Constructing High-Resolution, Bias-Corrected Climate Products: A Comparison of Methods

MAIKE F. HOLTHUIJZEN,^a BRIAN BECKAGE,^a PATRICK J. CLEMINS,^a DAVE HIGDON,^b AND JONATHAN M. WINTER^c

^a *University of Vermont, Burlington, Vermont*

^b *Virginia Tech, Blacksburg, Virginia*

^c *Dartmouth College, Hanover, New Hampshire*

(Manuscript received 28 October 2020, in final form 25 January 2021)

ABSTRACT: High-resolution, bias-corrected climate data are necessary for climate impact studies at local scales. Gridded historical data are convenient for bias correction but may contain biases resulting from interpolation. Long-term, quality-controlled station data are generally superior climatological measurements, but because the distribution of climate stations is irregular, station data are challenging to incorporate into downscaling and bias-correction approaches. Here, we compared six novel methods for constructing full-coverage, high-resolution, bias-corrected climate products using daily maximum temperature simulations from a regional climate model (RCM). Only station data were used for bias correction. We quantified performance of the six methods with the root-mean-square-error (RMSE) and Perkins skill score (PSS) and used two ANOVA models to analyze how performance varied among methods. We validated the six methods using two calibration periods of observed data (1980–89 and 1980–2014) and two testing sets of RCM data (1990–2014 and 1980–2014). RMSE for all methods varied throughout the year and was larger in cold months, whereas PSS was more consistent. Quantile-mapping bias-correction techniques substantially improved PSS, while simple linear transfer functions performed best in improving RMSE. For the 1980–89 calibration period, simple quantile-mapping techniques outperformed empirical quantile mapping (EQM) in improving PSS. When calibration and testing time periods were equivalent, EQM resulted in the largest improvements in PSS. No one method performed best in both RMSE and PSS. Our results indicate that simple quantile-mapping techniques are less prone to overfitting than EQM and are suitable for processing future climate model output, whereas EQM is ideal for bias correcting historical climate model output.

KEYWORDS: North America; Temperature; Bayesian methods; Regression analysis; Statistical techniques; Statistics; Climate models; Reanalysis data; Regional models

1. Introduction

High-resolution (≤ 1 km) gridded climate products with both fine spatial and temporal resolutions are crucial to assessing the effects of a changing climate on social and ecological systems at local scales (Flint and Flint 2012; Holden et al. 2011; Franklin et al. 2013). Such products are important for climate impact assessments (Zia et al. 2016), agricultural modeling (Hansen 2005), and ecological studies (Holden et al. 2011; Fridley 2009). General circulation models (GCMs) provide useful information about larger-scale climate, but their spatial resolution (100–450 km) is too coarse to gain insight into localized responses to climate change (Ekström et al. 2015; Lafon et al. 2013). In addition, GCMs simplify climate processes through parameterization schemes, resulting in the unrealistic representation of some climate processes (Maraun et al. 2017). Consequently, output from GCMs is characterized by a nontrivial degree of

bias (Lafon et al. 2013; Cannon et al. 2020; Maraun et al. 2017). Typically, postprocessing steps such as downscaling and bias correction are applied to climate model output prior to its use in applications or other downstream models.

In the downscaling process, output generated by climate models is transformed from a coarse resolution to finer resolution. The two main types of downscaling are *dynamical* and *statistical*. In *dynamical downscaling*, a regional climate model (RCM) is forced by GCM or reanalysis data. An RCM simulates climate processes at a finer resolution than forcing data by incorporating fine-scale landscape and atmospheric processes (Ekström et al. 2015; Caldwell et al. 2009; Leung et al. 2003; Wilby et al. 2004). RCMs are computationally intensive, although they typically require less processing power than GCMs (Feser et al. 2011; Giorgi et al. 2009). *Statistical downscaling*, in contrast, involves establishing statistical relationships between coarse-scale and fine-scale climate variables, often leveraging local, observed phenomena or attributes (Wilby et al. 2004). Statistical downscaling is computationally efficient and can be applied to both precipitation and temperature (Mearns et al. 2003; Fang et al. 2015). In contrast to dynamical downscaling, a substantial amount of observational data is necessary to derive statistical relationships necessary for statistical downscaling (Wilby et al. 2004). In addition, statistical downscaling can result in a reduction in the physical coherence of climate simulations Maraun et al. (2010). Approaches for statistical downscaling include regression-based methods (Ekström et al. 2015), principal components analysis (Huth 1999; Kettle and Thompson 2004), weather classification schemes, and weather

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JAMC-D-20-0252.s1>.

Corresponding author: Maïke Holthuijzen, maïke.holthuijzen@uvm.edu

DOI: 10.1175/JAMC-D-20-0252.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

generators (Wilby et al. 2004). Recently, machine learning methods such as artificial neural networks (Schoof and Pryor 2001), deep learning (Vandal et al. 2017), and random forests (Hutengs and Vohland 2016) have been used for downscaling both temperature and precipitation variables. Downscaling is especially important for accurate representation of temperature in regions characterized by topographically varied terrain (Hanssen-Bauer et al. 2005; Holden et al. 2011).

High-resolution climate data can also be generated by applying statistical downscaling to RCM output (Haas and Pinto 2012). While this combination of dynamical and statistical downscaling is complex, it is an effective workflow for generating high-resolution climate data simulations as it combines physical and statistical relationships (Engen-Skaugen 2007; Winter et al. 2016; Han et al. 2019).

Bias correction is another postprocessing procedure that can correct the mean, variance, and higher moments of climatological variables (Lafon et al. 2013; Cannon et al. 2020). Generally, bias-correction methods can be classified into four categories: 1) linear scaling (Lenderink et al. 2007; Hay et al. 2000), 2) nonlinear scaling (Leander and Buishand 2007), 3) distribution mapping (Piani et al. 2010), and 4) empirical (distribution free) quantile mapping (Teutschbein and Seibert 2012; Cannon et al. 2015; Wood et al. 2002). The techniques differ in their ability to correct higher-order moments of simulated climatological variables. For bias-correcting temperature variables, linear scaling and empirical quantile mapping (EQM) are often used (Maurer and Duffy 2005; Hayhoe et al. 2008; Wood et al. 2004; Bennett et al. 2014; Fang et al. 2015). EQM, a sophisticated technique, can correct the mean, variance, and higher moments of temperature and precipitation variables (Fang et al. 2015; Themeßl et al. 2011). Linear scaling is a simple technique in which the difference between monthly mean observed and simulated data is added to simulated data. Despite its simplicity, it is effective for bias correcting temperature variables (Shrestha et al. 2017; Lenderink et al. 2007). Most bias-correction methods assume stationarity of model errors over time (Roberts et al. 2019), and sufficient observational data are necessary to derive robust transfer functions.

Gridded, observational climate products (e.g., Livneh, Livneh et al. 2015; Daymet, Thornton et al. 2012; and PRISM, Daly et al. 2000) are often used for bias correction due to their extensive spatial and temporal coverage. However, the interpolation algorithms used to create gridded climate products can introduce bias (Behnke et al. 2016) and additional uncertainty when used for bias correcting climate model output (Walton and Hall 2018). In particular, Behnke et al. (2016) found that in the United States, gridded observational products (including Livneh, Daymet, and PRISM) generally exhibited a negative bias for maximum daily temperature and that biases were exacerbated in topographically complex regions. Similarly, Bishop and Beier (2013) found that in the northeastern United States, PRISM data products (Daly et al. 2000) demonstrated a cold bias for mean monthly temperature that increased at higher elevations.

A valuable alternative to gridded observational data products are long-term, curated station data, such as data from the Global Historical Climate Network (NOAA 2018). Station data represent direct climatological measurements and are

available globally (Peterson and Vose 1997; Durre et al. 2010). The use of station data, rather than gridded observational products, removes uncertainty during bias correction. Station data are often used to validate the accuracy of bias-corrected climate model output but can also be effectual for bias correcting output from climate models. Methods that account for the spatial autocorrelation of climate variables can improve the accuracy of gridded products created from sparsely distributed station data (Schabenberger and Gotway 2017). For instance, Mejia et al. (2012) downscaled monthly temperature and precipitation simulations from an RCM to climate stations and bias corrected the simulated climate variables with station data, resulting in appreciable improvement in the accuracy of a hydrological model. Poggio and Gimona (2015) showed that incorporating station data in a geostatistical downscaling and bias-correction approach resulted in full-coverage, high-resolution monthly temperature and precipitation data that better captured the complex topographical features of their study area. Recently, Xu et al. (2018) constructed 1km gridded datasets of monthly temperature over a region in China using a sophisticated geostatistical model, resulting in reduced uncertainty in the resulting datasets.

Despite the advantages of station data, its use in constructing full-coverage, bias-corrected, downscaled climate data, especially at high spatial and temporal resolutions, is limited. The density and spatial distribution of climate stations are often irregular, especially in mountainous and high-elevation regions (Daly et al. 2000). Another challenge is that for constructing full-coverage climate datasets, it is not sufficient to bias correct only at station locations, as bias correction must be applied at locations where stations are not present. There is a need for methods in which station data is leveraged to create full coverage, high-resolution bias-corrected climate data.

In this study, we leverage station data to develop and compare the performance of six downscaling and bias-correction methods for constructing high-resolution (1 km), daily gridded temperature climate products. All of the six methods are specifically developed to address the challenge of creating full-coverage climate products using only station data. The methods incorporate well-established interpolation and bias-correction techniques, but the workflows of the methods are novel and unique. We apply the methods to daily RCM simulations of 2-m maximum air temperature (TMAX) over a region in the northeastern United States. The relationship between elevation and temperature (lapse rate) is important to incorporate during downscaling, so we include fine-scale elevation during the downscaling process. However, in doing so, the adjustment of temperature due to elevation is difficult to disentangle with the adjustment due to bias correction. For this reason, all six methods were implemented with and without the incorporation of fine-scale elevation. We validate the methods using two calibration time periods and apply a spatial cross validation prior to calculating performance metrics to ensure that the ability of the methods to bias correct in a spatially coherent manner is accounted for.

This paper aims to address the following questions:

- 1) How do the different bias-correction and interpolation techniques used in the six methods affect performance,

FIG. 1. GHCND stations (black points) within the study area (red-outlined box).

as measured by the root-mean square error (RMSE) and Perkins skill score (PSS)?

- 2) Does performance among methods vary by month, and is performance among methods improved when elevation lapse rates are used during downscaling?
- 3) Is any one method particularly well-suited for high-resolution downscaling and bias correction with respect to both RMSE and PSS?

The article is organized as follows: in [section 2](#), we describe the study area, station and RCM data, and downscaling and bias-correction methods. In [section 2](#), we also provide specific justifications for each of the six methods and describe validation of the methods. In [section 3](#), we present our results, and in [section 4](#) we discuss our results and provide conclusions and recommendations.

