Ecological Applications, 26(5), 2016, pp. 1338–1351

Evaluation of downscaled, gridded climate data for the conterminous United States

R. Behnke,[1] S. Vavrus,[2,6] A. Allstadt,[3] T. Albright,[4] W. E. Thogmartin,[5] and V. C. Radeloff[3]

[1]*Numerical Terradynamic Simulation Group, University of Montana, 32 Campus Drive, Missoula, Montana 59812 USA*
[2]*Nelson Institute Center for Climatic Research, University of Wisconsin–Madison, 1225 West Dayton Street, Madison, Wisconsin 53511 USA*
[3]*SILVIS Lab, Department of Forest and Wildlife Ecology, University of Wisconsin–Madison, 1630 Linden Drive, Madison, Wisconsin 53706 USA*
[4]*Department of Geography, University of Nevada–Reno, 1664 North Virginia Street, Reno, Nevada 89557 USA*
[5]*Upper Midwest Environmental Sciences Center, United States Geological Survey, 2630 Fanta Reed Road, La Crosse, Wisconsin 54603 USA*

Abstract. Weather and climate affect many ecological processes, making spatially continuous yet fine-resolution weather data desirable for ecological research and predictions. Numerous downscaled weather data sets exist, but little attempt has been made to evaluate them systematically. Here we address this shortcoming by focusing on four major questions: (1) How accurate are downscaled, gridded climate data sets in terms of temperature and precipitation estimates? (2) Are there significant regional differences in accuracy among data sets? (3) How accurate are their mean values compared with extremes? (4) Does their accuracy depend on spatial resolution? We compared eight widely used downscaled data sets that provide gridded daily weather data for recent decades across the United States. We found considerable differences among data sets and between downscaled and weather station data. Temperature is represented more accurately than precipitation, and climate averages are more accurate than weather extremes. The data set exhibiting the best agreement with station data varies among ecoregions. Surprisingly, the accuracy of the data sets does not depend on spatial resolution. Although some inherent differences among data sets and weather station data are to be expected, our findings highlight how much different interpolation methods affect downscaled weather data, even for local comparisons with nearby weather stations located inside a grid cell. More broadly, our results highlight the need for careful consideration among different available data sets in terms of which variables they describe best, where they perform best, and their resolution, when selecting a downscaled weather data set for a given ecological application.

Key words: *climate; data set; ecoregions; extremes; gridded; resolution; weather.*

Introduction

Weather and climate constrain and drive ecological processes, including flowering and seed set phenology (Schwartz 1998, Wolfe et al. 2005, Polgar and Primack 2011), patterns in species occurrence and abundance (Bateman et al. 2012, Forcey et al. 2014), species group size (Thogmartin and McKann 2014), synchronization of population dynamics within and among species (the Moran effect) (Koenig 2002, Post and Forchhammer 2002), and disease transmission rates (Harvell et al. 2002). These processes often occur over fine spatial and temporal scales, are sensitive to spatial heterogeneity, and may entail nonlinear responses, making spatially continuous yet fine-resolution weather data important for ecological research and applied ecological predictions (Jones and Gladkov 2003, Parra et al. 2004). However, available weather and climate data may not match the information needs for ecological and conservation applications. Furthermore, the characteristics, limitations, and tradeoffs associated with different weather and climate data sets may not be easy for ecologists and conservation practitioners to discern. It is thus important that downscaled climate data sets are evaluated and characterized in terms of their suitability for ecological and conservation applications.

The most accurate climate data originate from measurements at quality-controlled weather stations, but many applications require a gridded weather product to provide complete spatial coverage (e.g., Abatzoglou 2013). There are many different ways to interpolate among weather stations, based on distance to stations and local geographic factors such as land cover, slope, and elevation. For example, the effects of elevation may be approximated by adjusting for the normal drop in temperature with altitude (lapse rate) or the usual local increase in precipitation up the slope of a mountain (Sheridan et al. 2010). However, data sets differ in how they make these assumptions, leading to differences in

Manuscript received 11 June 2015; revised 5 November 2015; accepted 23 November 2015. Corresponding Editor: D. S. Schimel.
[6]Corresponding Author. E-mail: sjvavrus@wisc.edu

the products. Despite these uncertainties, gridded station data are often treated as observations by users for a particular application. There is a risk that many users are unaware of the accuracy of the gridded product they choose and whether another choice might be more suitable for their purposes. Users will expect downscaled climate products to produce data that match local station data as closely as possible, particularly if gridded data have relatively high spatial resolution.

The overall goal of our study is to evaluate the ability of eight widely used, gridded data sets to represent actual weather conditions in the conterminous United States, including extreme weather events. Specifically, our purpose is to evaluate how closely the gridded data match the original station data and thus what information may be altered when a user substitutes a gridded data set in place of meteorological station data. This concern is especially pertinent to extreme weather, which tends to be more localized than general weather phenomena, especially precipitation extremes. To achieve our goal, we focus on four major research questions.

Our first question is how well gridded data sets capture temperature and precipitation, based on the expectation that performance may differ greatly between these variables, because precipitation is generally more heterogeneous than temperature in both space and time. Thus, we anticipated that gridded temperature data would resemble point measurements from nearby weather stations more closely than gridded precipitation data.

Our second question is how the accuracy of different downscaled data sets varies among ecoregions. Climate differs considerably among ecoregions, as does topography, posing challenges for any one-size-fits-all downscaling algorithm. Therefore, we expected that while one data set would perform best, for instance, in desert ecoregions, another would be optimal in temperate forest regions. Similarly, given that topography will increase the spatial heterogeneity of both temperature and precipitation, one data set may provide the highest accuracies in mountainous ecoregions, while another may be optimal in flatter areas.

Our third question is how well downscaled data sets capture means vs. extremes. Both mean climate and weather extremes are important for ecological research. Means are typically used to predict the effects of future climate change, but extremes may exert more immediate and profound effects on ecosystems (Jentsch et al. 2009). However, mean climate patterns are more homogeneous in both space and time than extremes, suggesting that differences in interpolation methods may result in different accuracies, and we expected accuracies for means to be higher than for extremes.

Our fourth and final question is how much the accuracy of the gridded products depends on their spatial resolution. Finer resolution is typically desirable for ecological applications, because many ecological patterns and processes vary at fairly fine resolutions (e.g., less than 1 km). However, while it is technically straightforward to create downscaled weather data sets at any spatial resolution, the question is whether the accuracy of the predicted patterns improves when interpolated to finer scales. We expect that as their resolution becomes finer, gridded products will more closely align with nearby weather stations and be better able to account for topography, thus matching observational data more realistically, but that improvement could come with a greater risk of false precision and overfitted interpolation models.

Data and Methods

Study area

Our study area is the conterminous United States, whose size, diverse climate regimes, and pronounced topographic variations provide a challenging test for gridded data sets. We evaluated these data sets based on their agreement with weather stations averaged across the entire domain and divided into 17 ecoregions that are designed to capture spatial climate variations (Fig. 1). These ecoregions are patterned after the classification scheme of Bukovsky (2011) and resemble the regions used by the National Ecological Observation Network (NEON) (Kampe et al. 2010).

