uses ANN and the three grid cell predictors. Model 3 uses GLM employing the seven grid cell predictors with first order interaction terms permitted (i.e. stepwise followed by L1 regularized regression). Model 4 uses the seven grid cell predictors and ANN. Models 5 and 6 use PC scores within GLM (stepwise and L1 regularized regression) and ANN, respectively.

To investigate the importance of domain size (number of grid cells) used to develop indices of the synoptic scale meteorology, PCA is conducted for seven different domain sizes for the 10 stations (i.e. 13×13 to 25×25 ERA-interim grid cells). The number of PCs retained (and rotated) typically increases with domain size presented to the PCA (see Table 1), and varies from eight in the case of the smallest domain at Birmingham and Pittsburgh, to 12 for the largest domain for Orlando. T_{min} and T_{max} daily anomalies are downscaled using a single ESD model; ANN with three hidden layers for each of the seven domain sizes. The root mean square error (RMSE) and Pearson correlation coefficient (*r*) relative to independent

observations in the test data set (i.e. data from even years) indicate the skill of ESD models that employ ANN and PC scores is relatively insensitive to the precise number of grid cells on which PCA is performed for the range of domain sizes considered herein (Figure 2). The range of maximum to minimum r and RMSE computed from the seven different ANN models relative to independent observations at the 10 locations are; 0.01 to 0.05 and 0.05-0.19 K for T_{min} and 0.01 to 0.03 and 0.04-0.18 K for T_{max} . Thus, once the domain is broadly of dimensions equal to the synoptic scale, the PCs appear to be relatively stable and to include an almost equal amount of information relevant to downscaling of near-surface temperatures. Transfer functions conditioned using PC scores computed for all seven variables over a domain comprising 19×19 grid cells (i.e. approximately 1500×1500 km) exhibits the highest r and lowest RMSE for T_{max} at eight of the ten stations (Figure 2) and is also associated with the most frequently occurring number of ‘significant’ PCs at eight of the ten stations (Table 1). Thus, in the following analyses PC scores are generated using this size of domain.

3.2. Skill metrics applied

Several key aspects of the ESD skill for T_{min} and T_{max} are evaluated. First, their ability to represent the entire probability distribution of daily anomalies (1st to 99th percentile values) versus observations in the test data (even years) is examined. The RMSE, mean bias (MB) and r computed from pairwise comparison of downscaled and observed (time synchronized) daily anomaly values are used to examine the model performance in a temporal context. We also assess the ability of the ESD models to capture the occurrence and time synchronization

of two key temperature ‘events’. In this analysis the mean seasonal cycle is added to predicted daily anomalies to identify days when $T_{min} < 0^{\circ}C$, or $T_{max} > 32.2^{\circ}C$. The former is referred to in CLIMDEX literature (Zhang et al., 2011) as ‘frost days’, while the latter relates to maximum daily temperatures above $90^{\circ}F$. For any value of relative humidity the heat index associated with air temperatures in excess of $90^{\circ}F$ is deemed as representing a risk for heat-related illness under guidance provided by the Occupational Safety and Health Act (see discussion in (Arbury et al., 2014)). Results are evaluated using Hit Rates (H) and False Alarm Rates (F) and the odds ratio (OR) (Stephenson, 2000):

$$OR = \frac{H}{1-H} / \frac{F}{1-F}$$

where;

$$H = \frac{a}{a+c}$$

$$F = \frac{b}{b+d}$$

a = number of correct predictions of an event

b = number of false alarms

c = number of misses

d = number of correct forecasts of no event

H and F (and thus OR) are computed using output for each day in the test data set, so for a hit to be recorded the prediction must be correct for that calendar date. $OR = 1$ implies independence of the predictions and observations. Values above 1 indicate increasing association between the “forecast” and observations and thus increasing skill of the prediction. OR is equal to the conditional joint probabilities of predictions and observations

and not their marginal probabilities, thus it is insensitive to model bias, so we contextualize the *OR* with rate of occurrence, H and F . Lastly, we evaluate ESD models with respect to the frequency of four CLIMDEX metrics (Zhang et al., 2011) applied after adding the mean seasonal cycle; number of frost days ($T_{\min} < 0^{\circ}\text{C}$), number of icing days ($T_{\max} < 0^{\circ}\text{C}$), number of summer days ($T_{\max} > 25^{\circ}\text{C}$) and number of tropical nights ($T_{\min} > 20^{\circ}\text{C}$). We only report statistics for the CLIMDEX indices for a location if the observations indicate that these conditions are met on an average of five or more days per year.

3.3. Results

Consistent with past research and *a priori* expectations none of the ESD models for daily T_{\min} and T_{\max} anomalies capture the full variability of observed values because the predictors do not represent all of the drivers of variability in the predictands. There is a warm bias in the low percentiles and cold bias in the upper percentiles of the daily T_{\min} and T_{\max} anomalies (Figures 3 and 4). The magnitude of these biases ranges from $\pm 2^{\circ}\text{C}$ in the 1st to 5th and 95th to 99th percentile anomalies at Phoenix and Yosemite in T_{\min} and T_{\max} anomalies, to $\pm 5^{\circ}\text{C}$ at the tails of the distribution at ARM. There is some weak qualitative evidence that the bias in the tails of the distribution of anomalies scale with the variance in the observed daily temperature anomalies. However, this is not uniformly the case. For example, biases in the ESD models for T_{\min} at ARM are large despite the comparatively narrow distribution of observed anomalies (i.e. daily deviations from the mean seasonal climatology) (Figure 3). Particularly at Fort Logan and Seattle, but to some degree at all sites, GLM with more predictors (i.e. L7 and LP) outperforms those with only three predictors (L3) particularly at

the tails of the T_{min} and T_{max} distributions of daily anomalies (Figures 3 and 4). Further, models that use ANN (A7, AP) exhibit higher skill in capturing the range of daily anomalies over the GLM (L7 and LP). For example, models built with ANN generally exhibit lower bias in the tails of the distribution of T_{min} anomalies (e.g. Birmingham, Fort Logan, Orlando and Seattle, Figure 3). This effect is also present, but is less pronounced, in T_{max}. There is less evidence for a consistent impact from use of grid cell specific variables versus PC scores as predictors.