2. Methods

a. Study area and data

The study area, the Lake Champlain Basin, consists of parts of Vermont, New Hampshire, eastern New York, and southern Quebec, Canada ([Fig. 1](#)). The region is topographically varied; the Green Mountains, Adirondack Mountains, and White Mountains span portions of Vermont, New York, and New Hampshire, respectively ([Winter et al. 2016](#)).

Elevations in the study area range from 30 to 1500 m above mean sea level (MSL).

Daily historical TMAX simulations over 1980–2014 were generated by the Advanced Weather and Research Forecasting (WRF) Model, version 3.9.1 ([Skamarock et al. 2019](#)). WRF is widely used as both a regional climate model and numerical weather prediction system ([Skamarock et al. 2019](#)). Initial and lateral boundary conditions were obtained from ERA-Interim, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA-Interim has an approximate spatial resolution of 80 km ([Dee et al. 2011](#)) and was downscaled to 4 km using three one-way nests (36, 12, 4 km) ([Huang et al. 2020](#)). Only output from the inner, 4-km-resolution domain was used in this study. Specific physics settings for WRF are shown in the online supplemental material. A total of 4387 WRF grid cells covered the study area.

Historical daily weather station data were obtained from the Global Historical Climate Network-Daily (GHCND; <https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND>). GHCND data records are adjusted to account for changes in instrumentation and other anomalies ([NOAA 2018](#); [Peterson and Vose 1997](#)). We retained only those stations with at least 70% complete records over the historical time period 1980–2014 (73 stations). In this study, WRF simulations were downscaled to a 1-km grid; elevation estimates at each 1-km grid cell were derived from a 30-m digital elevation model

FIG. 2. Workflows for the six bias-correction and downscaling methods described in this study. In EQM_IDW, EQM_krig, and EQM_grid, bias correction was done with EQM. EQM_grid differs with respect to EQM_krig and EQM_IDW in that bias correction was done at the *grid* rather than station level. In LTQM_grid_V and LTQM_grid_C, LT functions were constructed using rank-ordered WRF and GHCND station data. In LTQM_grid_V, interpolated LT parameters were used for bias correction at the fine-scale grid level, so LT parameters were allowed to *vary* spatially ($V = \text{vary}$). In LTQM_grid_C, the median values of interpolated LT parameters at the fine-scale grid level were calculated and subsequently used for bias correction, so LT parameters were *constant* over the fine-scale grid ($C = \text{constant}$). Interpolated parameters were also allowed to vary spatially over the fine-scale grid for method LT_grid, but LT functions were constructed using temporally ordered, rather than rank-ordered, data.

(U.S. Geological Survey 2019). Elevation values were interpolated to the 1-km grid using inverse distance weighting (IDW). The 1-km resolution was chosen on the basis of spatial resolution requirements for local climate impact assessments (Wang et al. 2012; Winter et al. 2016).

b. Description of downscaling and bias-correction methods

The six downscaling and bias-correction methods described in this paper can be divided into two groups: those that employ EQM for bias correction and those that employ linear transfer regression functions for bias correction. Within the two groups, methods differ mainly with respect to interpolation techniques (IDW; kriging) and procedures to transfer bias correction to locations that are void of stations.

Elevation has a major effect on climatological variables such as maximum temperature (Winter et al. 2016; Barry 1992). Therefore, during downscaling, it is important to account for lapse rates, especially in topographically rich regions, such as the Lake Champlain Basin (Winter et al. 2016). However, we found that when elevation was incorporated (using lapse rates) during downscaling, it became difficult to disentangle the effects of downscaling with those of bias correction. Therefore, all methods were implemented with and without the use of lapse rates or elevation covariate (depending on the interpolation method). When elevation was not accounted for, neither

lapse rates nor an elevation covariate was included during interpolation of WRF data. In this study, we will regard steps involving the interpolation of WRF to GHCND station locations or the fine-scale grid as downscaling.

1) EQM-BASED METHODS: EQM_KRIG, EQM_IDW, AND EQM_GRID

One way that station data can be leveraged for bias correcting WRF simulations in locations where stations are not present is to 1) interpolate WRF simulations to station locations, 2) bias correct interpolated WRF simulations at station locations using EQM, and 3) interpolate bias-corrected WRF simulations at station locations to the fine-scale grid. This general workflow is implemented in EQM_krig and EQM_IDW (Fig. 2; Table 1). As the suffixes suggest, the interpolation methods for EQM_krig and EQM_IDW were kriging and IDW, respectively. Both kriging, a geostatistical procedure, and IDW, a deterministic one, are common interpolation methods for downscaling (Boer et al. 2001; Jeffrey et al. 2001; Wikle et al. 2019; Poggio and Gimona 2015; Daly 2006).

For both methods EQM_krig and EQM_IDW, daily WRF simulations were first interpolated to GHCND station locations. For EQM_IDW, interpolation was completed using IDW with and without topographic downscaling (Winter et al. 2016). IDW with topographic downscaling combines IDW with

TABLE 1. Summary of six bias-correction and downscaling methods.

Method	Bias-correction technique	Details
EQM_krig	Empirical quantile mapping	Bias correction occurs at GHCND station locations; bias-corrected WRF is then interpolated to the fine-scale grid using kriging
EQM_IDW	Empirical quantile mapping	Bias correction occurs at GHCND station locations; bias-corrected WRF is then interpolated to the fine-scale grid using IDW
EQM_grid	Empirical quantile mapping	GHCND station and WRF data are interpolated to the fine-scale grid; bias correction occurs at the fine-scale grid
LTQM_grid_V	Linear transfer function: quantile mapping using rank-ordered regression	Linear transfer functions are constructed at GHCND station locations, and estimated parameters are kriged to the fine-scale grid; kriged parameters are used for bias correction at the fine-scale grid
LTQM_grid_C	Linear transfer function: quantile mapping using rank-ordered regression	Linear transfer functions are constructed at GHCND station locations, and estimated parameters are kriged to the fine-scale grid; the medians of kriged parameters over the fine-scale grid is calculated and used for bias correction (parameters are constant over the fine-scale grid)
LT_grid	Linear transfer function: temporally ordered regression	Linear transfer functions are constructed at GHCND station locations, and estimated parameters are kriged to the fine-scale grid; kriged parameters are used for bias correction at the fine-scale grid

lapse rates to adjust for fine-scale elevation and has been applied to high-resolution downscaling (Winter et al. 2016) (full details on topographic downscaling and IDW are given in the online supplemental material). For brevity, we will refer to IDW with topographic downscaling as topographic downscaling. Two parameters, the power p and number of nearest neighbor observations used in averaging n , control the smoothness of IDW interpolation. Higher values of p and n result in progressively smoother interpolated surfaces. Based on results from Winter et al. (2016), who used a similar study area and data, as well as our own assessment, we chose values of 2 and 9 for p and n , respectively. Elevational lapse rates were calculated using historical GHCND TMAX data within the study region following methods in (Winter et al. 2016).

For EQM_krig, WRF simulations were interpolated with kriging. To account for fine-scale elevation, elevation (either at station locations or at the fine-scale grid) was included as a covariate in universal kriging models. In the case when fine-scale elevation was not accounted for, ordinary kriging was used. IDW and kriging were implemented with the gstat package (Gräler et al. 2016) in R (R Core Team 2018). The prediction surface resulting from kriging depends on the location of observational data as well as the strength of spatial dependence among the data, which can be assessed with a variogram. Based on inspection of empirical variograms of daily WRF TMAX data, all kriging models were fit with the exponential covariance function. The effective range, partial sill, and nugget were set to 150 km, 15, and 0.2, respectively (full kriging details are described in the online supplemental material). We compared the two interpolation techniques, kriging and IDW, because we wanted to determine whether a geostatistical (kriging) or deterministic (IDW) interpolation technique would significantly influence performance. Kriging methods often work better for interpolating sparsely distributed data (Varouchakis and Hristopoulos 2013; Hofstra et al. 2008), such as the GHCND station data, but IDW is simple, computationally efficient, and generally better suited for interpolated densely gridded data (Wikle et al. 2019). However,

any interpolation method that incorporates relationships between temperature data and topographic features such as elevation is likely to produce more realistic predictions of climate variables, especially in regions of varying topography (Daly 2006).

Once WRF simulations were interpolated to GHCND station locations for all days in the historical time period, WRF interpolations were bias corrected at each GHCND station location using EQM [see Eq. (1)]. The EQM transfer function is expressed by the empirical cumulative distribution function (ecdf) and its inverse (ecdf⁻¹):

$$X_{\text{corr},t} = \text{ecdf}_{\text{obs},m}^{-1}[\text{ecdf}_{\text{raw},m}(X_{\text{raw},t})]. \quad (1)$$

In Eq. (1), $X_{\text{corr},t}$ is the corrected WRF TMAX value on day t , $\text{ecdf}_{\text{obs},m}^{-1}$ is the inverse ecdf of GHCND station data for month m , $\text{ecdf}_{\text{raw},m}$ is the ecdf of interpolated WRF TMAX simulations at a GHCND station location for month m , and $X_{\text{raw},t}$ is the interpolated, uncorrected WRF TMAX at a GHCND station location on day t . Thus, daily WRF simulations in a specific month were corrected with the corresponding monthly EQM transfer function. For example, a WRF simulated value of TMAX in January would be corrected with the EQM transfer function for January. EQM was implemented with the qmap package (Gudmundsson 2016) in R. Bias-corrected WRF simulations at GHCND station locations were interpolated to the fine-scale grid with the same method used to interpolate coarse-grid WRF simulations to GHCND station locations.

Despite the simplicity of EQM_krig and EQM_IDW, much of the original WRF data are not used, as ultimately only bias-corrected WRF simulations at GHCND station locations are interpolated to the fine-scale grid. Another approach to transferring information from stations to other locations for bias correcting WRF data is to 1) interpolate both GHCND station and WRF data to the fine-scale grid and 2) bias correct WRF interpolated data with interpolated station data on a gridcell-by-gridcell basis using EQM. The method EQM_grid

(Fig. 2; Table 1) has advantages over EQM_krig and EQM_IDW, since it preserves more spatial information from WRF data (i.e., the *grid* suffix indicates that bias correction is applied at the fine-scale grid, rather than station level).

In method EQM_grid, WRF simulations and GHCND station data were interpolated to the fine-scale grid. WRF and GHCND data were interpolated with IDW and kriging, respectively. Kriging, rather than IDW, was used for GHCND station data, as it is generally better for interpolating sparsely distributed data (Varouchakis and Hristopulos 2013). The covariance function and covariance parameters were identical to those used in EQM_krig. When elevation was accounted for, interpolation of WRF simulations was done via topographic downscaling. Interpolation of GHCND station data was done with universal kriging, which included an elevation covariate. Finally, after WRF simulations and GHCND station data were interpolated to the fine-scale grid, WRF interpolations were bias corrected with kriged GHCND station data grid cell by grid cell using EQM [Eq. (1)].