Data

As reference data, we used daily precipitation and temperature (maximum and minimum) records from the Global Historical Climate Network-Daily (GHCN-D) data set (Menne et al. 2012). We selected 3855 stations that had a minimum of 83% of daily records for each of the three variables for our reference period, 1981–2010 (Fig. 1). Values flagged by GHCN's quality control procedure (Durre et al. 2010) were removed. For each gridded data set, we extracted the temperature and precipitation time series only for grid cells containing at least one weather station. If a grid cell contained more than one station, we averaged the daily values to obtain a single time series for the cell. However, the overwhelming majority (> 90%) of grid cells contain only one weather station. Because the data sets vary in their representation of coastal areas, they occasionally do not include weather stations close to a coast and thus the number of weather stations varies slightly among data sets. We emphasize that our methodology does not directly address the spatially interpolated distribution of variables far from weather stations, but instead focuses on local comparisons between observed and downscaled weather data. A similar supplemental analysis was performed using observations independent of the GHCN measurements from three geographically distinct states: the California Irrigation Management Information System (CIMIS), the Florida Automated Weather Network (FAWN), and the Nebraska Automated Weather Data Network (NE_AWDN).

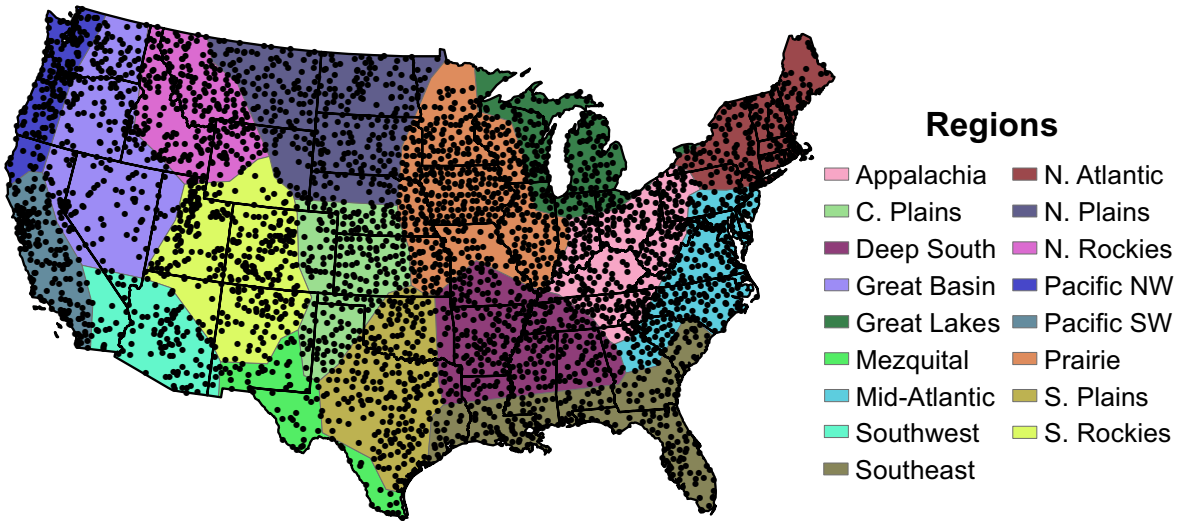

FIG. 1. The modified Bukovsky regions and the locations of stations used as reference data.

We investigated eight widely used gridded products that provide temperature and/or precipitation data on daily time scales throughout our study period: CPC, Daymet, Livneh, Maurer, NLDAS2, PRISM, TopoWx, and UIdaho (Table 1). Their spatial resolutions range from fine (TopoWx at 800 m and DayMet at 1 km) to fairly coarse (CPC at 0.25°, or ~25 km), allowing us to evaluate accuracy as a function of resolution (Question 4). Because gridded products differ in how they define a calendar day (e.g., local time relative to Coordinated Universal Time), appropriate lag correlations were applied through cross-correlation analysis to account for the several-hour offset in daily station data.

The downscaling and gridding methods of these data sets vary substantially and are summarized in Table 2. It is beyond the scope of this study to describe these methods in detail, but we provide a brief overview here. Broadly speaking, the various interpolation methods can be roughly categorized according to their sophistication into three groups. The simplest methods, lapse rate-based and gauge-based, apply a constant elevation correction for temperature (−6.5°C/km) and a basic regression or weighting algorithm to upscale precipitation measured at a rain gauge to the mean precipitation in a grid cell. This method is used to generate the CPC, Maurer, and Livneh products, which apply either the optimum interpolation algorithm (CPC) or the SYMAP computer mapping algorithm (Maurer, Livneh) that employs an inverse-distance weighting approach. An intermediate level, topoclimatic, incorporates environmental covariates such as elevation, aspect, slope, distance to coast, and land surface temperature. This procedure is used to generate PRISM, Daymet, and TopoWx. A third method, employed to generate NLDAS2 and UIdaho, employs a hybrid approach by using data from a combination of sources, including station data, reanalysis, other monthly or daily gridded data, or modeled data. For example, precipitation in the NLDAS2 data set is debiased according to observations, while its temperature fields are downscaled directly from reanalysis data (i.e.,

TABLE 1. Information on the eight gridded data products used in this study.

Data set	Variables used	Time span	Resolution (km)
Climate prediction center unified gauge-based analysis of daily precipitation (CPC)	prcp	1948–	28 × 21
Daymet	prcp, tmax, tmin	1980–2014	1 × 1
Livneh	prcp, tmax, tmin	1915–2013	7 × 5
Maurer	prcp, tmax, tmin	1949–2010	14 × 10
National land data assimilation system, version 2 (NLDAS2)	prcp, tmax, tmin	1979–	14 × 10
Parameter-elevation regressions on independent slopes model (PRISM (AN81d))	prcp, tmax, tmin	1981–2014	5 × 4
Topographical (TopoClimatic) weather (TopoWx)	tmax, tmin	1948–2014	0.8 × 0.8
UIdaho	prcp, tmax, tmin	1979–2014	4 × 4

Notes: Variables are daily precipitation (prcp), daily maximum temperature (tmax), and daily minimum temperature (tmin). Data sets for which resolutions are in degrees have been converted to an equivalent resolution expressed as latitude × longitude for a grid box at 40°N, approximately in the center of the analysis domain.

TABLE 2. Interpolation methods and references for the data products.

Data set	Interpolation or gridding method applied	Reference
CPC	modified optimum interpolation technique of Chen et al. (2002), orographic effects accounted for via PRISM methodology	Higgins et al. (2000)
Daymet	geographically weighted regression	Thornton et al. (1997)
Livneh	synergraphic mapping system (SYMAP) of Shepard (1984), precipitation scaled to match PRISM climatology	Livneh et al. (2013)
Maurer	synergraphic mapping system (SYMAP) of Shepard (1984), precipitation scaled to match PRISM climatology	Maurer et al. (2002)
NLDAS2	bilinear interpolation of NCEP-NCAR Reanalysis (Kalnay et al. 1996) adjusted for elevation using PRISM methodology, temporally disaggregated to one hour	Xia et al. (2012)
PRISM	geographically and elevation-weighted regression, station weighting by topography, distance to coast, atmospheric factors	Daly et al. (2008)
TopoWx	moving window regression kriging, geographically weighted regression	Oyler et al. (2015)
UIdaho	bilinear interpolation of NLDAS2 data, daily prcp data from NLDAS2 scaled by monthly prcp from PRISM, daily temperatures from NLDAS2 scaled by monthly temperatures from PRISM	Abatzoglou (2013)

simulations based on physical models), with no additional use of station data.