Mean RMSE of daily T_{max} and T_{min} anomalies averaged across all ten sites is comparable to, but slightly lower than, that reported in a previous analysis of 25 sites distributed across Canada (RMSE of 3.4-3.6°C) (Jeong et al., 2012). For daily T_{min} anomalies, the RMSE across the six ESD models and 10 stations range from 1.7 to 3.9°C, while $r = 0.62$ to 0.82 (Figure 5). There is substantial variability in model skill across the 10 sites, and relatively high consistency in terms of the performance of different model forms (Figure 5). All model functional forms exhibit comparatively poor performance for T_{min} and T_{max} at the same locations, indicating that the climate at those locations is particularly challenging to downscale even with non-linear techniques. For example, all models exhibit strong biases in the distributional tails of daily T_{min} anomalies (Figure 3), high RMSE (3.2 to 3.4°C) and relatively low r (0.63-0.69) at ARM (Figure 5). Nevertheless, the simplest models (L3) consistently exhibit lowest skill. The RMSE is highest and r is lowest L3 downscaling models for T_{min} anomalies at nine of 10 locations. Use of ANN improves skill for daily T_{min} anomalies over GLM consistent with previous research (Schoof & Pryor, 2001). The RMSE is smaller and r is higher at all 10 locations in ESD models for T_{min} anomalies using ANN

(A7) than in GLM (L7) built using grid cell predictors. For T_{min} daily anomalies, r is highest and RMSE is lowest at six of the ten stations in either the GLM or ANN transfer functions that using PC scores as predictors. Transfer functions built with ANN exhibit lowest RMSE and highest r at eight of the ten stations (versus two for the GLM). Thus, although there is a clear tendency for ANN models to outperform GLM there is a division between whether use of grid cell or PC score predictors is the optimal model form.

An advantage of the GLM applied with grid cell predictors is interpretability. All preferred GLM forms for T_{min} include at least one (and often multiple) linear interaction term(s) (Table 2a). Most frequently these interaction terms represent thermal or moisture advection (i.e. interaction terms between air temperature and the zonal and/or meridional wind components at 700 hPa or interaction terms between specific humidity and the zonal and/or meridional wind components at 700 hPa) (Table 2a). This is consistent with the finding that ANN derived results have slightly higher skill since machine learning inherently permits predictor interactions. All L7 models include both T700 and Q700 as predictors for daily T_{min} anomalies. While at 9 of 10 stations Z500 is a selected predictor for T_{max}, this predictor is included in models for T_{min} at only three of 10 stations (Table 2a). This may be due to predictor covariability, and hence that variability in Z500 is closely mirrored by that in T700 and/or Q700, or it may reflect the fact that nocturnal minimum air temperatures (T_{min}) are strongly influenced by near-surface humidity and cloud cover (and thus Q700) (Easterling et al., 1997). All GLM that employ PC scores as predictors also include one or more interaction terms, potentially indicating that even though the PCs represent the dominant

orthogonal modes in the data set, days that exhibit high association with multiple of those modes are important to describing the overall variability in T_{\max} and T_{\min} .

The frequency of frost days and tropical nights are generally well reproduced by all ESD models. However, there is a substantial negative bias in the number of frost days at Birmingham and Seattle particularly in models that do not employ PC scores as predictors (Figure 5). The number of tropical nights at Pittsburgh and Sioux City is also under-estimated by all ESD models (Figure 5). The AP model performs best for these challenging locations but still exhibits substantial negative bias in the frequency of occurrence of frost days (by up to 20%) and in tropical nights (by up to 21%). The OR for $T_{\min} < 0^{\circ}\text{C}$ exceeds 3 at all stations and ESD models (Figure 5). Although the L3 model shows highest OR for predicting the occurrence of $T_{\min} < 0^{\circ}\text{C}$ at Phoenix (Figure 5), the L3 models exhibit lowest OR at the other locations. The hit rate (H) and false alarm rate (F) for this threshold provide important information for interpreting the OR . At ARM, Boulder, Fort Logan, Pittsburgh, Sioux City and Yosemite, all model forms have $H > 0.8$ and $F < 0.12$. Orlando and Phoenix both exhibit a low actual occurrence of frost days (fewer than three per year on average in the Livneh dataset for Orlando and once per year in Phoenix). At Orlando and Phoenix $H < 0.3$, F is also low, and all ESD model forms underestimate the frequency of occurrence of $T_{\min} < 0^{\circ}\text{C}$ at these two sites. The model form with the lowest overall F is AP, while the ensemble mean H for the 10 locations is similar for; L7, A7, LP and AP in part because the AP model exhibits relatively poor performance at Orlando and Phoenix in terms of predicting the occurrence of frost days.

The summary of ESD for T_{min} indicates that ANN models most accurately represent the probability density function of daily anomalies. These models also generally perform slightly better for the frequency of frost days and tropical nights and the timing of frost days, particularly when PC scores are used as predictors (i.e. AP). There is a clear benefit to using more predictors than only Z500, Q700 and T700 irrespective of whether the transfer functions take a linear or non-linear form.

Consistent with results for T_{min}, L3 models for T_{max} daily anomalies exhibit highest RMSE (and mean bias (MB)) at nine (eight) of the 10 stations, while r with observed daily anomalies is lowest at eight of 10 stations for these models with the smallest number of predictors and a linear transfer function (i.e. A3 outperforms L3) (Figure 6). Use of ANN only slightly (but consistently) decreases the RMSE and increases r when only three grid cell predictors are used (Figure 6). Lowest RMSE and highest r is associated with A7 model at three stations, LP at three stations and at four stations for AP. Thus, there is evidence of clear benefit moving from the models that can draw on only three grid cell predictors, to more complex forms. However, as with the T_{min} daily anomalies, no single model form is associated with the best model performance for these metrics at all 10 stations, and there is considerable spatial consistency in terms of skill across model forms. For example, lowest RMSE relative to observations (and high r) is found for Phoenix for all six ESD models (Figure 6). The number of icing days derived from all ESD models are within $\pm 25\%$ at five of the six locations but exhibit substantial negative bias at ARM where an average of 10.5 days/year exhibit icing day conditions in the observations while the model-based predictions indicate ~3-4 days/year (Figure 6). This bias is smallest for models that employ the PC scores as

predictors, but is still substantial. The number of summer days (i.e. $T_{\max} > 25^{\circ}\text{C}$) is better reproduced at all stations for all model forms than the number of icing days. The frequency of $T_{\max} > 25^{\circ}\text{C}$ (i.e. summer days) is within $\pm 5\%$ of observations at nine of the 10 locations for all model forms. All ESD models perform most poorly for this metric at Seattle, where observations indicate this threshold is exceeded on approx. 42 days/year, while the models indicate an annual average occurrence rate of 33-39 days/year (Figure 6). All ESD models thus exhibit skill in predicting the pairwise occurrence of extreme heat ($> 32.2^{\circ}\text{C}$) even without application of variance inflation. The relatively high *OR* for these indices of extreme temperatures ($OR > 3$) demonstrates some evidence of ‘getting the answer right for the right reasons’. However, no clear inference can be drawn about the best model form for $T_{\max} > 32.2^{\circ}\text{C}$. Use of PC scores in both GLM and ANN (i.e. LP and AP) increase the *OR* for some sites (e.g. ARM and Yosemite), but A7 exhibits the highest *OR* at four sites (Figure 6). The summary of models for T_{\max} indicates that, as with results for T_{\min} , there is a clear benefit to using more predictors than only Z500, Q700 and T700. Models that employ ANN also, on average, most accurately represent probability density functions of daily T_{\max} anomalies. More complex models also generally perform slightly better for estimating the frequency of occurrence and timing of icing days and extreme heat, especially when PC scores are used as predictors (i.e. LP or AP).