2) LINEAR TRANSFER FUNCTION-BASED METHODS: QUANTILE MAPPING AND LINEAR REGRESSION (LTQM_GRID_C, LTQM_GRID_V, LT_GRID)

The linear transfer (LT) family of methods presents an alternative way to transfer information needed to bias correct WRF simulations at any location on the fine-scale grid. In methods LT_grid, LTQM_grid_V, and LTQM_grid_C, bias correction is done by applying LT functions derived from regression relationships between GHCND station data and WRF simulations (Fig. 2; Table 1). In these methods, simple regression parameters (slopes and intercepts) are estimated at GHCND station locations and interpolated to locations on the fine-scale grid where bias correction is to be performed. Thus, LT methods provide an alternative to the EQM methods (EQM_grid, EQM_krig, and EQM_IDW), as estimated parameters, rather than either bias-corrected data (EQM_krig, EQM_IDW) or GHCND station data (EQM_grid) are interpolated to the fine-scale grid and subsequently used to bias correct WRF data at the grid level.

The main difference between both LTQM methods (LTQM_grid_V and LTQM_grid_C) and LT_grid is the ordering of the data used to construct the simple regressions, which ultimately impacts the type of correction applied to WRF simulations. Two types of data ordering were considered: 1) temporally ordered (calendar order) (LT_grid) and 2) rank ordered (sorted from least to greatest) (LTQM_grid_V and LTQM_grid_C). In both cases 1 and 2, GHCND station data were expressed as a linear function of WRF data, and regression parameters (slope and intercept) were estimated via ordinary least squares (OLS). In the context of this study, resulting regression equations are applied to raw WRF data to complete the bias correction. The intercept adjusts the mean, while the slope scales the variance. Thus, since the regression equation is linear in form, the transfer function is linear.

If OLS assumptions are met, then by definition, OLS estimates are best linear unbiased estimators (BLUE; Seber and Lee 2012), and the regression line is the only such line that minimizes the mean square error. It follows that for case 1, in

which WRF and GHCND station data are temporally ordered (LT_grid), the LT function is guaranteed to improve daily discrepancies between WRF and GHCND station data (RMSE). However, the approach is not guaranteed to improve distributional discrepancies to the same degree. For case 2, in which data are rank ordered (LTQM_grid_V and LTQM_grid_C), the LT function acts as a simple type of quantile mapping and will thus improve distributional similarity (and PSS) between WRF and GHCND station data. However, RMSE is not guaranteed to improve. Since both LTQM_grid_C and LTQM_grid_V bias correct via a simple quantile-mapping technique, the “QM” in LTQM_grid_V and LTQM_grid_C refers to “quantile mapping.” The subtle difference between LTQM_grid_V and LTQM_grid_C will be discussed later.

While using rank-ordered data results in a simple form of quantile mapping, the quantile map between WRF and GHCND station data is modeled with a *linear* regression line. EQM is more flexible, as first, quantiles of observed and station data are estimated, and then the quantile map is approximated via linear or spline interpolation (Gudmundsson 2016). It is important to note that if OLS assumptions (linearity, homoscedasticity of residual errors, and independence of observations) are not met, the OLS estimates are no longer BLUE.

The first step for methods LT_grid, LTQM_grid_V, and LTQM_grid_C was identical: daily WRF simulations were interpolated to the fine-scale grid using IDW (or topographic downscaling). Daily WRF simulations were also interpolated to GHCND station locations, where LT functions were formulated [see Eq. (2)].

For all three methods (LTQM_grid_C, LTQM_grid_V, LT_grid), LT functions were constructed by regressing large-scale predictor variables (WRF data) on small-scale predictands (GHCND station data) at each GHCND station location. Separate LT functions were constructed for each month. The estimated regression parameters at each GHCND station location (slope and intercept coefficients) were kriged to the fine-scale grid, and interpolated WRF simulations on the fine-scale grid were bias corrected with the corresponding kriged regression parameters grid cell by grid cell. Therefore, the term “grid” in all three methods refers to bias correction taking place at the fine-scale grid center points, rather than station locations.

The LT function for LT_grid was a simple linear regression in which WRF interpolations at GHCND stations were predictor variables, and GHCND station data were the predictands [see Eq. (2)]. Data were sorted in temporal order. Twelve LT functions (one for each month) were constructed at each GHCND station location:

$$\text{TMAX}_{\text{station},i,m} = \beta_{0,i,m} + \beta_{1,i,m} \times \text{WRF}_{\text{IDW},i,m}. \quad (2)$$

In Eq. (2), $\text{TMAX}_{\text{station},i,m}$ is daily TMAX for GHCND station location i in month m , $\beta_{0,i,m}$ is the intercept for GHCND station location i in month m , $\beta_{1,i,m}$ is the slope for GHCND station location i in month m , and $\text{WRF}_{\text{IDW},i,m}$ represents daily interpolated WRF values at GHCND station location i in month m . Monthly parameter estimates of slopes and intercepts at each GHCND station location were kriged to the fine-scale grid with ordinary Bayesian kriging.

Empirical variograms of estimated monthly slope and intercept parameters showed some degree of spatial autocorrelation, although the association was stronger in cold-season months than in warm-season months. The exponential covariance function was used for all Bayesian kriging models. Based on the inspection of empirical variograms, we used noninformative prior distributions for the intercept β_0 , partial sill σ^2 , and nugget τ^2 , and an informative prior for the effective range ϕ :

$$\begin{aligned}\beta_0 &\sim N(0, 100), \\ \phi &\sim \text{Unif}\left(\frac{3}{D_{\max}}, \frac{3}{10}\right), \\ \sigma^2 &\sim \text{InvGamma}(2, 2), \quad \text{and} \\ \tau^2 &\sim \text{InvGamma}(2, 0.02).\end{aligned}$$

Here, $D_{\max}$ is the maximum distance between any two GHCND station locations (full details on Bayesian modeling are described in [section 1](#) of the online supplemental material). Bayesian kriging is preferable to non-Bayesian kriging when data are sparse, and there is some degree of uncertainty surrounding estimates of covariance function parameters ([Pilz and Spöck 2008](#)). Last, interpolated WRF simulations on the fine-scale grid were bias corrected grid cell by grid cell, using the corresponding kriged slope and intercept parameter estimates:

$$\text{TMAX}_{i,m}^* = \tilde{\beta}_{0,i,m} + \tilde{\beta}_{1,i,m} \times \text{WRF}_{1\text{km}-\text{interp},i,m}. \quad (3)$$

In Eq. (3), $\text{TMAX}_{i,m}^*$ is the bias-corrected, fine-scale WRF value for grid cell i in month m , $\tilde{\beta}_{0,i,m}$ is the kriged prediction for the intercept of grid cell i in month m , $\tilde{\beta}_{1,i,m}$ is the kriged slope parameter estimate at fine-scale grid cell i in month m , and $\text{WRF}_{1\text{km}-\text{interp},i,m}$ is the interpolated WRF value at the center of fine-scale grid cell i in month m .

For methods LTQM_grid_V and LTQM_grid_C, LT functions were constructed using rank-ordered WRF and GHCND station data. In these LT functions, the nearest WRF gridcell values to GHCND station locations were predictor variables and GHCND station data were predictands, similar to the approach of ([Berg et al. 2012](#)), who applied rank-ordered regression to bias correct temperature and precipitation simulations. [Berg et al. \(2012\)](#) found that modeling empirical quantiles of RCM and observed mean temperature data with a simple linear regression worked well if the quantile map between simulated and observed data was linear in form. Twelve LT functions were constructed at each GHCND station location:

$$\text{TMAX}_{i,m} = \beta_{0,i,m} + \beta_{1,i,m} \times \text{WRF}_{\text{NN}_{i,m}}. \quad (4)$$

In Eq. (4), $\text{TMAX}_{i,m}$ is daily TMAX at GHCND station location i in month k , $\beta_{0,i,m}$ is the intercept for GHCND station location i in month m , $\text{WRF}_{\text{NN}_{i,m}}$ are the one-nearest-neighbor gridcell WRF simulations relative to GHCND station location i in month m , and $\beta_{1,i,m}$ is the coefficient for station location i in month m .

There was one important difference between methods LTQM_grid_V and LTQM_grid_C. In method LTQM_grid_V,

monthly estimates of intercepts and slopes were kriged to the fine-scale grid with ordinary Bayesian kriging using the same priors as in LT_grid. Then, the kriged slopes and intercepts were used to bias correct interpolated WRF data on the fine-scale grid [Eq. (3)]. In method LTQM_grid_C, however, the monthly medians of kriged slopes and intercepts over the fine-scale grid were used to bias correct interpolated WRF data [Eq. (3)]. In LTQM_grid_V, the kriged slopes and intercepts used to bias correct WRF interpolations varied over the fine-scale grid (V for vary). In contrast to LTQM_grid_V, spatially constant (C for constant) slope and intercept values were used for bias correction in LTQM_grid_C. We implemented variations in which estimated slopes and intercepts varied spatially (LTQM_grid_V) and in which they were spatially constant (LTQM_grid_C), because monthly kriged surfaces of estimated slopes and intercepts over the fine-scale grid were not always spatially smooth. A rougher parameter surface could potentially result spatially incoherent corrections in some locations. Using constant monthly medians of kriged slope and intercept estimates alleviates issues related to a rough kriging surface but sacrifices flexibility in that any spatial dependence is no longer accounted for.

c. Performance measures

Bias-corrected WRF simulations should exhibit day-to-day, as well as distributional, correspondence to GHCND station data. Thus, we chose RMSE and PSS ([Perkins et al. 2007](#)) as performance metrics that 1) quantify daily discrepancies and 2) distributional similarity between WRF and GHCND station data, respectively. PSS ranges between 0 and 1, where 1 indicates a perfect distributional overlap between simulated and observed data, and 0 indicates no distributional overlap ([Perkins et al. 2007](#)). PSS is calculated by summing minimum densities of overlapping bins of discrete histograms of simulated and observed data. PSS is not highly influenced by outliers, but it is sensitive to bin size ([Perkins et al. 2007](#)). However, larger daily discrepancies between simulated and observed data have a comparatively larger influence on RMSE, due to the squared term in its calculation. Both PSS and RMSE metrics are widely in the climate literature for validation ([Keellings 2016](#); [Winter et al. 2016](#); [Perkins et al. 2007](#); [Poggio and Gimona 2015](#); [Bordoy and Burlando 2013](#); [Xu et al. 2018](#); [Vandal et al. 2019](#)).

To fairly assess the ability of the six methods to bias correct WRF simulations at locations where stations are not present, we implemented a fivefold spatial cross validation prior to calculating performance metrics. In each fold, 1) bias correction was based on approximately 70% of GHCND stations, and 2) bias correction was applied to WRF interpolations at the remaining 30% of GHCND station locations.

Because all of the six methods had slightly different workflows, the fivefold spatial cross validation was adjusted for each method to ensure that results were comparable.