Our classification into gauge-based, topoclimatic, and hybrid methods roughly corresponds to commonly employed strategies to downscale and interpolate, including the translation of coarse-resolution climate model output to finer scales by either statistical or dynamical downscaling. Such downscaling techniques have a long history, including simple inverse-distance weighting schemes such as Cressman (1959) and Shepard (1968), optimal interpolation algorithms that minimize mean-square interpolation errors in a network (Gandin 1965), sophisticated kriging methods that describe how spatial data are related as a function of distance and direction by modeling the covariance structure of the station and covariate data (Journel and Huijbregts 1978), and blending techniques that account for known spatial relationships among weather station data (Johns et al. 2003). Even more sophisticated interpolation methods involving Bayesian statistics have also been applied (e.g., Fuentes et al. 2006). Likewise, physically based approaches take advantage of known spatial relationships among weather variables across a measurement network, based on physical laws relating temperature, precipitation, solar radiation, humidity, etc. (e.g., Kittel et al. 1995). This reasoning also forms the basis for data assimilation algorithms that incorporate weather observations into numerical models that conform the measurements into physically realistic four-dimensional fields (Whitaker and Hamill 2002), a technique utilized in atmospheric reanalysis products such as MERRA (Rienecker et al. 2011), NCEP-NCAR Reanalysis (Kalnay et al. 1996), the 20th-Century Reanalysis (Compo et al. 2011), and the NLDAS2 data set used here. Physically based strategies also form the basis for regional climate model simulations, with the primary purpose to translate climate data at coarse scales (e.g., from a global model) to finer scales suitable for applications.

To address our four research questions, we calculated a variety of metrics on a national and regional basis. We used standard statistical metrics of mean bias, mean absolute error, root mean squared error, standard deviation, and linear correlation to compare the accuracy of the gridded products in terms of temperature and precipitation (Question 1) in each region (Question 2). We synthesized these results on a nationwide basis by using a Taylor diagram, a commonly used tool in climatology that measures agreement with observations among several metrics and variables (Taylor 2001), in this case temperature and precipitation. To assess regional (and national) performance among the gridded products, we compared their mean bias and mean absolute error in daily temperature and precipitation in each Bukovsky region through the use of "portrait plots" (Gleckler et al. 2008). To evaluate the representation of extreme weather (Question 3), we use 27 metrics collectively known as CLIMDEX or the ETCCDI (Expert Team on Climate Change Detection and Indices) (Zhang et al. 2011). These commonly used indices include measures of extreme temperature and precipitation using a variety of thresholds and time scales (Table 3), allowing for a fairly thorough evaluation of how the gridded products compare with station data under extreme temperature and precipitation conditions. The CLIMDEX indices have been used in many recent studies of past and future climate change (Alexander et al. 2006, Donat et al. 2013, Sillmann et al. 2013*a*,*b*).

We grouped the 27 CLIMDEX indices into three categories. The absolute indices are those that use an absolute threshold to define an extreme, such as the number of occurrences above or below a fixed value (e.g., days with a maximum temperature < 0°C). The second is the min_max category, which indicates how well a data set represents the minimum or maximum values of temperature or precipitation over the course of a month (e.g., the highest daily maximum temperature in a month). The final category is the relative indices, which measure extremes based on a relative scale, such as the frequency of days when precipitation exceeds the local 95th percentile. We did not assign total annual precipitation (prcptot), daily temperature range (dtr), and

Table 3. Description of CLIMDEX indices.

CLIMDEX index	Description	Units	Category
Temperature			
Extreme			
TNx	monthly maximum value of daily minimum temperature	°C	min_max
TXx	monthly maximum value of daily maximum temperature	°C	min_max
TXn	monthly minimum value of daily maximum temperature	°C	min_max
TNn	monthly minimum value of daily minimum temperature	°C	min_max
TN10p	monthly percentage of days when TN < 10th percentile	n/a	relative
TX10p	monthly percentage of days when TX < 10th percentile	n/a	relative
TN90p	monthly percentage of days when TN > 90th percentile	n/a	relative
TX90p	monthly percentage of days when TX > 90th percentile	n/a	relative
WSDI	annual warm spell duration index (count of at least 6 consecutive days when TX > 90th percentile)	days	relative
CSDI	annual cold spell duration index (count of at least 6 consecutive days when TN < 10th percentile)	days	relative
Threshold			
DTR	daily temperature range: monthly mean difference (TX – TN)	°C	
FD	annual count of days when TN < 0°C	days	absolute
SU	annual count of days when TX > 25°C	days	absolute
ID	annual count of days when TX < 0°C	days	absolute
TR	annual count of days when TN > 20°C	days	absolute
GSL	annual count of days between first span of at least 6 days with TM > 5°C and first span after 1 July of at least 6 days with TM < 5°C	days	absolute
Precipitation			
Extreme			
Rx1 day	monthly maximum 1-day precipitation	mm	min_max
Rx5 day	monthly maximum consecutive 5-day precipitation	mm	min_max
CDD	annual maximum length of dry spell (days with RR < 1 mm)	days	absolute
CWD	annual maximum length of wet spell (days with RR > 1 mm)	days	absolute
R95pTOT	annual total precipitation when RR > 95th percentile	mm	relative
R99pTOT	annual total precipitation when RR > 99th percentile	mm	relative
Threshold			
R1 mm	annual count of days when precipitation ≥ 1 mm	days	absolute
R10 mm	annual count of days when precipitation ≥ 10 mm	days	absolute
R20 mm	annual count of days when precipitation ≥ 20 mm	days	absolute
PRCPTOT	annual total precipitation	mm	
SDII mm	simple precipitation intensity index (mean precipitation on days with RR ≥ 1 mm)		

Notes: All monthly indices were aggregated to annual for this study. Description includes minimum temperature (TN), maximum temperature (TX), mean temperature (TM), and precipitation (RR). Indices with "n/a" units were unitless.

the simple precipitation intensity index (sdii) to any category.

Following Gleckler et al. (2008), we compared the gridded products based on the CLIMDEX indices using measures of relative error among data sets. We calculated root mean square error (RMSE) for each index and region, and assessed the performance of a given data set via the relative RMSE (rRMSE), which compares the RMSE in a data set ($RMSE_{ds}$) with the median RMSE ($RMSE_{Md}$) of all data sets

$$\mathrm{rRMSE} = \frac{\mathrm{RMSE}_{ds} - \mathrm{RMSE}_{Md}}{\mathrm{RMSE}_{Md}} \quad (1).$$

This method allows for direct comparison among the products. For example, an rRMSE value of −0.50 indicates that data set's RMSE is 50% lower (better) than the model median RMSE. In addition to the rRMSE metric for extremes, we also evaluated the gridded products based on their bias relative to observations near the middle of the temperature and precipitation distributions vs. the tails of those distributions.

To assess the role of resolution (Question 4), we compared these evaluation metrics with the spatial grain of each product. Doing so allowed us to determine the overall influence of spatial resolution, as well as specific relationships among temperature and precipitation, regional differences, and extremes.

We evaluated the eight gridded data products using the various measures described previously, beginning with broad-scale metrics (correlation, standard deviation, and RMSE) at the national level, followed by more targeted metrics (mean absolute error and mean bias) at

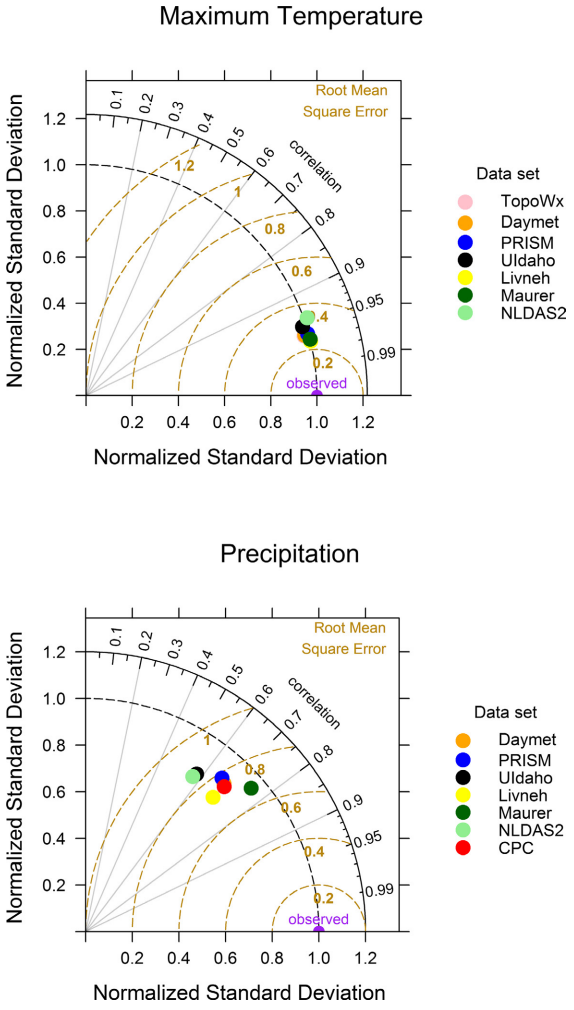

FIG. 2. Taylor diagram showing the ability of gridded data sets to represent station observations for maximum temperature and precipitation. Minimum temperature is not shown because it is very similar to maximum temperature.