ANN are initialized using random weights and more deep learning networks require larger training data. As discussed above, 100 independent ANN are constructed for T_{\min} and T_{\max} at each site using 100 samples randomly drawn from the training dataset (odd years) and the results used to evaluate the dispersion of the mean square error (MSE) between the model

Author Manuscript

predictions of Tmin and Tmax daily anomaly (from climatology) values in the test data set. The results of this analysis are used as a measure of the stability of the ANN. Model performance is relatively consistent across the 100 realizations indicating that over-fitting is not strongly manifest in these results (Figure 7). Further, the one-standard deviation dispersion of the model performance metric (MSE) across the 100 members from the three ANN model forms (A3, A7 and AP) at many sites are not inclusive. ANN trained using either the seven grid-cell predictors or the seven PC scores are associated with demonstrably smaller error in the independent samples of Tmin and Tmax than those that employ only the three grid cell predictors (A3). Recall MSE is strongly impacted by the range of daily anomalies from the climatological mean (Figure 3 and 4), hence MSE values are highest at Fort Logan for Tmin and Sioux City for Tmax. Nevertheless, Figure 7 also emphasizes the high-degree of site-to-site variability in downscaling model skill. While at most sites the MSE in either Tmin or Tmax is minimized for ANN that employ PC scores as predictors (AP), that is not uniformly the case and at three of the ten sites the model form for Tmin that employs the seven grid cell predictors (A7) have lowest mean MSE and a one-sigma range that lies below that of the AP model.

4. Precipitation downscaling

Daily precipitation represents a substantial challenge to both dynamical and statistical downscaling (Maraun et al., 2010). Typically, Regional Climate Models exhibit compensating errors in that they tend to “rain too frequently” but to underestimate the intensity of precipitation events (as recognized over 20 years ago (Mearns et al., 1995)),

leading to development of a range of bias correction approaches (Maraun, 2013; Themeßl et al., 2011). Most ESD approaches (particularly those based on GLM and not specifically designed to capture extreme events) exhibit similar behavior (Maraun et al., 2010) to a degree that is location specific (Vrac et al., 2007). Most (but not all) ESD approaches for modeling precipitation occurrence and amount employ a two-stage approach in which precipitation occurrence is first modeled and then amounts on wet days are assessed (Maraun et al., 2010). This framework is also used herein. As in the case of the downscaling of T_{min} and T_{max} we seek to examine the relative inherent skill of the different downscaling models and thus do not apply methods to inflate the variance of the downscaled variables.

4.1. Model form

The predictors are identical to those used in the temperature downscaling except no deseasonalization is undertaken since precipitation occurrence and amount may not exhibit a dominant (single) seasonal peak, and grid cell predictors (1) to (5) are transformed using a Box-Cox transformation to reduce skewness (Wilks, 2011). The wind components (predictors (6) and (7)) are closer in their raw form to being Gaussian distributed and are not transformed.

Previous studies have found ANN performed relatively poorly in reproducing the PoP (Fowler et al., 2007). Thus, the three model forms for PoP all employ logistic regression with 10-fold cross validation (Wilks, 2009). A wet day is declared on any day when the logistic regression derived PoP exceeds 0.5. GLM and ANN are then applied to modeling of wet day amounts. Previous research applying bias correction techniques has also demonstrated a clear dependence on the threshold used to define a wet-day (Gutmann et al., 2014). Three

thresholds are used here; > 0 mm/day, > 0.1 mm/day and >1 mm/day that are inclusive of the typical range used in ESD (Gutmann et al., 2014) and the 1 mm/day used in the CLIMDEX indices. Hence, nine logistic regression models are derived for PoP at each location. Models L3-1 to -3 use only three predictors (grid cell Z500, T700 and Q700) for the three different wet day thresholds. Models L7-1 to -3 use all seven grid cell predictors, stepwise and regularized logistic regression for the three thresholds. Models LP-1 to -3 use PC scores with stepwise and regularized logistic regression also for the three thresholds.

We develop and apply models for wet day amount based on output from models conditioned to predict the PoP estimates according to each threshold. We consider two distributional forms to fit the wet day amounts; Poisson (Schoof & Pryor, 2001) and the gamma distribution with a log-link function (Fealy & Sweeney, 2007), wherein both are applied with a GLM framework solved using maximum likelihood methods again with 10-fold cross validation. Finally, models of wet day amounts are also developed using ANN with three hidden layers. Accordingly, there are a total of 27 models for wet day amount for each location. Nine are for a threshold of > 0 mm/day, nine are for a wet-day threshold 0.1 mm/day and nine are for a wet-day threshold of 1 mm/day. These models are referred to herein using three letter abbreviations wherein the first letter indicates whether the predictors are grid cell values (G) or are principal component scores (P). The second letter denotes the number of predictors used where 3 denotes use of the three grid-cell predictors, S all seven grid-cell values or PC scores. The final letter denotes use of gamma (G) or Poisson (P) distributions in the GLM or ANN (A). To analyze the stability of ANN for the temperature parameters, 100 ANN realizations are developed for each station and precipitation threshold.

4.2. Skill metrics applied

Some ESD approaches model short-term temporal dependence of precipitation occurrence using Markov-chains (Bellone et al., 2000; Hughes et al., 1999; Schoof & Pryor, 2008). In this analysis temporal sequencing is not explicitly treated in the downscaling models but is used as an independent skill metric to determine if use of output from different model formulations (e.g., using PCA to incorporate spatial information) inherently manifests improvements in the simulation of lag-1 persistence (Hertig et al., 2018). We also compute *OR* for predictions of a wet day. Other indices used derived in part from the CLIMDEX metrics and are; a simple precipitation intensity index (average wet day precipitation, mm/day), mean bias (mm/day), annual total precipitation (mm) (and interannual variability), and the R90pTOT (i.e. sum of annual total precipitation received on wet days when the rain rate exceeds the 90th percentile value).

4.3. Results

OR values for time-wise predictions of wet day occurrence are above 1 for all nine ESD models indicating some degree of skill at all sites (Figure 8). They vary from a minimum value of 1.5 for model L3-1 at Fort Logan to 3.6 for model LP-3 at Yosemite (Figure 8). The Pearson correlation coefficient (*r*) between the mean *OR* (across all 9 models) and mean PoP (mean of values for the three thresholds) from the 10 stations is -0.35 (Spearman rank correlation = -0.19) which indicates that sites with a higher PoP have a slight tendency towards exhibiting lower *OR*. All models for Phoenix and Seattle exhibit relatively high *OR*

(Figure 8), but Phoenix has a comparatively low overall PoP (approx. 9-19%) while the PoP in Seattle is 37-60% depending on the threshold applied (Figure 9).