For EQM_krig and EQM_IDW methods, the cross validation was performed as follows for each of the $i = 1, \dots, k$, $k = 5$, folds: for fold i , bias-corrected WRF interpolations at GHCND station locations in fold $k \neq i$ were used as training data and were interpolated (via kriging or IDW) to GHCND station locations in fold i .

FIG. 3. Monthly average TMAX (°C) of WRF interpolations at GHCND locations (blue) and GHCND station data (red) from 1980 to 2014 showing a distinct cold bias in the WRF simulations for months that make up winter and early spring (months 1–5 and 11–12).

For EQM_grid, TMAX values at GHCND station locations in the $k \neq i$ fold were used as training data and were interpolated using ordinary kriging to station locations in fold i . Then, interpolated WRF data at GHCND station locations in the i th were bias corrected using kriged GHCND station values. This was repeated for the $i = 1, \dots, k$, $k = 5$ folds.

For LT_grid, LTQM_grid_V, and LTQM_grid_C methods, LT functions (2 and 4) were constructed at GHCND station locations in folds $k \neq i$; Bayesian kriging was used to krig estimated LT parameters (slopes and intercepts) to GHCND station locations in fold i . Interpolated WRF values at GHCND station locations in the i th fold were bias corrected with kriged estimated LT parameters. This was repeated for the $i = 1, \dots, k$, $k = 5$ folds.

Like all cross-validation approaches, GHCND stations in each of the five folds were randomly selected prior to spatial cross validation; thus, for each method, the stations in folds $k = 1, \dots, k$ were the same to ensure that results would be comparable.

Spatially cross-validated, daily RMSE values were calculated by method and month using the following formula:

$$E_k(Y) = \sqrt{\frac{1}{n_k} \sum_{i \in \text{fold } k} \left\{ Y \left[s_j - \widehat{Y}(s_j)^2 \right] \right\}} \quad \text{and}$$

$$\text{RMSE} = \frac{1}{K} \sum_{k=1}^K E_k(Y),$$

where $Y(s_j)$ is the TMAX value at GHCND station s_j , $\widehat{Y}(s_j)^2$ is the bias-corrected WRF TMAX value at GHCND station location s_j , n_k is the number of observations in fold k , and $K = 5$.

To calculate PSS, discrete probability density functions (PDFs) were constructed for bias-corrected WRF and GHCND station data using bin widths of 0.5°C as recommended by (Perkins et al. 2007). Spatially cross-validated PSS was calculated by method and month using the following formula:

$$E_k(\text{PSS}) = \sum_i^{b_k} \min(Z_i, Z_i^*) \quad \text{and}$$

$$\text{PSS}_m = \frac{1}{K} \sum_{k=1}^K E_k(\text{PSS}),$$

where Z_i is the normalized density of the PDF of GHCND station data in bin i , Z_i^* is the normalized density of the PDF of bias-corrected WRF data in bin i , b_k is the number of bins used to construct the PDFs of GHCND station and bias-corrected WRF data in fold k , and $K = 5$.

d. Validation

We validated the six methods using two calibration time periods. Bias correction was applied to 1) 1980–2014 WRF simulations using 1980–2014 GHCND station data and 2) 1990–2014 WRF simulations using 1980–89 GHCND station data. The former approach helps evaluate performance of methods for processing historical simulations, while the latter approach assesses potential performance of methods for processing future projections. For clarity, we name these cases by referring to the subset of GHCND station data that are used for bias correction (e.g., “1980–2014” and “1980–89”).

e. Analysis of performance metrics

Performance metrics of the six methods were analyzed with two linear analysis of variance (ANOVA) models (one for RMSE and one for PSS). Our observations show that raw WRF interpolations at GHCND station locations exhibit a distinct cold bias in winter and early spring in comparison with summer and early autumn months (Fig. 3), so we controlled for monthly variation in the two ANOVA models. We also controlled for whether or not elevation was accounted for in downscaling to help disentangle the effects of downscaling with those of bias correction. Last, we controlled for the calibration time period (1980–89 or 1980–2014) used to bias correct WRF simulations. We used linear ANOVA models to evaluate performance among methods, as they are easy to interpret and provide

information on how PSS and RMSE differ among methods while controlling for variables. With the incorporation of interaction effects, linear ANOVA models can also help expand knowledge of more complex relationships among performance metrics, the six methods, and controlling variables (described below).

Prior to ANOVA model fitting, spatially cross-validated RMSE and PSS were averaged over the six methods and months. Full models for PSS and RMSE were fit with the following four fixed effects:

- Method: identifier for the downscaling and bias-correction method (EQM_krig, EQM_IDW, EQM_grid, LT_grid, LTQM_grid_V, and LTQM_grid_C),
- Month: month of the year (1–12),
- Elevation: binary variable to denote whether the effect of elevation was included with the use of elevational lapse rates (“YES”) or not (“NO”), and
- Bias_correction_years: binary variable to denote if 1990–2014 WRF simulations were bias corrected with 1980–89 GHCND station calibration dataset (“1980–89”) or whether 1980–2014 WRF simulations were bias corrected with the 1980–2014 GHCND station calibration dataset (“1980–2014”).

In addition, the initial full model fits included reasonable and informed two- and three-way interactions: Month $\times$ Method, Elevation $\times$ Method $\times$ Bias_correction_years, Elevation $\times$ Method, Elevation $\times$ Bias_correction_years, Bias_correction_years $\times$ Method, and Method $\times$ Bias_correction_years $\times$ Elevation. After full ANOVA models were fit, all variables with a p value < 0.05 were eliminated, and both ANOVA models were fit again with remaining variables. After fitting final ANOVA models, pairwise comparisons, as well as estimated marginal means (necessary for interaction plots) were calculated with the R package emmeans (Lenth 2020). Pairwise comparisons were carried out with the Bonferroni correction for multiple comparisons. We also calculated η^2 for all effects in the final models for PSS and RMSE; η^2 quantifies the proportion of variance associated with main effects and interactions in a linear model and is a useful indicator of effect size and strength of association in linear models (Levine and Hullett 2002; Muller and Peterson 1984). Values for η^2 range between 0 and 1, where higher values indicate greater variable importance; η^2 is calculated as the sum of squares of an independent variable (SS_{between}) divided by the total sum of squares (TSS) of the model:

$$\eta^2 = SS_{\text{between}} / TSS.$$

3. Results

a. Overall performance

Raw WRF interpolations at GHCND station locations exhibited a cold bias, which was most pronounced in months 11 and 12 and 1–4 (Fig. 3). Generally, the average day-to-day correspondence between GHCND station data and bias-corrected WRF data, as measured by mean RMSE, varied little among methods, ranging between 3.1 and 3.5 (Fig. 4a). Distributional similarity between GHCND station data and bias-corrected WRF data, measured by PSS, ranged between 0.94 and 0.96 (Fig. 4b). All methods performed better than

uncorrected WRF: RMSE of uncorrected WRF interpolations at GHCND station locations ranged between 3.6 and 3.9, while mean PSS ranged between 0.90 and 0.91 (Figs. 4a,b).

Performance metrics for all methods exhibited monthly variation: both mean monthly RMSE and PSS were worse in months 11, 12, and 1–4 than in months 5–10 (Figs. 5a,b), although monthly variation was much more pronounced for RMSE than PSS. Overall, methods LT_grid and LTQM_grid_V performed best and worst, respectively, in terms of mean RMSE (Fig. 4a), while methods EQM_grid and LTQM_grid_V performed best and worst, respectively in terms of mean PSS (Fig. 4b).

Mean RMSE and PSS improved when bias correction was based on 1980–2014 GHCND station data (and the correction was applied to 1980–2014 WRF data) when compared with when bias correction was based on 1980–89 GHCND station data (and the correction was applied to 1990–2014 WRF simulations) (Figs. 4a,b, respectively). Generally, when elevation was accounted for during downscaling (with lapse rates), mean RMSE decreased (Fig. 4a), but Elevation did not have an appreciable impact on mean PSS (Fig. 4b). There was no consistent relationship between low RMSE and high PSS. An example of a downscaled, bias-corrected data product for one particular day is shown in Fig. 6 (only one example is shown, as downscaled, bias-corrected data for all methods were ocularly indistinguishable). Figure 6 clearly shows that the downscaled, bias-corrected data capture the fine-scale topographical variation of TMAX over the study region and are much more realistic when compared with raw WRF.

b. Statistical analysis of error metrics

The final ANOVA model for RMSE included the main effects Month, Bias_correction_years, Elevation, and Method as well as the interactions Month $\times$ Method, Method $\times$ Bias_correction_years, and Method $\times$ Elevation. The p values for all variables in the final RMSE ANOVA model were less than 10^{-5} (Table 2; see appendix Table A1 for the full ANOVA table). The final model for PSS included the main effects Month, Method, and Bias_correction_years and the interaction terms Month $\times$ Method and Method $\times$ Bias_correction_years. The p values for all variables in the final PSS ANOVA model were less than 10^{-5} (Table 3; see appendix Table A2 for the full model ANOVA). In contrast to results for RMSE, the effect of Elevation was not significant in the full model for PSS ($p = 0.857$; see appendix Table A2).

Because of the significance of interaction as well as main effects, main effects are discussed in the context of interaction effects. Results for pairwise comparisons for interaction terms present in RMSE and PSS ANOVA models, as well as an alternative metric to PSS for quantifying distributional similarity, are shown in the online supplemental material.

1) MONTH, METHOD, AND MONTH $\times$ METHOD

(i) RMSE

The η^2 for Month was 0.94, whereas η^2 for Month $\times$ Method and Method were 0.014 and 0.0092, respectively (Table 4). The large η^2 for Month indicates that Month was the most

FIG. 4. Mean (a) RMSE (°C) and (b) PSS by Method and Bias_correction_years, where “1980–89” and “1980–2014” refer to GHCND station datasets used to bias correct 1990–2014 and 1980–2014 WRF simulations, respectively. Error bars represent standard errors over five spatial cross-validation folds. “WRF_interp” denotes the raw WRF simulations interpolated to station locations and is shown as a comparison with bias-corrected WRF data.

important variable in the model despite the statistical significance of the interaction Month $\times$ Method. Thus, RMSE varied substantially by month. Indeed, the monthly pattern of RMSE was consistent for all methods (Fig. 7a). The interaction plot for Month $\times$ Method shows that mean marginal RMSE of all methods was greater (3.2°–4.2°C) in months 1–4, 11, and 12 than in months 5–10 (2.5°–3°C) (Fig. 7a). Overall, mean marginal RMSE of LT_grid was lower than that of all other methods and was significantly lower during months 2–6, 11, and 12.