both the national and regional levels. These evaluations captured the ability of the gridded data products to represent observed weather conditions over the entire distribution of temperature and precipitation. We then evaluated for extreme weather conditions using the CLIMDEX indices and assessed the influence of spatial resolution. All of the gridded data sets provide both temperature and precipitation data, except for CPC (precipitation only) and TopoWx (temperature only).

RESULTS

Accuracy of temperature vs. precipitation

Across the conterminous United States, gridded values of temperature matched weather station records closely, with very high correlation coefficients (> 0.9) and nearly identical temporal variability (Fig. 2). By contrast, the agreement for precipitation was much weaker, as all the data sets exhibited correlations < 0.8 with much larger RMSE and weaker temporal variability than observed. According to the Taylor diagram, the Maurer data provided the best match with observations in terms of correlation, variability (daily standard deviation), and RMSE, whereas NLDAS2 and UIdaho showed the largest discrepancies. Thus, at least on a nationwide basis, temperature was represented far more realistically and uniformly than precipitation in all of the gridded products (Question 1).

Regional variations

Regional variations in the performance of the gridded products, as displayed in the portrait plots, showed considerable variability in mean bias and mean absolute error for temperature and precipitation among the Bukovsky regions, as well as the national average (Fig. 3). Mean regional biases in temperature ranged from −2.5°C to +4°C and were almost all negative for the daily maximum (gridded data were cooler than observed), whereas biases of both signs occurred for the daily minimum. Although no product stood out as superior across ecoregions, Daymet had the smallest mean bias nationwide (−0.06°C). Conversely, NLDAS2 overestimated minimum temperature (mean bias = 2.46°C) and underestimated maximum temperature (mean bias = −1.23°C) much more than the other data sets. Mean absolute errors for both minimum and maximum temperature generally clustered between 1°C and 2°C in all regions, except for NLDAS2, which had values > 3°C for maximum temperature and > 4°C for minimum temperature. The influence of topography was not straightforward. With respect to the mean bias, regions with high topography, such as the Northern and Southern Rockies, exhibited the largest differences with observations, while flatter and more topographically uniform regions, such as the Plains, showed the smallest differences. However, this topographic effect was much less clear in terms of mean absolute error, in that the Northern and Southern Rockies had among the largest differences of any region, yet so did the much flatter Northern and Central Plains. Overall, the magnitude of the mean bias and mean absolute error, when averaged over all data sets and regions, was comparable between the daily maximum and daily minimum temperature (−0.31°C and 0.26°C for mean bias and 1.69°C and 1.49°C for mean absolute error, respectively).

For precipitation, PRISM had a clear negative mean bias in most regions, while the mean bias for other data sets was slightly positive. Livneh exhibited the smallest mean bias nationwide (0.03 mm) and Daymet the largest (0.14 mm). However, the magnitude of the mean biases among the data sets was very small, < 0.2 mm, compared with the mean absolute errors that sometimes exceeded 3 mm. The spread of the mean absolute errors was more substantial, ranging from a low of 1.17 mm in Maurer

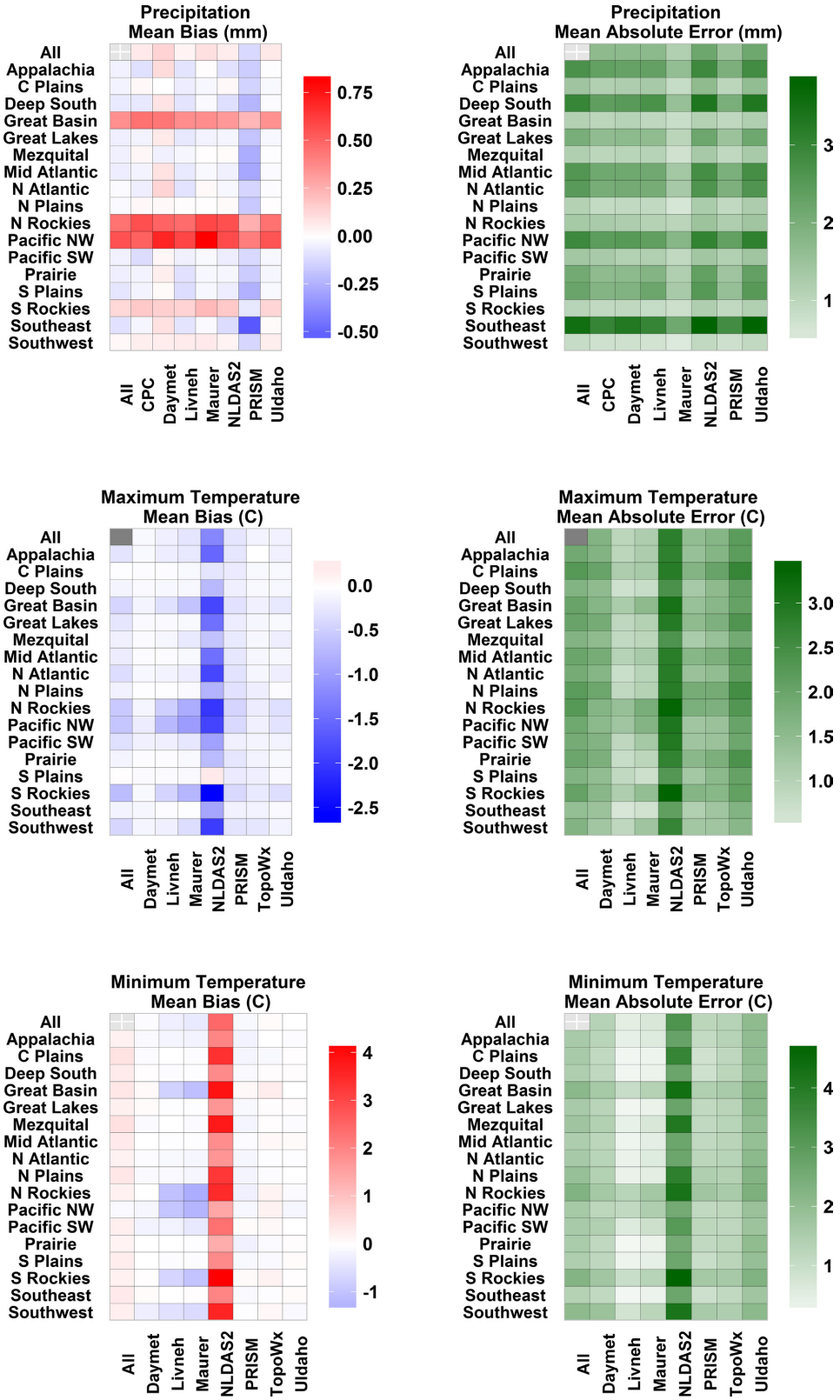

FIG. 3. Portrait plot showing the annual mean bias and mean absolute errors by region for gridded temperature and precipitation compared to observed values. The All categories represent the mean bias and mean absolute error for the average temperature and precipitation values of all the gridded data sets in a region (*x*-axis) or of all the regions in a data set (*y*-axis).

to a high of 2.19 mm in NLDAS2. The data sets consistently indicated more precipitation than observations in highly elevated regions, such as the Northern Rockies, Pacific Northwest, and Great Basin. By contrast, the mean absolute errors were more closely associated with precipitation amount, with the largest discrepancies in the relatively wet Deep South, Southeast, and Pacific Northwest, while the smallest errors occurred in the arid to semiarid climates of the Southwest, Southern Rockies, and Northern Plains.