The *OR*, and thus model skill in predicting the time-wise occurrence of a wet day, tends to increase with increased model complexity (e.g. *OR* is higher for models L7-X and LP-X than in L3-X). At all ten locations ESD models that use the seven grid cell predictors outperform those that employ three, and at nine locations models built using PC scores as predictors exhibit higher *OR* than models using grid cell predictors (Figure 8). *OR* also tend to increase with use of a higher threshold (e.g. model L3-3 typically exhibit higher *OR* than model L3-1, and model LP-3 has a higher *OR* than model LP-1) (Figure 8), possibly because days with larger amounts of precipitation are more distinct from dry days than days when only very light rain occurs. It is important to recall that *OR* is insensitive to model bias as illustrated using the case of Phoenix. Phoenix has a low PoP irrespective of which threshold is used (Figure 9), and relatively high *OR* for all model formulations (Figure 8). The PoP for a wet day threshold of 0.1 mm/day is 0.16 in the training sample (i.e. 814 out of 5110 days in the odd years), and 0.18 in the test sample (854 out of 4752 days). The PoP for the ESD logistic regression model predictions for the independent data from the even years for a threshold of 0.1 mm/day is:

- Three grid cell predictors. PoP = 0.08 (373 out of 4752 days)
- Seven grid cell predictors, stepwise and regularized regression. PoP = 0.09 (428 days)
- PC scores, stepwise and regularized regression. PoP = 0.10 (486 days)

Thus, there is a modest gain in accuracy from using PC scores as predictors, all models are systematically biased towards predicting a dry day at Phoenix (Figure 9), resulting in a

negative bias in lag-1 persistence of wet days in all ESD models for this location (Figure 10). The mean precipitation rate (intensity) on wet days in the test data set that are correctly predicted by all downscaling models exceeds that on all wet days by an average of 27%. This is consistent with the expectation that days with heavier precipitation are more different in terms of the prevailing synoptic scale meteorology than days characterized by light precipitation.

The observed probability that a dry day is followed by a dry day and that a wet day follows a wet day exceeds 60% at all ten stations in the test data sets. The ESD models generally capture this high degree of persistence though no information was provided to ensure this behavior. An *a priori* expectation was use of PC scores, which represent a broader spatial footprint than the grid-cell based predictors, may increase the skill in reproducing persistence in precipitation regimes. Dry-dry and wet-wet persistence probabilities from ESD models using PC scores are indeed generally in closest accord with the observations (Figure 10). For eight stations the LP-2 model best represents the lag-1 persistence of dry days (Figure 10). For nine stations the L3-2 model form is least well able to capture the lag-1 persistence of dry days. However, discrepancies with the observed values across the three model forms is relatively small and indeed in data from Orlando transfer functions built using the PC scores as predictors perform marginally worse than the other two model forms for both dry-dry and wet-wet persistence probabilities (Figure 10). Seasonal variation in PoP at all sites is relatively represented by the ESD models (despite no model input designed to cause this behavior), but all overestimate the PoP in the summer at ARM, at Orlando during summer, and at Yosemite in the early months of the year (Figure 9).

Mid-tropospheric circulation variables (i.e. Z500, U700 and V700) appear to be almost universally important predictors of PoP along with Q700 (Table 2b). In contrast to *a priori* expectations, only one site has an interaction term between specific humidity and the horizontal wind component (i.e. advective moisture fluxes (Yang et al., 2010)) in the PoP models and only two stations include these interaction terms for precipitation amount on a wet day (Table 2b). There is also a tendency towards simpler (i.e. few predictor) models for amount than in the models for PoP at all stations (Table 2b), consistent with past research in Canada (Jeong et al., 2012).

Mean intensity on a wet day is generally well reproduced by all ESD models for all wet day thresholds (Figure 11). For example, for a wet day threshold of 0.1 mm/day the mean intensity is within 2.1 mm/day of the observations for all station and model forms, and the mean bias is < 1 mm/day (Figure 11b). The sum of precipitation on wet days with rainfall rates above the 90th percentile value is underestimated by all ESD models conditioned using all three thresholds (0 mm/day, 0.1 mm/day, 1 mm/day), as is the annual total precipitation at most sites for the lower thresholds (Figures 11a-b). These biases are present, but less pronounced, in the more complex models (GS-X and PS-X). For a wet day threshold of 1 mm/day all ESD models are systematically biased towards over prediction of mean intensity and annual total, but are negatively biased in amount above the 90th percentile (P90) (Figure 11c).

The distribution chosen to represent wet day amounts has a much smaller impact than use of more grid cell predictors (cf. GS-X suite of models vs. G3-X) and the improvement in accuracy achieved by use of PC scores as predictors. Downscaled estimates from the GS-X

(where X is P or G) models of mean intensity on a wet day for a threshold of 0.1 mm/day lie within 22% at seven of the ten locations and are within 60% at all ten locations. The PS-P models exhibit smallest bias in terms of amount above P90. When averaged over all 10 locations it is 40% for a wet day threshold of 0 mm/day, 45% for a wet day threshold of 0.1 mm/day and 64% for a wet day threshold of 1 mm/day. The biases are an average of 1.25 times higher in all models built only using the three grid cell predictors irrespective of the transfer function form.

Two consistent signals are present in the ESD output for annual total precipitation and interannual variability. Models using PC scores as predictors are associated with the smallest bias when the 10 locations are treated as an ensemble (i.e. viewed as a whole), and models that use PC scores and a Poisson distribution to represent wet day amounts (PS-P) perform best for both the mean annual total and interannual variability (Figure 11). All models are systematically, negatively, biased in terms of reproducing the interannual variability, but again irrespective of the threshold applied to define a wet day the models that use PC scores as predictors are associated with smallest bias.

The 100 independent ANN conditioned on wet day precipitation amount indicate relatively low 10th to 90th percentile ranges, and thus high repeatability in terms of performance when applied to independent data (Figure 12). The dispersion of mean square errors (MSE) when measured in absolute terms is largest at Yosemite in part as a consequence of the relatively high mean annual total precipitation (approx. 1200 mm) (Figure 11b). There is some evidence that the models exhibit largest variability when only the three grid cell predictors are used, and the spread of MSE from the 100 ANN developed for Birmingham along with

the divergence of ANN model weights and biases implies the training data set is insufficient to develop deep robust (repeatable) deep learning algorithms at this site, particularly when the three grid cell predictors are applied (Figure 12).