(ii) PSS

In contrast, for the PSS ANOVA model, the influence of Method ($\eta^2 = 0.43$) was greater than that of Month $\times$ Method

($\eta^2 = 0.28$) and Month ($\eta^2 = 0.11$) (Table 5). This means that PSS varied more among the six methods than among months (Fig. 7b). Mean marginal PSS for EQM_IDW, EQM_krig, and EQM_grid varied between 0.92 and 0.95, regardless of month (Fig. 7b). However, mean marginal PSS for LTQM_grid_C and LTQM_grid_V ranged between 0.88 and 0.90 in months 1–4 and then increased to between 0.94 and 0.96 in months 5–12 (Fig. 7b). Mean marginal PSS for LT_grid followed a similar pattern as LTQM_grid_V and LTQM_grid_C for months 1–4; however, in months 5–10, mean marginal PSS for LT_grid was significantly lower than that of all other methods for every month, ranging between 0.90 and 0.91.

FIG. 5. Mean (a) RMSE ($^{\circ}\text{C}$) and (b) PSS by Method, Month, and Bias_correction_years, where “1980–89” and “1980–2014” refer to GHCND station datasets used to bias correct 1990–2014 and 1980–2014 WRF simulations, respectively. Error bars represent standard errors over five spatial cross-validation folds. “WRF_interp” denotes raw WRF simulations interpolated to station locations and are shown to indicate relative improvement of all methods over raw WRF interpolated values.

2) BIAS_CORRECTION_YEARS AND BIAS_CORRECTION_YEARS $\times$ METHOD

(i) RMSE

In the RMSE ANOVA model, η^2 values for Bias_correction_years $\times$ Method and Bias_correction_years ($\eta^2 = 0.0018$ and 0.0063 , respectively) indicate that the main effect of Bias_correction_years was slightly more important than Bias_correction_years $\times$ Method. Overall, mean marginal RMSE slightly improved when bias correction was based on 1980–2014 GHCND station data (3.15° – 3.25°C) when compared with the 1980–89 GHCND subset (3.18° and 3.57°C), although

there were slight differences among methods (Fig. 8a). The improvement was more pronounced for the EQM-based methods than for the LT-based methods. Mean marginal RMSE of LT_grid was significantly lower (3.18°C) than that of all other methods (3.3° – 3.56°C) when bias correction was based on the 1980–89 GHCND dataset (Fig. 8a). When bias correction was done with the 1980–2014 GHCND dataset, mean marginal RMSE of LT_grid was lowest overall, but it did not differ significantly from that of EQM_grid. Mean marginal RMSE of LTQM_grid_V was significantly greater than that of all other methods. It is important to note that η^2 values for Bias_correction_years and Bias_correction_years $\times$ Method

FIG. 6. (left) Original WRF simulations for TMAX (°C) and (right) downscaled WRF TMAX (°C) using method EQM_IDW for 5 Aug 1982.

were much smaller than η^2 of Month, which means that Month was relatively more important than Bias_correction_years and Bias_correction_years $\times$ Method.

(ii) PSS

Mean PSS generally increased when the 1980–2014, as compared with the 1980–89 GHCND dataset, was used for bias correction, but the amount of increase varied among methods. Use of the 1980–2014 GHCND dataset for the bias-correction set resulted in a consistent improvement in PSS for all EQM-based methods. The interaction Bias_correction_years $\times$ Method was particularly evident for methods LTQM_grid_C and LTQM_grid_V. Mean marginal PSS for both LTQM_grid_C and LTQM_grid_V were significantly higher than that of all other methods when the 1980–1989 GHCND dataset was used for bias correction (Fig. 8b). However, when the 1980–2014 GHCND dataset was used for bias correction, mean marginal PSS of LTQM_grid_C and LTQM_grid_V fell significantly below that of EQM-based methods. In contrast to results for RMSE, LT_grid performed worst overall. Mean marginal PSS of LT_grid was significantly lower than that of all other methods, regardless of which GHCND dataset was used for bias correction (Fig. 8b). The main effect Bias_correction_years and interaction Bias_correction_years $\times$ Method ($\eta^2 = 0.09$ and 0.14 , respectively) were comparatively

less influential in the model than the effects Method and Month $\times$ Method ($\eta^2 = 0.43$ and 0.28 , respectively) (Table 5).

3) ELEVATION

(i) RMSE

In general, RMSE improved when fine-scale elevation was accounted for during the downscaling process (Fig. 9a). The η^2 for Elevation was nearly 19 times that of Elevation $\times$ Method ($\eta^2 = 0.013$ and 0.00069 , respectively; Table 4), indicating that the main effect of Elevation was more important in the RMSE ANOVA model than the interaction term. The relative importance of Elevation in the model was similar to the interaction Month $\times$ Method ($\eta^2 = 0.014$). Mean marginal RMSE of LT_grid was significantly less than that of all other methods, and mean marginal RMSE of LTQM_grid_V was significantly greater than that of all other methods, regardless of whether elevation was accounted for.

(ii) PSS

The effect of Elevation was not significant in the full model for PSS (Table A2), and Elevation did not have any appreciable effect on PSS (Fig. 9b).

TABLE 2. Summary table for the final RMSE ANOVA model.

	Degrees of freedom	Sum squares	Mean square	F value	Pr(>F)
Month	11	84.43	7.68	1485.08	0.0000
Bias_correction_years	1	0.57	0.57	109.68	0.0000
Elevation	1	1.14	1.14	219.73	0.0000
Method	5	0.82	0.16	31.74	0.0000
Month $\times$ Method	55	1.28	0.02	4.51	0.0000
Method $\times$ Bias_correction_years	5	0.16	0.03	6.18	0.0000
Method $\times$ Elevation	5	0.06	0.01	2.40	0.0388
Residuals	204	1.05	0.01		

TABLE 3. Summary table for the final PSS ANOVA model.

	Degrees of freedom	Sum squares	Mean square	<i>F</i> value	Pr(> <i>F</i>)
Month	11	0.02	0.00	18.58	0.0000
Method	5	0.05	0.01	95.13	0.0000
Bias_correction_years	1	0.03	0.03	253.94	0.0000
Month $\times$ Method	55	0.06	0.00	9.48	0.0000
Method $\times$ Bias_correction_years	5	0.02	0.00	34.42	0.0000
Residuals	210	0.02	0.00		

4. Discussion

In this study, we developed six novel strategies for constructing high-resolution, bias-corrected gridded climate products of daily historical TMAX simulations from a regional climate model, where all bias correction was based solely on station data. The six methods we present result in a substantial improvement over raw WRF simulations, are straightforward to implement, and can be applied to historical simulations as well as future projections of temperature variables. We found that no one method could concomitantly minimize RMSE and maximize PSS, which highlights the difficulty in correcting both the overall distributional discrepancies as well day-to-day discrepancies between simulated and observed data. Although performances were similar, the methods differ considerably in their ability to correct overall distributional discrepancies and day-to-day discrepancies between simulated and observed data. This is due mainly to the type of bias correction, rather than the spatial interpolation technique implemented in the methods. Distributional similarity, as measured by PSS, is achieved by matching the quantile-mapping techniques such as EQM. However, enhancing day-to-day correspondence, as measured by RMSE, is accomplished most effectively through an LT function obtained by temporally ordered linear regression between simulated and observed data. The most effective method thus depends on what is deemed most important in a particular application, day-to-day correspondence or distributional similarity, of simulated and observed data. Performance is further affected by seasonal bias of raw WRF simulations, the calibration period used for bias correction, and the inclusion of elevation information during the downscaling step.

a. Monthly variation in performance

Our results show that performance of the six methods is affected by seasonal bias of WRF simulations. Raw WRF simulations for TMAX exhibited a distinct cold bias in winter and early spring and a slight warm bias in summer relative to GHCND station data. Huang et al. (2020) compared TMAX simulations resulting from several parameterizations of WRF with Daymet gridded TMAX data and found that all parameterizations resulted in an annual cold bias during the cold season (months 11–5). The underestimation of mean temperature between months 11 and 5 is mainly due to the radiation scheme (Huang et al. 2020), and the best parameterization is reflected in the WRF data used for this study. Meng et al. (2018)

found that WRF run with a similar radiation scheme as used in this study overestimated the albedo of snow over the Tibetan Plateau, which is consistent with the winter cold bias in WRF simulations over the Lake Champlain Basin. As a result of the cold bias, performance in terms of both RMSE and to a lesser degree, PSS declined in cold-season months. Ideally, bias-correction techniques should improve both RMSE and PSS despite pronounced seasonal bias in raw WRF data.

1) PSS

Although raw WRF simulations exhibited considerable seasonal bias, EQM-based methods consistently improved PSS. However, for methods LT_grid_C and LT_grid_V, in which bias correction was done with a simple quantile-mapping LT function, monthly PSS varied substantially throughout the year and, relative to PSS of EQM-based methods, was especially low in months 1–4. Rank-ordered regression (LTQM_grid_C and LTQM_grid_V) is a quantile-mapping technique similar in functionality to EQM. However, the resulting LT function (an estimated regression line) has limitations when the quantile–quantile map of simulated and observed data is nonlinear, as in our study. In cold-season months, the quantile–quantile map of WRF and GHCND station data was nonlinear, especially in the tails, while during warmer months, it was relatively linear (Fig. 10). The nonlinear quantile–quantile map of cold-season months contributed to the poor performance of LTQM_grid_C and LTQM_grid_V in improving PSS. This result supports findings by (Berg et al. 2012), who also found that a rank-ordered regression approach could not adequately correct seasonal biases of simulated data when quantile–quantile maps between observed and simulated data were nonlinear.

However, nonlinear quantile–quantile maps can be successfully approximated in most implementations of EQM through linear interpolation, splines, or other smoothing techniques (Gudmundsson 2016). Accordingly, we found

TABLE 4. The η^2 for the final RMSE ANOVA model.

Predictor	η^2
Month	0.94
Month $\times$ Method	0.014
Elevation	0.013
Method	0.0092
Bias_correction_years	0.0063
Bias_correction_years $\times$ Method	0.0018
Elevation $\times$ Method	0.00069

FIG. 7. Interaction plots for Method $\times$ Month showing estimated mean marginal (a) RMSE and (b) PSS for each month. Error bars represent 95% confidence intervals.

that EQM-based methods substantially outperformed rank-ordered regression-based methods, LTQM_grid_C and LTQM_grid_V, in months ~11–4. Ranked-order regression, in contrast to EQM, can only correct the first and second moments of a distribution, also contributing to the poor performance of LTQM_grid_C and LTQM_grid_V in months 11–4.

We did not find appreciable differences in the performance between the two simple quantile-mapping-based methods, LTQM_grid_C and LTQM_grid_V, despite the fact that LT parameter estimates varied over the fine-scale grid in LTQM_grid_V, whereas in LTQM_grid_C, LT parameter estimates were spatially constant. Although inspection of empirical

variograms of estimated regression parameters suggested some degree of spatial autocorrelation, the amount of autocorrelation was likely not sufficient to produce a measurable difference in results. In our study, kriging LT parameter estimates

TABLE 5. The η^2 for the final PSS ANOVA model.