Extremes

As a bridge between these evaluations comprising all the daily data and the detailed CLIMDEX indices focused on extremes, we first compared differences between the gridded products and weather stations across the whole distribution of temperatures and precipitation amounts over the entire domain (Fig. 4). In terms of temperature, the gridded products were fairly realistic across most of the distribution, but we found discrepancies at both tails. The sign of these biases differed, however, such that the gridded products underestimated the magnitude of both extremely hot and cold weather. Overall, the magnitude of the mean bias was considerably larger for minimum temperature than maximum temperature, especially under the coldest conditions. Biases in daily maxima were below 7°C for maximum temperature, whereas all the data sets except Livneh exceeded this bias in minimum temperature at some point in the distribution. We emphasize, however, that very rarely do temperatures in the continental United States reach the extremes plotted at the tails of these distributions, and locations in the vast majority of the domain never experience them.

Under the coldest conditions, Maurer and Livneh showed little bias, but both underestimated the magnitude of the most extreme heat (above 45°C). In contrast, Daymet, TopoWx, PRISM, and UIdaho had greater mean biases for extremely cold weather than the very warmest days, for which their mean bias remained relatively small (≤ -4°C). NLDAS2 displayed noticeable biases across the entire temperature distribution, consisting of an overall cool bias for maximum temperature and a warm bias for minimum temperature, with a sharp increase in biases at both tails of each temperature distribution.

The biases for precipitation were fairly uniform, in that the mean bias increased nearly linearly in each data set until the observed daily precipitation reached an extremely high value of ~175 mm, beyond which the behavior was more erratic. Consistent with the Taylor diagram and portrait plots, Maurer most closely matched station observations across the entire precipitation distribution, while the mean biases were consistently largest in NLDAS2 and UIdaho.

The results for extreme temperature and precipitation from the CLIMDEX indices were generally consistent with our findings for the entire distribution, as summarized among the three categories (Fig. 5) and illustrated for every index (Appendix S1: Figs. S1 and S2). We found the best overall match with observations for the Maurer product, and the largest biases for NLDAS2. In particular, NLDAS2 was erroneous in its depiction of temperature extremes based on absolute thresholds, consistent with its systematic biases across the entire distribution (Fig. 4). With respect to the other two categories of extreme temperature, the Livneh and Maurer products were very comparable in matching point observations better than the other data sets. Maurer stood out even more clearly in representing precipitation extremes accurately, producing the smallest errors in every category. Livneh was similarly accurate for relative measures of extremes, but fared less credibly in depicting extremes in the other two categories, especially absolute indices. The limitation of using relative measures of error in constructing the portrait plots prevented a clear regional assessment, but we note that for extreme precipitation there was a larger spread across products in their performance within regions in the absolute precipitation category than in the min-max and relative categories (Fig. 5). The same conclusion also holds for extreme temperature, but the magnitude of intra-regional variability in the absolute temperature category was largely dictated by the outlier NLDAS2 data set.

Spatial resolution

We assessed the influence of spatial resolution by comparing the relationship between resolution and performance among the gridded products in these results. Our results showed that the Maurer product agreed best overall with weather station observations, both for temperature and precipitation, whereas NLDAS2 generally had the largest discrepancies. This result is noteworthy, because both of these data sets have the same spatial resolution (14 × 10 km), which was the coarsest except for CPC's precipitation-only data. Conversely, the two products with the finest resolution, TopoWx (800 m) and DayMet (1 km), generally fell in the middle of the pack in terms of their agreement with observations. This was the case even for weather extremes (Fig. 5) and in the most topographically complex regions (Fig. 3), where we had expected that a fine resolution would matter most. Particularly surprising was the lack of an influence of resolution on precipitation extremes (Fig. 4), because we expected larger grid cells would dilute the intensity of heavy rainfall interpolated from weather station data. Yet even for extreme precipitation, Maurer produced the smallest biases throughout virtually the entire range of precipitation amounts.

Independent weather stations

The analyses described previously relied on comparisons among the gridded products and some weather

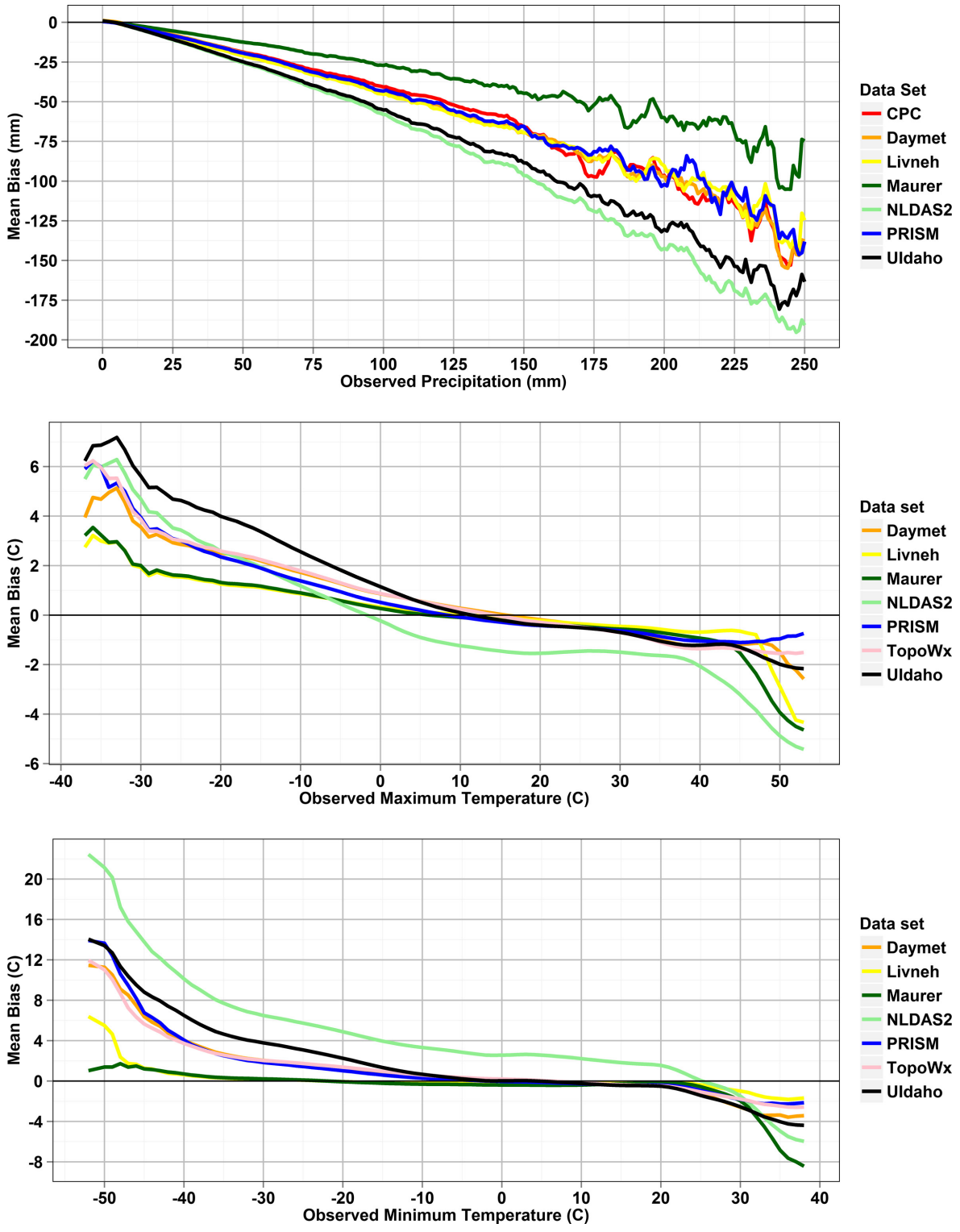

FIG. 4. Line plots showing the mean bias of gridded precipitation and temperature data relative to observed values. A nine-point and a five-point running mean were applied to the precipitation and temperature data mean biases, respectively.