5. Evaluation in the context of the VALUE end-user survey

The VALUE project (Gutiérrez et al., 2018; Maraun et al., 2018b; Maraun et al., 2015) included a stakeholder/end user survey (62 participants) that found temperature and precipitation were the top two climate variables of interest. The required accuracy was $\pm 10-20\%$ for temperature and $\pm 10-50\%$ for precipitation and temporal and spatial resolutions of daily and point or < 1 km are desired (Rössler et al., 2018). Due to large fractional errors that can arise close to 0°C , we assume a mean air temperature of 15°C and equate a $\pm 15\%$ uncertainty to $\pm 2.25^{\circ}\text{C}$. Downscaled daily anomalies for T_{min} and T_{max} from each ESD model are used to quantify the number of days during the independent test sample that this acceptable threshold of uncertainty is exceeded in the absence of any use of variance correction. Although LP and AP models (which use PC scores as predictors) perform best, for all ten locations all six ESD models exceed this threshold on over one-fifth of downscaled days. The minimum number of differences in T_{min} anomaly values (19%) is found for the AP model at Seattle. Conversely, at ARM, 43-46% of daily T_{min} anomalies differ from observed values by this threshold or more for the six ESD models. The smallest number of days with T_{max} differences between downscaled and observed anomalies in excess of 2.25°C is found for the A7 model for Phoenix (17%). A deviation of $\pm 30\%$ between downscaled and observed daily precipitation on a wet day is exceeded on approximately one-quarter of

wet days at most locations for most ESD models that employ a threshold of 0.1 mm/day. Fewer than 5% of wet days for all ESD models exceed this threshold for Phoenix, but up to 43% do for Pittsburgh. These results emphasize the discrepancy between the needs of climate information end users and the inherent skill of current-generation ESD approaches. While noting that no statistical model will incorporate all sources of predictand variability and that the models developed herein do not include variance inflation to account for unexplained variance, given they also do not include all sources of uncertainty (e.g. use of perfect predictors), the results illustrate the urgent need for further investment in improving downscaling approaches to meet stakeholder requirements.

6. Summary and concluding remarks

Our results, derived from a systematic experiment across diverse climates, are in contrast to some earlier work that has indicated simple ESD methods perform as well as more sophisticated methods in reproducing mean characteristics (Fowler et al., 2007), and that ANN-based transfer functions do not exhibit higher skill for T_{min} and T_{max} (Khan et al., 2006). Results presented herein indicate use of non-linear model forms does enhance ESD model skill. Given the same predictors, ANNs generally perform better than GLM for downscaling daily T_{max} , T_{min} , and precipitation intensity on a wet day. Use of ANN also leads to improved performance in downscaling extreme temperature metrics (even in the absence of variance enhancement) particularly at sites where the overall ESD performance is comparatively poor.

Enhancing model complexity by adding more predictors (while employing measures to reduce over-training) also generally enhances downscaling skill for T_{min}, T_{max}, PoP and wet day accumulated precipitation. Use of only Z500, T700 and Q700 as predictors either in GLM or ANN for T_{min} or T_{max} is associated with consistently poorer performance relative to models employing a larger (potential) array of grid cell predictors or indices (PCs) derived from spatial fields of the predictors. The improvement in downscaling skill associated with larger sets of predictors extends to measures associated with extremes. The simplest GLM with fewest predictors perform poorly in reproducing the tails of the probability distributions of daily T_{min} and T_{max} anomalies (Figures 3 and 4). The precise domain over which PCA is applied to geophysical fields to derive the PC scores that are used as predictors in the ESD models has little impact on the overall downscaling skill for domains of approximately 1000×1000 to 2500×2500 km (Figure 2), and is thus of lesser importance to the overall skill of ESD models than use of GLM versus ANN model forms.

All logistic regression models (even with stepwise and regularized regression) perform relatively poorly in terms of reproducing the seasonal variability in PoP. Precipitation amounts on a wet day, amount total, and rainfall rates on high precipitation days, are better simulated particularly for more complex models that use PC scores as predictors. In general, using a higher threshold to characterize a wet day (1 mm/day vs. 0.1 or 0 mm/day) leads to improvement in PoP downscaling as measured using odds ratios. While all of the precipitation occurrence models perform similarly, at most stations models employing PCs derived from predictor spatial fields leads to improved representation of the seasonal cycle (Figure 9) and in model performance as measured by *OR*. While none of the approaches

tested here fully reproduce the lag-1 persistence of dry days and wet days, models using PCs as predictors generally perform better than those based on values from a single grid point. Each of ESD models reproduces the mean wet day precipitation intensity with a small bias, but underestimates the amount of precipitation occurring on days with rainfall rates exceeding the 90th percentile value. However, the underestimation is generally smallest in more complex models (those with larger numbers of predictors or using ANN rather than Poisson or Gamma representations of wet day precipitation intensity within a GLM framework).

There are some important caveats that should be applied to our findings:

- This analysis is a perfect predictor study, and is not inclusive of all possible sources of skill variability of ESD models (e.g. it does not include multiple reanalysis datasets (Manzanas et al., 2015)). Further, application of downscaling in a climate change context introduces several additional sources of uncertainty. These include emissions (or radiative forcing) pathways and climate model characteristics. Factors such as the predictor selection are also key to determining the ability to respond to climate non-stationarity. Predictors are usually selected based on an association with high frequency variability in the predictand (as here) but may not sufficiently manifest factors that induce low frequency variance.
- Our systematic assessment of model skill across diverse climates and a range of evaluation metrics produces some generalizable findings across an array of climate contexts. Nevertheless, some are inevitably partly specific to the data sets used (i.e.

Livneh and ERA-interim) and the relatively short data period (1979-2005) used for model training and testing.

- We consider statistical skill i.e. association between the predictions and observations in independent test data sets drawn from alternating years to those use to develop the transfer functions. This has two important implications:
 1. Samples used in the training and testing differ very little. This approach of building the transfer functions on odd years and testing on even years, though widely used in the downscaling community, implicitly conveys very little about the elasticity of models in the presence of an evolving climate. It has been suggested that this could be addressed by, for example, training precipitation downscaling models on dry years and testing them on wet years (Schoof, 2015; Wilks, 1999), but this approach likely focusses on capturing internal climate variability rather than a response to changes in the radiation balance.
 2. Model skill as presented herein is a measure solely of the statistical performance. The value or utility of any downscaling product will necessarily be determined based on the application (see discussion in (Fowler et al., 2007; Maraun et al., 2015)).

Despite these caveats, by considering the role of downscaling model form (GLM vs. ANN) and model complexity (small/large predictor sets, utility of PCA-derived predictors representing a larger spatial footprint) on skill across a range of variables and metrics, we have demonstrated that nonlinear models (ANN) are generally more skillful than GLM and that use of predictors with a larger spatial footprint (via PCA) also leads to more skillful

prediction. This is particularly true for extremes, for which simple approaches typically exhibit the largest errors.