Predictor	η^2
Method	0.43
Month $\times$ Method	0.28
Bias_correction_years	0.14
Month	0.11
Bias_correction_years $\times$ Method	0.09

FIG. 8. Interaction plots for Method $\times$ Bias_correction_years (“1980–89” and “1980–2014” refer to GHCND station datasets used to bias correct 1990–2014 and 1980–2014 WRF simulations, respectively). Plots show estimated mean marginal (a) RMSE and (b) PSS for the 1980–89 and 1980–2014 calibration time periods. Error bars represent 95% confidence intervals.

over the fine-scale grid had no noticeable advantage over the use of spatially constant LT parameter estimates.

2) RMSE

The bias-correction techniques implemented in the six methods generally reduced RMSE in comparison with raw WRF simulations, but in contrast to PSS, bias-corrected WRF data RMSE from all methods exhibited a strong seasonal pattern. Quantile-mapping techniques (e.g., EQM and rank-ordered regression) perform well in correcting distributional discrepancies between

simulated and observed data. However, the temporal ordering of data is not preserved in quantile-mapping techniques, which has a negative impact on the improvement of day-to-day discrepancies (RMSE). This has also been noted in other studies. For example, Goodess et al. (2007) found that quantile-mapping bias-correction techniques resulted in the greatest rank-ordered correlation (highest PSS) between simulated and observed data but did not appreciably improve RMSE.

Bias-correction techniques that preserve the temporal order of data, such as that implemented in LT_grid,

FIG. 9. Interaction plots showing estimated mean marginal (a) RMSE and (b) PSS for Method $\times$ Elevation. Error bars represent 95% confidence intervals.

outperformed quantile-mapping techniques in reducing day-to-day discrepancies. However, even after correction with a temporally ordered regression, seasonal patterns in RMSE persisted. In LT_grid, the transfer function was derived by fitting a (temporally ordered) linear regression. Because of the strong cold bias in raw WRF data in cold-season months, the daily discrepancy between WRF and GHCND station data was, on average, larger during cold-season months than during warm-season months. Therefore, regression lines fit in cold-season months were characterized by a relatively large residual standard error. Consequently, the resulting LT function was not as effective in reducing RMSE, and seasonal patterns in RMSE persisted after bias correction. If raw simulated data exhibit pronounced seasonal bias, substantial improvements in RMSE are difficult to achieve. Of all the bias-correction techniques evaluated in this study, temporally ordered linear regression (LT grid), was most effective at

reducing day-to-day discrepancies (RMSE) between simulated and observed data.

b. Implications for future and historical downscaling

To gain further insight into the suitability of methods for future or historical bias correction, we used 1) 1980–89 GHCND station data to bias correct 1990–2014 WRF simulations and 2) 1980–2014 GHCND station data to bias correct 1980–2014 WRF simulations. The 1980–2014 WRF simulations bias corrected with 1980–2014 GHCND station data generally exhibited improved PSS and RMSE relative to 1990–2014 WRF simulations that were bias corrected with the 1980–89 GHCND subset.

With respect to distributional similarity (PSS), EQM outperformed rank-ordered regression when the 1980–2014 GHCND dataset was used for bias correction, while the converse was true when the 1980–89 GHCND dataset was

FIG. 10. Quantile–quantile maps of GHCND station and WRF TMAX at a sample GHCND station location for (a) September and (b) February; 100 estimated quantiles derived from 1980–2014 data are shown.

used for bias correction. The bias–variance trade-off, a well-known concept in statistical learning (Friedman et al. 2001), can help to explain this result. Simple statistical models with few parameters, such as temporally and rank-ordered linear regression, have high bias but low variance. Highly flexible techniques, that require estimation of more parameters, such as EQM, have low bias but high variance (Friedman et al. 2001). Highly flexible models result in low training errors, as they can fit training data very well. However, they are also more prone to overfitting on training data, making them less able to generalize to new data (Friedman et al. 2001). The EQM transfer function will result in a nearly perfect quantile mapping if observational and simulated data of the same time period are used in its construction. When the quantile-mapping transfer function is subsequently used to bias correct simulated data from the same time period, the correction will adjust simulated quantiles to very closely match those of observed quantiles. This explains why EQM performed very well when bias correction was based on the 1980–2014 GHCND station dataset and the correction was also applied to 1980–2014 WRF data.

However, when bias correction was based on the 1980–89 GHCND dataset, and the correction was applied to 1990–2014 WRF data, EQM-based methods did not perform as well in terms of correcting distributional similarity (PSS). This indicates some degree of overfitting and lack of robustness of EQM for bias correcting future projections. Our results agree with Piani et al. (2010) and Berg et al. (2012), who found that as the calibration time period increased, so did the risk of overfitting EQM transfer functions. Bias correction via rank-ordered regression (implemented in LTQM_grid_V and LTQM_grid_C) resulted in a higher PSS relative to EQM when the 1980–89 GHCND dataset was used to correct 1990–2014 WRF simulations. Rank-ordered regression is less prone to overfitting, which means the resulting transfer function is better able to generalize to bias correcting future projections.

In the context of this study, relatively simple bias-correction techniques (rank-ordered regression) may be better suited to bias correct future projections. Flexible techniques such as EQM work well when the transfer function is applied to simulated data from the same time period that was used to construct it. Our results support studies in which simple methods such as linear scaling (Shrestha et al. 2017), rank-ordered regression (Berg et al. 2012), and multiple linear regression (Fowler et al. 2007) performed as well as more sophisticated techniques for bias correcting future temperature projections.

In terms of improving day-to-day correspondence, temporally ordered regression, LT_grid outperformed all other methods when the 1980–89 calibration set was used for bias correction. When bias correction was based on the 1980–2014 calibration dataset, both LT_grid and EQM_grid performed significantly better than all other methods, regardless of which GHCND dataset was used for bias correction.

c. Elevation

Elevation had comparatively little impact on performance of the methods relative to monthly variation. The distribution of elevation among GHCND stations in our study area was not particularly representative of the elevation range and topography study area. Elevations in the Lake Champlain Basin range from 30 to 1500 m, whereas the majority of GHCND station elevations ranged from 30 and 400 m, and only one station had an elevation greater than 1000 m. This imbalance is unfortunate but not uncommon in climate studies (Daly et al. 2000; Winter et al. 2016; Holden et al. 2011; Roberts et al. 2019). Including fine-scale elevation during downscaling is generally recommended, but likely due to the lack of higher-elevation stations, its corrective effect on temperature was not as pronounced. Stahl et al. (2006) also found that when interpolating daily minimum and maximum temperatures from climate stations over a topographically complex region, interpolations were most accurate when station elevations

TABLE A1. ANOVA table for the full RMSE model.

	Degrees of freedom	Sum squares	Mean square	<i>F</i> value	Pr(> <i>F</i>)
Month	11	84.43	7.68	1526.98	0.0000
Method	5	0.82	0.16	32.64	0.0000
Bias_correction_years	1	0.57	0.57	112.77	0.0000
Elevation	1	1.14	1.14	225.93	0.0000
Month × Method	55	1.28	0.02	4.63	0.0000
Method × Bias_correction_years	5	0.16	0.03	6.35	0.0000
Method × Elevation	5	0.06	0.01	2.46	0.0343
Bias_correction_years × Elevation	1	0.02	0.02	3.71	0.0555
Method × Bias_correction_years × Elevation	5	0.04	0.01	1.61	0.1593
Residuals	198	1.00	0.01		

accurately represented the topography of the study region. Including fine-scale elevation data in the downscaling process did have minor positive influence on performance of the methods. It resulted in improved day-to-day discrepancies (RMSE), but it had no appreciable impact on distributional similarity (PSS). It could be that quantile-mapping bias-correction techniques had a greater positive influence on PSS than adjustments due to lapse rates.

5. Conclusions

The six high-resolution downscaling and bias-correction methods we presented in this study are straightforward and easy to implement, and depending on the method, result in substantially improved RMSE and PSS when compared with uncorrected WRF simulations. Although we applied these methods to historical (1980–2014) daily maximum temperature simulations, methods could be applied to future climate projections and any type of temperature variable (minimum, maximum, or average). Although the ranges of performance metrics among methods were narrow, there were statistically significant differences in performance among methods, and we found that performance variation was mainly due to differences in bias-correction techniques. The selection of the most appropriate method for constructing high-resolution, bias-corrected temperature products using station data depends primarily on the intended use of the resulting data product.

We did not find that one method could simultaneously minimize RMSE and maximize PSS. Maximizing PSS is achieved by matching the quantiles of simulated and observed data (EQM), which, in turn increases distributional similarity between simulated and observed data. However, minimizing RMSE is achieved by decreasing the discrepancy between daily modeled and observed data, which is done most effectively via a linear regression between simulated and observed data. Thus, bias-correction methods such as EQM improve PSS but not necessarily RMSE, whereas simple linear regression transfer functions improve RMSE but generally not PSS. While the objectives of minimizing RMSE and maximizing PSS are not mutually exclusive, they may be difficult to attain concomitantly in practice.

From our results, for processing historical simulations, the most effective method for improving day-to-day correspondence (RMSE) of simulated and observed data is LT_grid.

In this method, simulations are bias corrected at the fine-scale grid by transfer functions obtained from temporally ordered regressions generated at station locations. In topographically varied regions performance is further enhanced by accounting for fine-scale elevation data in the downscaling process. Unlike rank-ordered regression, bias correction with temporally ordered regression is more sensitive to seasonal bias in simulated data.

To achieve optimal distributional similarity (PSS) of simulated and observed data in historical downscaling, quantile mapping-based methods (EQM_IDW, EQM_krig and EQM_grid) are most effective. In EQM_IDW and EQM_krig, WRF simulations are bias corrected at station locations and then interpolated to the fine-scale grid. In EQM_grid, bias correction of interpolated WRF simulations occurs at the fine-scale grid, but this slight difference in workflow did not result in appreciable differences in performance relative to EQM_IDW and EQM_krig. The interpolation techniques (IDW, kriging) did not affect performance of EQM-based methods, which suggests that in creating full-coverage temperature products, deterministic interpolation techniques (IDW) perform as well as geostatistical techniques (kriging). Moreover, performance of EQM-based methods did not benefit from the inclusion of fine-scale elevation data during downscaling. For historical downscaling, EQM-based methods are generally more resistant to seasonal bias in simulated data. However, EQM is susceptible to overfitting on calibration data and may not provide robust bias correction for future projections.

Quantile-mapping methods in which LT functions are constructed through rank-ordered regression (LTQM_grid_V and LTQM_grid_C) are less prone to overfitting on calibration data. Therefore, such methods are better suited improving distributional similarity (PSS) of future temperature projections than EQM-based methods. In LTQM_grid_V and LTQM_grid_C, simulated data are bias corrected at the fine-scale grid through transfer functions obtained from rank-ordered regressions generated at station locations. Similar to EQM-based methods, including fine-scale elevation data during downscaling does not significantly improve distributional similarity (PSS). We did not find differences in performance between LTQM_grid_V and LTQM_grid_C. As LTQM_grid_V accounts for spatial autocorrelation among LT parameter estimates at station locations, it would likely perform better if LT parameter estimates at station locations exhibit strong spatial autocorrelation.