stations from which the gridded data were ultimately derived. It was not possible to verify our analyses using a complete network of independent weather stations, because there is no such alternative data set of sufficient spatial and temporal coverage available. In lieu of this, we compared the gridded products against three sets of

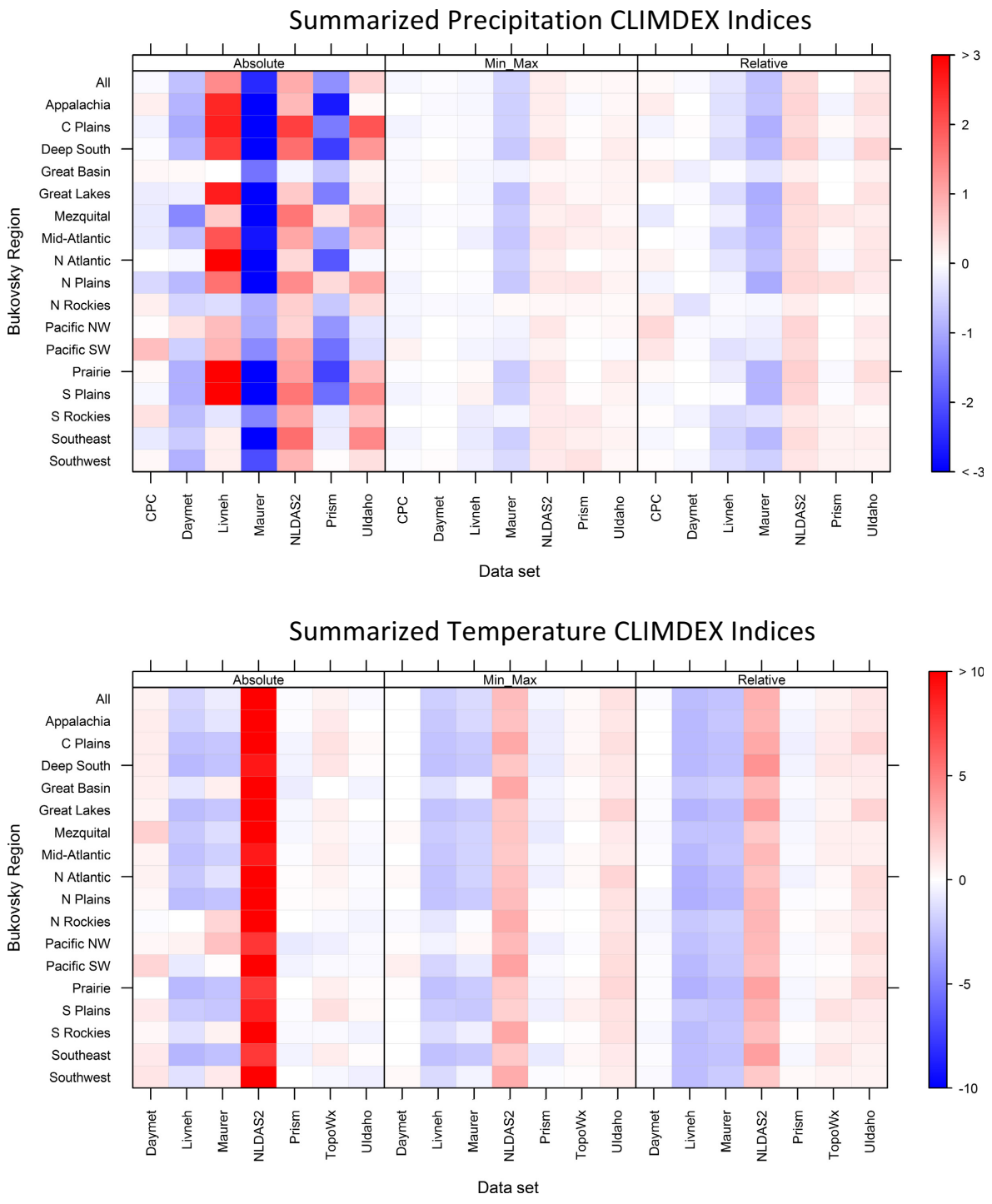

FIG. 5. Summarized precipitation and temperature CLIMDEX indices. Values are equal to the sum of the individual CLIMDEX indices, as categorized in Table 3.

regional weather networks from very different parts of the conterminous United States, comprised of stations independent of the GHCN measurements (Data and Methods, Appendix S2: Fig. S1). As summarized in Appendix S2, the conclusions from these supplemental comparisons were similar to those obtained from the GHCN observations for each of our four primary research questions. Again, the gridded data sets represented temperature much more accurately than precipitation among all three networks (Appendix S2: Fig. S2).

We also found the same kind of complex relationships between accuracy and topographic variability among the regional networks as in the GHCN comparisons. The largest mean biases in temperature occurred over the most geographically diverse network in California, but the precipitation biases in that state were somewhat lower than in the other two networks (Appendix S2: Figs. S2 and S4). However, the much higher rainfall amounts in Florida probably boosted the magnitude of errors in that state. As with the GHCN comparisons, climatic averages in the gridded products were represented much more accurately than extremes, and daily maximum temperatures were better captured than daily minima (Appendix S2: Figs. S2–S4). Furthermore, the resolution of the gridded data sets was not an important factor in explaining their accuracy in the comparison among regional weather networks. The finest-resolution product, Livneh, generally fell within the middle of the pack, while NLDAS2 again displayed the largest biases, particularly for temperature (Appendix S2: Figs. S2–S5), even though it is not the coarsest data set.

Discussion and Conclusions

We evaluated the performance of commonly used gridded weather data sets for the conterminous United States. Our study was designed for users of these products for ecological research and applied ecological predictions, with particular emphasis on extreme temperature and precipitation. Our study was not intended to serve as a "beauty contest," in part because each user has a unique purpose for applying these products and also because the agreement between the data sets and observations differs greatly across variables and regions. In fact, no data set was "best" everywhere and for all the variables we analyzed. Instead, our purpose was to present how well the gridded data sets agree with station observations at overlapping locations and to assess the relative agreement between variables (temperature vs. precipitation), across regions, and between means and extremes, so that ecologists and conservation biologists can select the data set that is best suited for their purposes. We believe that ours is the first study to comprehensively evaluate a large set of these data sets on a large geographic scale.