Acknowledgments

This research was funded by the US Department of Energy, Office of Science (DE-SC0016438) and was enabled by access to computational resources supported by the NSF Extreme Science and Engineering Discovery Environment (XSEDE) (award TG-ATM170024). We acknowledge Seth McGinnis and Rachel Macrary of NCAR for providing the Livneh and ERA-interim datasets analyzed herein, and discussions with our colleagues in the FACETS project. We also acknowledge the thoughtful comments and suggestions of two reviewers.

References

- Amendola, S., Maimone, F., Pasini, A., Ciciulla, F., & Pelino, V. (2017). A neural network ensemble downscaling system (SIBILLA) for seasonal forecasts over Italy: winter case studies. *Meteorological Applications*, 24(1), 157-166.
- Arbury, S., Jacklitsch, B., Farquah, O., Hodgson, M., Lamson, G., Martin, H., & Profitt, A. (2014). Heat illness and death among workers-United States, 2012-2013. *MMWR. Morbidity and mortality weekly report*, 63(31), 661-665.
- Beck, C., Philipp, A., & Streicher, F. (2016). The effect of domain size on the relationship between circulation type classifications and surface climate. *International Journal of Climatology*, 36(7), 2692-2709.
- Behnke, R., Vavrus, S., Allstadt, A., Albright, T., Thogmartin, W. E., & Radeloff, V. C. (2016). Evaluation of downscaled, gridded climate data for the conterminous United States. *Ecological Applications*, 26(5), 1338-1351.

- Bellone, E., Hughes, J. P., & Guttorp, P. (2000). A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Climate research*, 15, 1-12.
- Brinkmann, W. A. (2002). Local versus remote grid points in climate downscaling. *Climate research*, 21, 27-42.
- Brinkmann, W. A. R. (2000). Modification of a correlation-based circulation patterns classification to reduce within-type variability of temperature and precipitation. *International Journal of Climatology*, 20, 839-852.
- Cavazos, T., Comrie, A. C., & Liverman, D. M. (2002). Intraseasonal variability associated with wet monsoons in southeast Arizona. *Journal of Climate*, 15(17), 2477-2490.
- Dee, D. P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., . . . Bauer, P. (2011). The ERA-Interim reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, 137(656), 553-597.
- Dwyer, J. G., Biasutti, M., & Sobel, A. H. (2012). Projected changes in the seasonal cycle of surface temperature. *Journal of Climate*, 25(18), 6359-6374.
- Easterling, D. R., Horton, B., Jones, P. D., Peterson, T. C., Karl, T. R., Parker, D. E., . . . Jamason, P. (1997). Maximum and minimum temperature trends for the globe. *Science*, 277(5324), 364-367.
- Estrada, F., Guerrero, V. M., Gay-García, C., & Martínez-López, B. (2013). A cautionary note on automated statistical downscaling methods for climate change. *Climatic change*, 120(1-2), 263-276.
- Fealy, R., & Sweeney, J. (2007). Statistical downscaling of precipitation for a selection of sites in Ireland employing a generalised linear modelling approach. *International Journal of Climatology*, 27(15), 2083-2094.
- Fowler, H. J., Blenkinsop, S., & Tebaldi, C. (2007). Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, 27, 1547-1578.
doi:10.1002/joc.1556
- Gutiérrez, J., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., . . . Kotlarski, S. (2018). An intercomparison of a large ensemble of statistical downscaling methods

- over Europe: results from the VALUE perfect predictor cross
International Journal of Climatology, *In press*. Available online:
<https://doi.org/10.1002/joc.5462>
- Gutmann, E., Pruitt, T., Clark, M. P., Brekke, L., Arnold, J. R., Raff, D. A., & Rasmussen, R. M. (2014). An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resources Research*, *50*(9), 7167-7186.
- Hanssen-Bauer, I., Achberger, C., Benestad, R., Chen, D., & Førland, E. (2005). Statistical downscaling of climate scenarios over Scandinavia. *Climate research*, *29*(3), 255-268.
- Hertig, E., Maraun, D., Bartholy, J., Pongracz, R., Vrac, M., Mares, I., . . . Soares, P. M. (2018). Comparison of statistical downscaling methods with respect to extreme events over Europe: Validation results from the perfect predictor experiment of the COST Action VALUE. *International Journal of Climatology*, *In press*.
<https://doi.org/10.1002/joc.5469>.
- Hessami, M., Gachon, P., Ouarda, T. B., & St-Hilaire, A. (2008). Automated regression-based statistical downscaling tool. *Environmental Modelling & Software*, *23*(6), 813-834.
- Hughes, J. P., Guttorp, P., & Charles, S. P. (1999). A non
-homogeneous hid
model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(1), 15-30.
- Huth, R. (1999). Statistical downscaling in central Europe: evaluation of methods and potential predictors. *Climate research*, *13*, 91-101.
- Jacob, D., Elizalde, A., Haensler, A., Hagemann, S., Kumar, P., Podzun, R., . . . Wilhelm, C. (2012). Assessing the transferability of the Regional Climate Model REMO to different COordinated Regional Climate Downscaling EXperiment (CORDEX) regions. *Atmosphere*, *3*, 181-199.
- Jeong, D., St-Hilaire, A., Ouarda, T., & Gachon, P. (2012). Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada. *Stochastic environmental research and risk assessment*, *26*(5), 633-653.

- Karl, T., Wang, W., Schlesinger, M., Knight, R., & Portman, D. (1990). A method of relating General Circulation Model simulated climate to the observed local climate. Part I: Seasonal statistics. *Journal of Climate*, 3, 1053-1079.
- Khan, M. S., Coulibaly, P., & Dibike, Y. (2006). Uncertainty analysis of statistical downscaling methods. *Journal of Hydrology*, 319(1-4), 357-382.
- Kundzewicz, Z. W., & Stakhiv, E. Z. (2010). Are climate models “ready for prime time” in water resources management applications, or is more research needed? *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, 55(7), 1085-1089.
- Lakshmanan, V., Gilleland, E., McGovern, A., & Tingley, M. (2015). *Machine Learning and Data Mining Approaches to Climate Science*: Springer, ISBN 978-3-319-17219-4.
- Livneh, B., Rosenberg, E. A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K. M., . . . Lettenmaier, D. P. (2013). A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions. *Journal of Climate*, 26(23), 9384-9392.
- Manzanas, R., Brands, S., San-Martín, D., Lucero, A., Limbo, C., & Gutiérrez, J. (2015). Statistical downscaling in the tropics can be sensitive to reanalysis choice: a case study for precipitation in the Philippines. *Journal of Climate*, 28(10), 4171-4184.
- Maraun, D. (2013). Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate*, 26(6), 2137-2143.
- Maraun, D., Huth, R., Gutiérrez, J. M., Martín, D. S., Dubrovsky, M., Fischer, A., . . . Pongrácz, R. (2018a). The VALUE perfect predictor experiment: evaluation of temporal variability. *International Journal of Climatology*, *In press*. doi: 10.1002/joc.5222.
- Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., . . . Thiele-Eich, I. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48, 10.1029/2009rg000314. doi:10.1029/2009rg000314
- Maraun, D., Widmann, M., & Gutierrez, J. M. (2018b). Statistical downscaling skill under present climate conditions: A synthesis of the VALUE perfect predictor experiment.