TABLE A2. ANOVA table for the full PSS model.

	Degrees of freedom	Sum squares	Mean square	<i>F</i> value	Pr(> <i>F</i>)
Month	11	0.02	0.00	17.88	0.0000
Method	5	0.05	0.01	91.55	0.0000
Bias_correction_years	1	0.03	0.03	244.40	0.0000
Elevation	1	0.00	0.00	0.03	0.8579
Month × Method	55	0.06	0.00	9.13	0.0000
Method × Bias_correction_years	5	0.02	0.00	33.13	0.0000
Method × Elevation	5	0.00	0.00	0.74	0.5910
Bias_correction_years × Elevation	1	0.00	0.00	0.24	0.6266
Method × Bias_correction_years × Elevation	5	0.00	0.00	0.02	0.9997
Residuals	198	0.02	0.00		

Acknowledgments. This material is based upon work supported by the National Science Foundation under Vermont EPSCoR Grant NSF OIA 1556770.

Data availability statement. Code will be available in a public Github repository. Please send an e-mail message to the corresponding author for access to the WRF data that were used in this study.

APPENDIX

Summary Tables for Full RMSE and PSS ANOVA Models

Tables A1 and A2 show the ANOVA tables for the full RMSE and PSS models, respectively.

REFERENCES

- Barry, R. G., 1992: *Mountain Weather and Climate*. Routledge, 402 pp.
- Behnke, R., S. Vavrus, A. Allstadt, T. Albright, W. E. Thogmartin, and V. C. Radeloff, 2016: Evaluation of downscaled, gridded climate data for the conterminous United States. *Ecol. Appl.*, **26**, 1338–1351, <https://doi.org/10.1002/15-1061>.
- Bennett, J. C., M. R. Grose, S. P. Corney, C. J. White, G. K. Holz, J. J. Katzfey, D. A. Post, and N. L. Bindoff, 2014: Performance of an empirical bias-correction of a high-resolution climate dataset. *Int. J. Climatol.*, **34**, 2189–2204, <https://doi.org/10.1002/joc.3830>.
- Berg, P., H. Feldmann, and H.-J. Panitz, 2012: Bias correction of high resolution regional climate model data. *J. Hydrol.*, **448–449**, 80–92, <https://doi.org/10.1016/j.jhydrol.2012.04.026>.
- Bishop, D. A., and C. M. Beier, 2013: Assessing uncertainty in high-resolution spatial climate data across the US Northeast. *PLOS ONE*, **8**, e70260, <https://doi.org/10.1371/JOURNAL.PONE.0070260>.
- Boer, E. P., K. M. de Beurs, and A. D. Hartkamp, 2001: Kriging and thin plate splines for mapping climate variables. *Int. J. Appl. Earth Obs. Geoinf.*, **3**, 146–154, [https://doi.org/10.1016/S0303-2434\(01\)85006-6](https://doi.org/10.1016/S0303-2434(01)85006-6).
- Bordoy, R., and P. Burlando, 2013: Bias correction of regional climate model simulations in a region of complex orography. *J. Appl. Meteor. Climatol.*, **52**, 82–101, <https://doi.org/10.1175/JAMC-D-11-0149.1>.
- Caldwell, P., H.-N. S. Chin, D. C. Bader, and G. Bala, 2009: Evaluation of a WRF dynamical downscaling simulation over California. *Climatic Change*, **95**, 499–521, <https://doi.org/10.1007/s10584-009-9583-5>.
- Cannon, A. J., S. R. Sobie, and T. Q. Murdock, 2015: Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *J. Climate*, **28**, 6938–6959, <https://doi.org/10.1175/JCLI-D-14-00754.1>.
- , C. Piani, and S. Sippel, 2020: Bias correction of climate model output for impact models. *Climate Extremes and Their Implications for Impact and Risk Assessment*, Elsevier, 77–104.
- Daly, C., 2006: Guidelines for assessing the suitability of spatial climate data sets. *Int. J. Climatol.*, **26**, 707–721, <https://doi.org/10.1002/joc.1322>.
- , G. Taylor, W. Gibson, T. Parzybok, G. Johnson, and P. Pasteris, 2000: High-quality spatial climate data sets for the United States and beyond. *Trans. ASAE*, **43**, 1957–1962, <https://doi.org/10.13031/2013.3101>.
- 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>.
- Durre, I., M. J. Menne, B. E. Gleason, T. G. Houston, and R. S. Vose, 2010: Comprehensive automated quality assurance of daily surface observations. *J. Appl. Meteor. Climatol.*, **49**, 1615–1633, <https://doi.org/10.1175/2010JAMC2375.1>.
- Ekström, M., M. R. Grose, and P. H. Whetton, 2015: An appraisal of downscaling methods used in climate change research. *Climatic Change*, **6**, 301–319.
- Engen-Skaugen, T., 2007: Refinement of dynamically downscaled precipitation and temperature scenarios. *Climatic Change*, **84**, 365–382, <https://doi.org/10.1007/s10584-007-9251-6>.
- Fang, G., J. Yang, Y. Chen, and C. Zammit, 2015: Comparing bias correction methods in downscaling meteorological variables for a hydrologic impact study in an arid area in China. *Hydrol. Earth Syst. Sci.*, **19**, 2547–2559, <https://doi.org/10.5194/hess-19-2547-2015>.
- Feser, F., B. Rockel, H. von Storch, J. Winterfeldt, and M. Zahn, 2011: Regional climate models add value to global model data: A review and selected examples. *Bull. Amer. Meteor. Soc.*, **92**, 1181–1192, <https://doi.org/10.1175/2011BAMS3061.1>.
- Flint, L. E., and A. L. Flint, 2012: Downscaling future climate scenarios to fine scales for hydrologic and ecological modeling and analysis. *Ecol. Process.*, **1**, 2, <https://doi.org/10.1186/2192-1709-1-2>.
- Fowler, H. J., S. Blenkinsop, and C. Tebaldi, 2007: Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.*, **27**, 1547–1578, <https://doi.org/10.1002/joc.1556>.
- Franklin, J., F. W. Davis, M. Ikegami, A. D. Syphard, L. E. Flint, A. L. Flint, and L. Hannah, 2013: Modeling plant species distributions under future climates: How fine scale do climate

- projections need to be? *Global Change Biol.*, **19**, 473–483, <https://doi.org/10.1111/gcb.12051>.
- Fridley, J. D., 2009: Downscaling climate over complex terrain: High finescale (less than 1000 m) spatial variation of near-ground temperatures in a montane forested landscape (Great Smoky Mountains). *J. Appl. Meteor. Climatol.*, **48**, 1033–1049, <https://doi.org/10.1175/2008JAMC2084.1>.
- Friedman, J., T. Hastie, and R. Tibshirani, 2001: *The Elements of Statistical Learning*. 2nd ed. Springer Series in Statistics, Springer, 745 pp.
- Giorgi, F., and Coauthors, 2009: Addressing climate information needs at the regional level: The CORDEX framework. *WMO Bull.*, **58**, 175–183.
- Goodess, C., and Coauthors, 2007: An intercomparison of statistical downscaling methods for Europe and European regions—Assessing their performance with respect to extreme temperature and precipitation events. University of East Anglia Climatic Research Unit Research Publ. 11, 68 pp.
- Gräler, B., E. Pebesma, and G. Heuvelink, 2016: Spatio-temporal interpolation using gstat. *R J.*, **8**, 204–218, <https://doi.org/10.32614/RJ-2016-014>.
- Gudmundsson, L., 2016: qmap: Statistical transformations for post-processing climate model output, version 1.0-4. R package, <https://rdrr.io/cran/qmap/>.
- Haas, R., and J. G. Pinto, 2012: A combined statistical and dynamical approach for downscaling large-scale footprints of European windstorms. *Geophys. Res. Lett.*, **39**, L23804, <https://doi.org/10.1029/2012GL054014>.
- Han, Z., Y. Shi, J. Wu, Y. Xu, and B. Zhou, 2019: Combined dynamical and statistical downscaling for high-resolution projections of multiple climate variables in the Beijing–Tianjin–Hebei region of China. *J. Appl. Meteor. Climatol.*, **58**, 2387–2403, <https://doi.org/10.1175/JAMC-D-19-0050.1>.
- Hansen, J. W., 2005: Integrating seasonal climate prediction and agricultural models for insights into agricultural practice. *Philos. Trans. Roy. Soc.*, **360B**, 2037–2047, <https://doi.org/10.1098/rstb.2005.1747>.
- Hanssen-Bauer, I., C. Achberger, R. Benestad, D. Chen, and E. Førland, 2005: Statistical downscaling of climate scenarios over Scandinavia. *Climate Res.*, **29**, 255–268, <https://doi.org/10.3354/cr029255>.
- Hay, L. E., R. L. Wilby, and G. H. Leavesley, 2000: A comparison of delta change and downscaled GCM scenarios for three mountainous basins in the United States. *J. Amer. Water. Resour. Assoc.*, **36**, 387–397, <https://doi.org/10.1111/j.1752-1688.2000.tb04276.x>.
- Hayhoe, K., and Coauthors, 2008: Regional climate change projections for the northeast USA. *Mitig. Adapt. Strategies Global Change*, **13**, 425–436, <https://doi.org/10.1007/s11027-007-9133-2>.
- Hofstra, N., M. Haylock, M. New, P. Jones, and C. Frei, 2008: Comparison of six methods for the interpolation of daily, European climate data. *J. Geophys. Res.*, **113**, D21110, <https://doi.org/10.1029/2008JD010100>.
- Holden, Z. A., J. T. Abatzoglou, C. H. Luce, and L. S. Baggett, 2011: Empirical downscaling of daily minimum air temperature at very fine resolutions in complex terrain. *Agric. For. Meteorol.*, **151**, 1066–1073, <https://doi.org/10.1016/j.agrformet.2011.03.011>.
- Huang, H., J. M. Winter, E. C. Osterberg, J. Hanrahan, C. L. Bruyère, P. Clemins, and B. Beckage, 2020: Simulating precipitation and temperature in the Lake Champlain Basin using a regional climate model: Limitations and uncertainties. *Climate Dyn.*, **54**, 69–84, <https://doi.org/10.1007/s00382-019-04987-8>.
- Hutengs, C., and M. Vohland, 2016: Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sens. Environ.*, **178**, 127–141, <https://doi.org/10.1016/j.rse.2016.03.006>.
- Huth, R., 1999: Statistical downscaling in central Europe: Evaluation of methods and potential predictors. *Climate Res.*, **13**, 91–101, <https://doi.org/10.3354/cr013091>.
- Jeffrey, S. J., J. O. Carter, K. B. Moodie, and A. R. Beswick, 2001: Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environ. Modell. Software*, **16**, 309–330, [https://doi.org/10.1016/S1364-8152\(01\)00008-1](https://doi.org/10.1016/S1364-8152(01)00008-1).
- Keellings, D., 2016: Evaluation of downscaled CMIP5 model skill in simulating daily maximum temperature over the south-eastern United States. *Int. J. Climatol.*, **36**, 4172–4180, <https://doi.org/10.1002/joc.4612>.
- Kettle, H., and R. Thompson, 2004: Statistical downscaling in European mountains: Verification of reconstructed air temperature. *Climate Res.*, **26**, 97–112, <https://doi.org/10.3354/cr026097>.
- Lafon, T., S. Dadson, G. Buys, and C. Prudhomme, 2013: Bias correction of daily precipitation simulated by a regional climate model: A comparison of methods. *Int. J. Climatol.*, **33**, 1367–1381, <https://doi.org/10.1002/joc.3518>.
- Leander, R., and T. A. Buishand, 2007: Resampling of regional climate model output for the simulation of extreme river flows. *J. Hydrol.*, **332**, 487–496, <https://doi.org/10.1016/j.jhydrol.2006.08.006>.
- Lenderink, G., A. Buishand, and W. Deursen, 2007: Estimates of future discharges of the river Rhine using two scenario methodologies: Direct versus delta approach. *Hydrol. Earth Syst. Sci.*, **11**, 1145–1159, <https://doi.org/10.5194/hess-11-1145-2007>.
- Lenth, R., 2020: emmeans: Estimated marginal means, aka least-squares means, version 1.4.4. R package, <https://CRAN.R-project.org/package=emmeans>.
- Leung, L. R., L. O. Mearns, F. Giorgi, and R. L. Wilby, 2003: Regional climate research: Needs and opportunities. *Bull. Amer. Meteor. Soc.*, **84**, 89–95, <https://doi.org/10.1175/BAMS-84-1-89>.
- Levine, T. R., and C. R. Hullett, 2002: Eta squared, partial eta squared, and misreporting of effect size in communication research. *Hum. Commun. Res.*, **28**, 612–625, <https://doi.org/10.1111/j.1468-2958.2002.tb00828.x>.
- Livneh, B., T. J. Bohn, D. W. Pierce, F. Munoz-Arriola, B. Nijssen, R. Vose, D. R. Cayan, and L. Brekke, 2015: A spatially comprehensive, hydrometeorological data set for Mexico, the U.S., and southern Canada 1950–2013. *Sci. Data*, **2**, 150042, <https://doi.org/10.1038/sdata.2015.42>.
- Maraun, D., and Coauthors, 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Global Change Biol.*, **48**, RG3003, <https://doi.org/10.1029/2009RG000314>.
- , and Coauthors, 2017: Towards process-informed bias correction of climate change simulations. *Nat. Climate Change*, **7**, 764–773, <https://doi.org/10.1038/nclimate3418>.
- Maurer, E. P., and P. B. Duffy, 2005: Uncertainty in projections of streamflow changes due to climate change in California. *Geophys. Res. Lett.*, **32**, L03704, <https://doi.org/10.1029/2004GL021462>.
- Mearns, L., F. Giorgi, P. Whetton, D. Pabon, M. Hulme, and M. Lal, 2003: Guidelines for use of climate scenarios developed from regional climate model experiments. Data Distribution Centre of the Intergovernmental Panel on Climate Change Doc., 38 pp., https://www.ipcc-data.org/guidelines/dgm_no1_v1_10-2003.pdf.
- Mejia, J. F., J. Huntington, B. Hatchett, D. Koracin, and R. G. Niswonger, 2012: Linking global climate models to an integrated hydrologic model: Using an individual station down-