For the first three research questions, our findings matched expectations, though the degree of differences between gridded products and weather station records were intriguing. We found that the gridded data matched observations of temperatures much better than those of precipitation (Question 1; Fig. 2), and regional variation in performance indicated that topographically complex regions are most difficult for these models (Question 2; Fig. 3). The gridded products reproduced average weather conditions more accurately than extremes for both temperature and precipitation (Question 3; Fig. 4), but we found larger mean biases in daily minimum than daily maximum temperatures. This feature may stem from reduced atmospheric mixing at night, which can produce more spatially variable temperatures associated with urban heat islands and cold-air drainage in mountainous terrain (Hocevar and Martsolf 1971, Oke 1995). In terms of precipitation magnitude, we found a nearly linear increase in the deviation from observations in all data sets, some of which exhibited a bias of more than 50% for very heavy rainfalls (> 200 mm/d). For both temperature and precipitation, biases were quite large for the most extreme values, but such outliers are very unusual and do not occur at many locations.

Contrary to our expectations for Question 4, we found no clear relationship between the resolution of gridded products and their agreement with observations, either for average conditions (Figs. 2 and 3) or extremes (Figs. 4 and 5). For example, the moderate-resolution PRISM (5 × 4 km resolution) and high-resolution DayMet (nearly the finest coverage at 1 km) data sets had the largest nationwide mean biases in precipitation, whereas the coarsest product, CPC at 25 km, fell right in the middle. Mean biases and absolute errors in temperature were largest in NLDAS2, whereas the domain-averaged, mean absolute error in temperature was second-smallest in Maurer, even though both products have the same resolution. The lack of a relationship between accuracy and resolution was even more surprising for extreme precipitation, for which depicted intensity typically declines with increasing grid box area (Chen and Knutson 2008). Nevertheless, the Maurer data set clearly agreed best with observations in representing extreme precipitation, even though its resolution is the same as that of NLDAS2, which showed the worst overall match with observations. In fact, CPC is the only product coarser than Maurer. Therefore, differing assumptions and methodologies among the data sets must be responsible for overriding the presumably beneficial effect of resolution.

In this study, two products stood out in their overall tendency to be closest to (Maurer) and farthest from (NLDAS2) observed measurements. This assessment holds for both extremes and moderate weather conditions, although there are exceptions for certain variables and regions. The precise reason(s) why these two are outliers is beyond the scope of this study, but we demonstrated previously that resolution alone was not the cause. A possible factor is the contrasting structure of these two data sets. The Maurer product is, in part, derived from the same weather station data used in our evaluation, and it applies interpolation algorithms that do not use inputs besides elevation as predictors (*Data and Methods*). By contrast, NLDAS2 uses data from a combination of sources to construct gridded estimates. In addition, the daily maximum and minimum temperatures in NLDAS2 are based on hourly data obtained from the temporal disaggregation of three-hour NARR (North American Regional Reanalysis) data, rather than the actual high and low temperatures recorded by weather stations and used by the other gridded products. Consequently, NLDAS2 would be expected to underestimate

(overestimate) the daily temperature maximum (minimum), consistent with its systematic bias identified here.

Because our analysis was limited to grid cells containing weather stations, we did not evaluate the ability of each data set to interpolate temperature or precipitation elsewhere. However, our findings should still be relevant to unsampled locations for several reasons. First, temperature and precipitation patterns typically do not change greatly from one grid cell to the next, except under unusual conditions such as extreme elevation changes. Second, the 3855 stations that we used in our analysis represent only a fraction of the total used in creating the data sets. For example, Livneh uses ~30000 stations, PRISM at least 10000 stations, and TopoWx over 14000 stations (Daly et al. 2008, Oyler et al. 2015). Because of these dense measurement networks, the average distance between stations is fairly small. Third, errors similar to those revealed by our local evaluations of weather station data can also be expected to occur elsewhere, as a reflection of inherent biases in the interpolation algorithm employed by a particular gridded product. Therefore, we believe that a realistic representation of weather conditions at nearby weather stations is a prerequisite for a product to be considered trustworthy elsewhere, and that errors prevalent in grid cells co-located with weather stations suggest that similar biases occur throughout the spatial domain. While limited in scope, our analysis of three independent networks supports this interpretation.

One purpose of this study was to analyze a broad range of extreme weather conditions among the products, so we chose the widely used CLIMDEX indices. While there are many advantages to using these particular metrics, they do pose some challenges for evaluation, especially because not all the indices measure what is considered an extreme event at many locations. For example, a daily rainfall of 20 mm (R20 mm) may constitute an extreme event in some places, but it is a fairly common occurrence in the eastern half of the United States.

An ideal analysis would base evaluation upon a network of weather stations completely independent of the gridded data products. While at this time no such alternative data set with sufficient spatial and temporal coverage is available, the Meteorological Assimilation Data Ingest System (MADIS) network is growing in both the number of participating stations and length of record, and may be useful for such an assessment in the future. However, the general agreement between the results from the nationwide analysis with GHCN weather stations and the regional networks (Results and Appendix S2) indicates that our conclusions are likely robust. We suggest our findings be interpreted as lower bounds for expected errors in the application of these data sets if the products were evaluated at all grid cells in the domain.

We hope that our study will encourage ecologists to consider carefully the details of available gridded data products, rather than treating them as "black boxes." For example, our analysis has demonstrated that selecting the highest resolution product may not translate to the most accurate data. Most of the gridded data sets we evaluated would be appropriate for ecological applications typically using long-term climate measures (e.g., species distribution modeling) or seasonal weather conditions (e.g., annual population models). However, the effects of shorter-term, extreme weather events are increasingly included in these applications (e.g., Altwegg et al. 2006, Bateman et al. 2012, Descamps et al. 2015), so gridded data sets should be selected with care. As more weather records become available at new locations and new downscaling techniques are developed, gridded data sets should continue to improve in accuracy and precision.

Acknowledgments

This research was supported by a grant from the NASA Biodiversity Program through a subaward from the U. S. Fish and Wildlife Service, F12AP00423, and by grants from the United States Geological Survey (G14AP00182) and the National Science Foundation (NSF EPSCoR Track-1 NSF-IIA-1443108). The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the U. S. Fish and Wildlife Service. Any use of trade, product, or firm names are for descriptive purposes only and do not imply endorsement by the U.S. Government. We thank one anonymous reviewer and editor for helpful comments on earlier versions of this manuscript, and B. Bateman, P. Heglund, and A. Pidgeon for feedback and discussions.

Literature Cited

Abatzoglou, J. T. 2013. Development of gridded surface meteorological data for ecological applications and modelling. International Journal of Climatology 33:121–131.

Alexander, L. V., et al. 2006. Global observed changes in daily climate extremes of temperature and precipitation. Journal of Geophysical Research: Atmospheres 111:DO5109. http://dx.doi.org/10.1029/2005JD006290

Altwegg, R., A. Roulin, M. Kestenholz, and L. Jenni. 2006. Demographic effects of extreme winter weather in the barn owl. Oecologia 149:44–51.

Bateman, B. L., J. VanDerWal, and C. N. Johnson. 2012. Nice weather for bettongs: using weather events, not climate means, in species distribution models. Ecography 35:306–314.

Bukovsky, M. S. 2011. Masks for the Bukovsky regionalization of North America, Regional Integrated Sciences Collective, Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, CO.

Chen, C., and T. Knutson. 2008. On the verification and comparison of extreme rainfall indices from climate models. Journal of Climate 21:1605–1621.

Chen, M., P. Xie, J. E. Janowiak, and P. A. Arkin. 2002. Global land precipitation: a monthly analysis based on gauge observations. Journal of Hydrometeorology 3:249–266.

Compo, G. P., et al. 2011. The twentieth century reanalysis project. Quarterly Journal of the Royal Meteorological Society 137:1–28.

Cressman, G. P. 1959. An operational objective analysis system. Monthly Weather Review 87:367–374.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris. 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. International Journal of Climatology 28:2031–2064.

Descamps, S., A. Tarroux, O. Varpe, N. G. Yoccoz, T. Tveraa, and S.-H. Lorentsen. 2015. Demographic effects of extreme weather events: snow storms, breeding success, and population growth rate in a long-lived Antarctic seabird. Ecology and Evolution 5:314–325.