International Journal of Climatology, In press. Available online;

<https://doi.org/10.1002/joc.5877>.

Maraun, D., Widmann, M., Gutiérrez, J. M., Kotlarski, S., Chandler, R. E., Hertig, E., . . .

Wilcke, R. A. (2015). VALUE: A framework to validate downscaling approaches for climate change studies. *Earth's Future*, 3(1), 1-14.

McKendry, I., Stahl, K., & Moore, R. (2006). Synoptic sea

-level pressure

by a general circulation model: comparison with types derived from NCEP/NCAR re

Analysis and Implications for Downscaling.

26(12), 1727-1736.

Mearns, L., Giorgi, F., McDaniel, L., & Shields, C. (1995). Analysis of daily variability of precipitation in a nested regional climate model: comparison with observations and doubled CO₂ results. *Global and Planetary Change*, 10(1-4), 55-78.

Melillo, J. M., Richmond, T. C., & Yohe, G. W. (Eds.). (2014). *Climate Change Impacts in the United States: The Third National Climate Assessment*: Government Printing Office, Available online at: <https://nca2014.globalchange.gov/>. doi: 10.7930/J0Z31WJ2.

Ng, A. Y. (2004). *Feature selection, L1 versus L2 regularization, and rotational invariance*. Paper presented at the ICML'04: proceedings of the 21st International Conference on Machine Learning, Available from:

<https://icml.cc/impls/conferences/2004/proceedings.html>.

Pryor, S. C., & Schoof, J. T. (2008). Changes in the seasonality of precipitation over the contiguous USA. *Journal of Geophysical Research*, 113(D21), 10.1029/2008jd010251. doi:10.1029/2008jd010251

Pryor, S. C., Sullivan, R. C., & Schoof, J. T. (2017). Modeling the contributions of global air temperature, synoptic-scale phenomena and soil moisture to near-surface static energy variability using artificial neural networks. *Atmospheric Chemistry and Physics*, 17, 14457-14471. doi:10.5194/acp-2017-367

Reichert, B. K., Bengtsson, L., & Åkesson, O. (1999). A statistical modeling approach for the simulation of local paleoclimatic proxy records using general circulation model output. *Journal of Geophysical Research: Atmospheres*, 104(D16), 19071-19083.

- Richman, M. B. (1986). Rotation of principal components. *Journal of Climatology*, 6, 293-335.
- Rössler, O., Fischer, A. M., Huebener, H., Maraun, D., Benestad, R. E., Christodoulides, P., . . . Kanamaru, H. (2018). Challenges to link climate change data provision and user needs—perspective from the COST ~~International Journal of~~ *Journal of Climatology*, In press. <https://doi.org/10.1002/joc.5060>
- Sauter, T., & Venema, V. (2011). Natural three-dimensional predictor domains for statistical precipitation downscaling. *Journal of Climate*, 24(23), 6132-6145.
- Schoof, J. T. (2015). High ~~resolution~~ -resolution projection of precipitation over the contiguous US. *Journal of Geophysical Research: Atmospheres*, 120(8), 3029-3042.
- Schoof, J. T., & Pryor, S. C. (2001). Downscaling temperature and precipitation: a comparison of regression-based methods and artificial neural networks. *International Journal of Climatology*, 21, 773-790.
- Schoof, J. T., & Pryor, S. C. (2008). On the proper order of Markov chain model for daily precipitation occurrence in the contiguous United States. *Journal of Applied Meteorology and Climatology*, 47, 2477-2486.
- Schoof, J. T., Pryor, S. C., & Suprenant, J. (2010). Development of daily precipitation projections for the United States based on probabilistic downscaling. *Journal of Geophysical Research*, 115(D13106), doi:10.1029/2009JD013030.
- Schoof, J. T., Shin, D., Cocke, S., LaRow, T., Lim, Y. K., & O'Brien, J. (2009). Dynamically and statistically downscaled seasonal temperature and precipitation hindcast ensembles for the southeastern USA. *International Journal of Climatology*, 29(2), 243-257.
- Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461-464.
- Soleh, A. M., Wigena, A. H., Djuraidah, A., & Saefuddin, A. (2015). Statistical downscaling to predict monthly rainfall using linear regression with L1 regularization (LASSO). *Applied Mathematical Sciences*, 9(108), 5361-5369.
- Stephenson, D. B. (2000). Use of the “odds ratio” for diagnosing forecast skill. *Weather and Forecasting*, 15(2), 221-232.

Takle, E. S., Roads, J., Rockel, B., Gutowski Jr., W. J., Arritt, R. W., Meinke, I., . . . Zadra, A. (2007). Transferability intercomparison: An opportunity for new insight on the global water cycle and energy budget. *Bulletin of the American Meteorological Society*, 88, 375-384.

Themeßl, J. M., Gobiet, A., & Leuprecht, A. (2011). Empirical error correction of daily precipitation from regional climate models. *International Journal of Climatology*, 31(10), 1530-1544. - statistical down

von Storch, H. (1999). On the use of "inflation" in statistical downscaling. *Journal of Climate*, 12(12), 3505-3506.

Vrac, M., Stein, M., & Hayhoe, K. (2007). Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. *Climate research*, 34, 169-184. doi:10.3354/cr00696

Wallace, C. J., & Osborn, T. J. (2002). Recent and future modulation of the annual cycle. *Climate research*, 22(1), 1-11.

White, D., Richman, M., & Yarnal, B. (1991). Climate regionalization and rotation of principal components. *International Journal of Climatology*, 11, 1-25.

Wilby, R. L., Dawson, C. W., & Barrow, E. M. (2002). SDSM— a decision support tool for the assessment of regional climate change impacts. *Environmental Modelling & Software*, 17(2), 145-157. doi:[http://dx.doi.org/10.1016/S1364-8152\(01\)00060-3](http://dx.doi.org/10.1016/S1364-8152(01)00060-3)

Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., & Wilks, D. S. (1998). Statistical downscaling of general circulation model output: a comparison of methods. *Water Resources Research*, 34(11), 2995-3008.

Wilks, D. S. (1999). Multisite downscaling of daily precipitation with a stochastic weather generator. *Climate research*, 11, 125-136.

Wilks, D. S. (2009). Extending logistic regression to provide full MOS forecasts. *Meteorological Applications*, 16(3), 361-368. - probability - dis

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Oxford, UK: Academic press.

Yang, W., Bárdossy, A., & Caspary, H.-J. (2010). Downscaling daily precipitation time series using a combined circulation-and regression-based approach. *Theoretical and Applied Climatology*, 102(3-4), 439-454.