- scaling approach. *J. Contemp. Water Res. Educ.*, **147**, 17–27, <https://doi.org/10.1111/j.1936-704X.2012.03100.x>.
- Meng, X., and Coauthors, 2018: Simulated cold bias being improved by using MODIS time-varying albedo in the Tibetan Plateau in WRF model. *Environ. Res. Lett.*, **13**, 044028, <https://doi.org/10.1088/1748-9326/AAB44A>.
- Muller, K. E., and B. L. Peterson, 1984: Practical methods for computing power in testing the multivariate general linear hypothesis. *Comput. Stat. Data Anal.*, **2**, 143–158, [https://doi.org/10.1016/0167-9473\(84\)90002-1](https://doi.org/10.1016/0167-9473(84)90002-1).
- NOAA, 2018: Global Historical Climatology Network Daily. NCEI, accessed 30 September 2017, <https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND>.
- Perkins, S., A. Pitman, N. Holbrook, and J. McAneney, 2007: Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *J. Climate*, **20**, 4356–4376, <https://doi.org/10.1175/JCLI4253.1>.
- Peterson, T. C., and R. S. Vose, 1997: An overview of the Global Historical Climatology Network temperature database. *Bull. Amer. Meteor. Soc.*, **78**, 2837–2850, [https://doi.org/10.1175/1520-0477\(1997\)078<2837:AOOTGH>2.0.CO;2](https://doi.org/10.1175/1520-0477(1997)078<2837:AOOTGH>2.0.CO;2).
- Piani, C., J. Haerter, and E. Coppola, 2010: Statistical bias correction for daily precipitation in regional climate models over Europe. *Theor. Appl. Climatol.*, **99**, 187–192, <https://doi.org/10.1007/s00704-009-0134-9>.
- Pilz, J., and G. Spöck, 2008: Why do we need and how should we implement Bayesian kriging methods. *Stochastic Environ. Res. Risk Assess.*, **22**, 621–632, <https://doi.org/10.1007/s00477-007-0165-7>.
- Poggio, L., and A. Gimona, 2015: Downscaling and correction of regional climate models outputs with a hybrid geo-statistical approach. *Spat. Stat.*, **14**, 4–21, <https://doi.org/10.1016/j.spasta.2015.04.006>.
- R Core Team, 2018: *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing, <https://www.R-project.org/>.
- Roberts, D. R., W. H. Wood, and S. J. Marshall, 2019: Assessments of downscaled climate data with a high-resolution weather station network reveal consistent but predictable bias. *Int. J. Climatol.*, **39**, 3091–3103, <https://doi.org/10.1002/joc.6005>.
- Schabenberger, O., and C. A. Gotway, 2017: *Statistical Methods for Spatial Data Analysis*. CRC, 512 pp.
- Schoof, J. T., and S. Pryor, 2001: Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks. *Int. J. Climatol.*, **21**, 773–790, <https://doi.org/10.1002/joc.655>.
- Seber, G. A., and A. J. Lee, 2012: *Linear Regression Analysis*. Vol. 329, John Wiley and Sons, 582 pp.
- Shrestha, M., S. C. Acharya, and P. K. Shrestha, 2017: Bias correction of climate models for hydrological modelling—Are simple methods still useful? *Meteor. Appl.*, **24**, 531–539, <https://doi.org/10.1002/met.1655>.
- Skamarock, W. C., and Coauthors, 2019: A description of the Advanced Research WRF Model version 4. NCAR Tech. Note NCAR/TN-556+STR, 145 pp., <https://doi.org/10.5065/1dfh-6p97>.
- Stahl, K., R. Moore, J. Floyer, M. Asplin, and I. McKendry, 2006: Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. *Agric. For. Meteorol.*, **139**, 224–236, <https://doi.org/10.1016/j.agrformet.2006.07.004>.
- Teutschbein, C., and J. Seibert, 2012: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *J. Hydrol.*, **456–457**, 12–29, <https://doi.org/10.1016/j.jhydrol.2012.05.052>.
- Themeßl, M. J., A. Gobiet, and A. Leuprecht, 2011: Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *Int. J. Climatol.*, **31**, 1530–1544, <https://doi.org/10.1002/joc.2168>.
- Thornton, P. E., M. M. Thornton, B. W. Mayer, N. Wilhelm, Y. Wei, R. Devarakonda, and R. Cook, 2012: Daymet: Daily surface weather on a 1 km grid for North America, 1980–2008. Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics, accessed 15 February 2019, <https://doi.org/10.3334/ORNLDAAAC/1219>.
- U.S. Geological Survey, 2019: 3DEP 1/3 arc-second DEM. USGS, accessed 9 September 2018, <https://viewer.nationalmap.gov/basic/>.
- Vandal, T., E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, 2017: DeepSD: Generating high resolution climate change projections through single image super-resolution. *Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Halifax, NS, Canada, Knowledge Discovery and Data Mining, 1663–1672, <https://doi.org/10.1145/3097983.3098004>.
- , —, and A. R. Ganguly, 2019: Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation. *Theor. Appl. Climatol.*, **137**, 557–570, <https://doi.org/10.1007/s00704-018-2613-3>.
- Varouchakis, E., and D. Hristopulos, 2013: Comparison of stochastic and deterministic methods for mapping groundwater level spatial variability in sparsely monitored basins. *Environ. Monit. Assess.*, **185** (1), 1–19, <https://doi.org/10.1007/s10661-012-2527-y>.
- Walton, D., and A. Hall, 2018: An assessment of high-resolution gridded temperature datasets over California. *J. Climate*, **31**, 3789–3810, <https://doi.org/10.1175/JCLI-D-17-0410.1>.
- Wang, T., A. Hamann, D. L. Spittlehouse, and T. Q. Murdock, 2012: ClimateWNA—High-resolution spatial climate data for western North America. *J. Appl. Meteor. Climatol.*, **51**, 16–29, <https://doi.org/10.1175/JAMC-D-11-043.1>.
- Wikle, C. K., A. Zammit-Mangion, and N. Cressie, 2019: *Spatio-Temporal Statistics with R*. CRC Press, 380 pp.
- Wilby, R. L., S. Charles, E. Zorita, B. Timbal, P. Whetton, and L. Mearns, 2004: Guidelines for use of climate scenarios developed from statistical downscaling methods. IPCC Doc., 27 pp., https://www.ipcc-data.org/guidelines/dgm_no2_v1_09_2004.pdf.
- Winter, J. M., B. Beckage, G. Bucini, R. M. Horton, and P. J. Clemins, 2016: Development and evaluation of high-resolution climate simulations over the mountainous northeastern United States. *J. Hydrol.*, **17**, 881–896, <https://doi.org/10.1175/JHM-D-15-0052.1>.
- Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.*, **107**, 4429, <https://doi.org/10.1029/2001JD000659>.
- , L. R. Leung, V. Sridhar, and D. Lettenmaier, 2004: Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, **62**, 189–216, <https://doi.org/10.1023/B:CLIM.0000013685.99609.9e>.
- Xu, C., J. Wang, and Q. Li, 2018: A new method for temperature spatial interpolation based on sparse historical stations. *J. Climate*, **31**, 1757–1770, <https://doi.org/10.1175/JCLI-D-17-0150.1>.
- Zia, A., and Coauthors, 2016: Coupled impacts of climate and land use change across a river–lake continuum: Insights from an integrated assessment model of Lake Champlain's Missisquoi Basin, 2000–2040. *Environ. Res. Lett.*, **11**, 114026, <https://doi.org/10.1088/1748-9326/11/11/114026>.