Donat, M. G., L. V. Alexander, H. Yang, I. Durre, R. Vose, and J. Caesar. 2013. Global land-based data sets for monitoring climatic extremes. Bulletin of the American Meteorological Society 94:997–1006.

Durre, I., B. E. Gleason, T. G. Houston, and R. S. Vose. 2010. Robust automated quality control of daily surface observations. Journal of Applied Meteorology and Climatology 49:1615–1633.

Forcey, G. M., W. E. Thogmartin, G. M. Linz, and P. C. McKann. 2014. Land use and climate affect black tern, northern harrier, and marsh wren abundance in the Prairie Pothole Region of the United States. The Condor 116:226–241.

Fuentes, M., T. G. F. Kittel, and D. Nychka. 2006. Sensitivity of ecological models to their climate drivers: statistical ensembles for forcing. Ecological Applications 16:99–116.

Gandin, L. S. 1965. Objective analysis of meteorological fields. Pages 242. Translated from Russian, Jerusa-lem, Israel Program for Scientific Translations, Gidrometeoizdat, Leningrad.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux. 2008. Performance metrics for climate models. Journal of Geophysical Research: Atmospheres 113:D06104. http://dx.doi.org/10.1029/2007JD008972.

Harvell, C. D., C. E. Mitchell, J. R. Ward, S. Altizer, A. P. Dobson, R. S. Ostfeld, and M. D. Samuel. 2002. Climate warming and disease risks for terrestrial and marine biota. Science 296:2158–2162.

Higgins, R. W., W. Shi, E. Yarosh and R. Joyce. 2000. Improved United States precipitation quality control system and analysis. NCEP/Climate Prediction Center ATLAS No. 7. Pages 40. Camp Springs, MD 20746, USA.

Hocevar, A., and J. D. Martsolf. 1971. Temperature distribution under radiation frost conditions in a central Pennsylvania valley. Agricultural Meteorology 8:371–383.

Jentsch, A., J. Kreyling, J. Boettcher-Treschkow, and C. Beierkuhnlein. 2009. Beyond gradual warming: extreme weather events alter flower phenology of European grassland and heath species. Global Change Biology 15:837–849.

Johns, C. J., D. Nychka, T. G. F. Kittel, and C. Daly. 2003. Infilling sparse records of spatial fields. Journal of the American Statistical Association 98:796–806.

Jones, P. G. and A. Gladkov. 2003. FloraMap: a computer tool for predicting the distribution of plants and other organisms in the wild. Version 1.02, Centro Internacional de Agricultura Tropical, Cali, Colombia.

Journel, A., and C. Huijbregts. 1978. Mining geostatistics. Academic Press, New York, USA.

Kalnay, E., et al. 1996. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society 77:437–471.

Kampe, T. U., B. R. Johnson, M. Kuester, and M. Keller. 2010. NEON: the first continental-scale ecological observatory with airborne remote sensing of vegetation canopy biochemistry and structure. Journal of Applied Remote Sensing. 4(043510): http://dx.doi.org/10.1117/1.3361375.

Kittel, T. G. F., N. A. Rosenbloom, T. H. Painter and D. S. Schimel, and VEMAP Modelling Participants. 1995. The VEMAP integrated database for modeling United States ecosystem/vegetation sensitivity to climate change. Journal of Biogeography 22:857–862.

Koenig, W. D. 2002. Global patterns of environmental synchrony and the Moran effect. Ecography 25:283–288.

Livneh, B., E. A. Rosenberg, C. Lin, B. Nijssen, V. Mishra, K. M. Andreadis, E. P. Maurer, and D. P. Lettenmaier. 2013. A long-term hydrologically based data set of land surface fluxes and states for the conterminous United States: update and extensions. Journal of Climate 26:9384–9392.

Maurer, E., A. Wood, J. Adam, D. Lettenmaier, and B. Nijssen. 2002. A long-term hydrologically based data set of land surface fluxes and states for the conterminous United States. Journal of Climate 15:3237–3251.

Menne, M. J., I. Durre, R. S. Vose, B. E. Gleason, and T. G. Houston. 2012. An overview of the Global Historical Climatology Network-Daily Database. Journal of Atmospheric and Oceanic Technology 29:897–910.

Oke, T. R. 1995. The heat island characteristics of the urban boundary layer: characteristics, causes and effects. Pages 81–107 *in* J. E. Cermak, A. G. Davenport, E. J. Plate and D. X. Viegas, editors. Wind climate in cities. Kluwer Academic, Netherlands.

Oyler, J. W., A. Ballantyne, K. Jencso, M. Sweet, and S. W. Running. 2015. Creating a topoclimatic daily air temperature data set for the conterminous United States using homogenized station data and remotely sensed land skin temperature. International Journal of Climatology 35:2258–2279.

Parra, J. L., C. C. Graham, and J. F. Freile. 2004. Evaluating alternative data sets for ecological niche models of birds in the Andes. Ecography 27:350–360.

Polgar, C. A., and R. B. Primack. 2011. Leaf-out phenology of temperate woody plants: from trees to ecosystems. New Phytologist 191:926–941.

Post, E., and M. C. Forchhammer. 2002. Synchronization of animal population dynamics by large-scale climate. Nature 420:168–171.

Rienecker, M. M., et al. 2011. MERRA: NASA's modern-era retrospective analysis for research and applications. Journal of Climate 24:3624–3648.

Schwartz, M. D. 1998. Green-wave phenology. Nature 394:839–840.

Shepard, D. 1968. A two dimensional interpolation function for irregularly spaced data. Pages 517–524. Proceedings of the 23rd National Conference of the Association for Computing Machinery, Princeton, NJ. ACM.

Shepard, D. S. 1984. Computer mapping: the SYMAP interpolation algorithm. Spatial Statistics and Models 40:133–145.

Sheridan, P., S. Smith, A. Brown, and S. Vosper. 2010. A simple height-biased correction for temperature downscaling in complex terrain. Meteorological Applications 17:329–339.

Sillmann, J., V. V. Kharin, X. Zhang, F. W. Zwiers, and D. Bronaugh. 2013a. Climate extremes indices in the CMIP5 multimodel ensemble: Part 1: model evaluation in the present climate. Journal of Geophysical Research: Atmospheres 118:1716–1733.

Sillmann, J., V. V. Kharin, F. W. Zwiers, X. Zhang, and D. Bronaugh. 2013b. Climate extreme indices in the CMIP5 multi-model ensemble. Part 2: future climate projections. Journal of Geophysical Research: Atmospheres 118:2473–2493.

Taylor, K. 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres 106:7183–7192.

Thogmartin, W. E., and P. C. McKann. 2014. Large-scale climate variation modifies winter group size in the endangered Indiana bat. Journal of Mammalogy 95:117–127.

Thornton, P., S. Running, and M. White. 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. Journal of Hydrology 190:214–251.

Whitaker, J. S., and T. M. Hamill. 2002. Ensemble data assimilation without perturbed observations. Monthly Weather Review 130:1913–1924.

Wolfe, D., M. Schwartz, A. Lakso, Y. Otsuki, R. Pool, and N. Shaulis. 2005. Climate change and shifts in spring phenology of three horticultural woody perennials in northeastern USA. International Journal of Biometeorology 49:303–309.

Xia, Y., et al. 2012. Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. Journal of Geophysical Research: Atmospheres 117:D03110. http://dx.doi.org/10.1029/2011JD016051.

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

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1890/15-1061/suppinfo

Data Availability

Data associated with this paper have been deposited in Dryad: http://dx.doi.org/10.5061/dryad.7tv80