Zhang, X., Alexander, L., Hegerl, G. C., Jones, P., Tank, A. K., Peterson, T. C., . . . Zwiers, F. W. (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change*, 2(6), 851-870.

Captions

Figure 1: Map of continental USA showing the ten locations from which the predictands are drawn. The letters in parentheses after each location name are the abbreviations used in Figures 2, 5, 6, 7, 10, 11 and 12. The background colors used for the individual states denote the regions used within the 2014 National Climate Assessment. The abbreviation ARM denotes the U.S. Department of Energy Atmospheric Radiation Measurement Climate Research Facility.

Figure 2: Pearson correlation coefficient (r) and Root Mean Square Error (RMSE) between independent observed daily Tmin and Tmax anomalies and downscaled predictions for each of the ten locations. The ESD transfer functions are built using ANN (using 3 hidden layers) where the predictors are PC scores from different domain sizes (where the number of grid cells used in the spatial domain presented to the PCA is shown in the legend). Results from a domain of 19×19 grid cells that is used within this manuscript is highlight by the black squares. The locations are referred to using the abbreviations introduced in Figure 1.

Figure 3: Difference in the 1st to 99th percentile of daily anomalies of minimum temperature (T_{\min}) from climatology in observations from the independent data set (even years) and six different transfer functions. The solid light-blue lines denote the 1st to 99th percentile observed daily Tmin anomalies (the scale is shown on the right-hand axes). The abbreviations used to identify the transfer functions are as follows; L3 denotes a transfer function built using multiple linear regression with three grid cell predictors (Z500, T700 and Q700), A3

indicates transfer functions built using artificial neural networks with three grid cell predictors (Z500, T700 and Q700), L7 is for a transfer function built using the seven grid cell predictors (Z500, T700, Q700, T500, Q500, U700, V700) but allowing for first order interactions of the predictors and stepwise procedure for selecting the predictors (based on BIC) and the regularized multiple linear regression. A7 is for a transfer function built using artificial neural networks with seven grid cell predictors (Z500, T700, Q700, T500, Q500, U700, V700). LP is for a transfer function built using the PC scores, allowing for first order interactions of the predictors and stepwise procedure for selecting the predictors (based on BIC) and the regularized multiple linear regression. AP is for a transfer function built using artificial neural networks and PC scores.

Figure 4: As Figure 3 but for daily anomalies of maximum temperature (T_{\max}) from climatology in observations from the independent data set (even years) and six different transfer functions. The solid light-blue lines denote the 1st to 99th percentile observed daily Tmax anomalies (shown on the right-hand axes).

Figure 5: Skill assessment of the different ESD models as applied to Tmin. RMSE, r and MB are computed for the daily. Bias in the number of frost days and tropical nights and *OR* for $T_{\min} < 0^{\circ}\text{C}$ is computed once the climatology has been added to the predicted anomalies. A value of -0.1 for tropical nights indicate the downscaled values underestimate the number of tropical nights by 10% relative to observations. *OR* are computed from the hit rate and false alarm rate and thus test time synchronicity. The abbreviations used to identify the ESD models are as in Figure 3 and 4.

Figure 6: As in Figure 5. Skill assessment of the different ESD methods as applied to Tmax. RMSE, r and MB are computed for the daily deviations from climatology. The bias in number of icing days and summer days in the 13 years and the *OR* for the occurrence of $T_{\max} > 32.2^{\circ}\text{C}$ are computed once the climatology has been added.

Figure 7: Mean square error (MSE) of (a) Tmin and (b) Tmax at each of the ten locations (see legend) for 100 ANN derived from the training data set and applied to the independent

(test) data. The results are shown for ANN that employ the three grid cell predictors (A3), seven grid cell predictors (A7) and the PC scores (AP). The symbols denote the mean of the ensemble of 100 ANN while the range denotes plus or minus one standard deviation.

Figure 8: The odds ratios (*OR*) for the probability of precipitation (PoP) at the ten stations for the nine ESD models applied to the independent test data. Model abbreviations are as follows: L3 denotes logistic regression using grid cell values of Z500, T700 and Q700. L7 denotes stepwise logistic regression with L1 regularization wherein the predictors are Z500, T700, Q700, T500, Q500, U700, V700. LP indicates stepwise logistic regression with L1 regularization with PC scores as predictors. -1 indicates a wet day threshold of > 0 mm/day (red), -2 indicates a threshold of 0.1 mm/day (yellow) and 3 is used for a threshold of > 1 mm/day (red).

Figure 9: PoP in the test data from the ten stations. The annual average PoP derived from the observations for the three thresholds is shown by the dashed horizontal lines, while equivalent estimates from the three different ESD models are indicated by the symbols close to y-axis. The monthly mean PoP from the observations and each of the models for a threshold for a wet day of > 0.1 mm/day are shown by the lines with symbols on them. Model abbreviations are as in Figure 8.

Figure 10: Lag-1 persistence estimates from the observed precipitation time series at each station and those from ESD models for a wet day threshold of > 0.1 mm/day (abbreviated as -2). The model abbreviations are as in Figure 8.

Figure 11: Results for the ten stations of four aspects of the precipitation climate; mean intensity on a wet day, mean bias on a wet day, mean annual total (and interannual variability) and the amount (in mm) of precipitation above the 90th percentile value. Note in the panel showing annual total precipitation and interannual variability therein, the observed values are slightly displaced on the horizontal axis to aid legibility. Frame (a) shows results for a precipitation threshold of 0 mm/day, (b) for 0.1 mm/day and (c) for 1 mm/day for a wet day. Nine transfer function (ESD models) are shown the first two letters denote the

predictors; G3 = 3 predictors are used (grid cell values of Z500, T700 and Q700), GS = 7 grid cell predictors, and PS denotes models built using the PC scores as predictors. The final letter denotes the transfer function form; P is Poisson for the link function in the logistic regression, G is for gamma for the link function in the logistic regression, and A is ANN.

Figure 12: Mean square error (MSE) of precipitation amounts on a wet day (for a wet day threshold of 0.1 mm/day) at each of the ten locations (see legend) for each of the 100 ANN as trained on the training data (odd years) and applied to the independent (test) data. The results are shown for ANN that employ the three grid cell predictors (L3-2), seven grid cell predictors (L7-2) and the PC scores (LP-2). The symbols denote the median value of the ensemble of 100 ANN while the range denotes the span from the 10th to 90th percentile values.

Table 1: Number of principal components selected as a function of domain size (specified in terms of number of grid cells, where a value of 19 indicates the domain is an area of 19 by 19 grid cells centered on the station of interest) and station

Table 2: Predictors selected for (a) the linear models of Tmin and Tmax using a combination of stepwise and L1 regularization of the regression models and (b) logistic regression of precipitation occurrence using a threshold of 0.1 mm/day to define a wet day using a combination of stepwise and L1 regularization of the regression models.