DOCKETED	
Docket Number:	19-ERDD-01
Project Title:	Research Idea Exchange
TN #:	231300-1
Document Title:	An Assessment of High-Resolution Gridded Temperature Datasets over California
Description:	Peer-reviewed journal paper by Walton et al, published in Journal of Climate
Filer:	Susan Wilhelm
Organization:	California Energy Commission
Submitter Role:	Commission Staff
Submission Date:	12/20/2019 11:46:23 AM
Docketed Date:	12/20/2019

An Assessment of High-Resolution Gridded Temperature Datasets over California

DANIEL WALTON

Institute of the Environment and Sustainability, University of California, Los Angeles, Los Angeles, California

ALEX HALL

Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, Los Angeles, California

(Manuscript received 14 June 2017, in final form 23 December 2017)

ABSTRACT

High-resolution gridded datasets are in high demand because they are spatially complete and include important finescale details. Previous assessments have been limited to two to three gridded datasets or analyzed the datasets only at the station locations. Here, eight high-resolution gridded temperature datasets are assessed two ways: at the stations, by comparing with Global Historical Climatology Network–Daily data; and away from the stations, using physical principles. This assessment includes six station-based datasets, one interpolated reanalysis, and one dynamically downscaled reanalysis. California is used as a test domain because of its complex terrain and coastlines, features known to differentiate gridded datasets. As expected, climatologies of station-based datasets agree closely with station data. However, away from stations, spread in climatologies can exceed 6°C. Some station-based datasets are very likely biased near the coast and in complex terrain, due to inaccurate lapse rates. Many station-based datasets have large unphysical trends ($>1°\mathrm{C}\ \mathrm{decade}^{-1}$) due to unhomogenized or missing station data—an issue that has been fixed in some datasets by using homogenization algorithms. Meanwhile, reanalysis-based gridded datasets have systematic biases relative to station data. Dynamically downscaled reanalysis has smaller biases than interpolated reanalysis, and has more realistic variability and trends. Dynamical downscaling also captures snow–albedo feedback, which station-based datasets miss. Overall, these results indicate that 1) gridded dataset choice can be a substantial source of uncertainty, and 2) some datasets are better suited for certain applications.

1. Introduction

High-resolution gridded temperature datasets are widely used because they are spatially complete and include finescale variations due to topography and other features. Such detail is important for many modeling applications in fields such as hydrology, ecology, and agriculture (Thornton et al. 1997; Mote et al. 2005; Abatzoglou 2013; Stoklosa et al. 2015). Gridded datasets are also used to compute historical trends (e.g., Hamlet and Lettenmaier 2005; Vose et al. 2014), evaluate regional climate models (e.g., Caldwell et al. 2009; Walton et al. 2015), and train statistical models to downscale low-resolution climate information to higher resolution (e.g., Hidalgo et al. 2009; Pierce et al. 2014).

There are a variety of approaches for generating high-resolution gridded temperature data. One approach is to interpolate or smooth data from irregularly spaced stations to a regular grid. Datasets generated in this manner are termed ‘‘station based’’ datasets. A key distinction between these datasets is that some datasets fit smooth temperature curves to the station data (e.g., Thornton et al. 1997; Hijmans et al. 2005), while others use interpolation algorithms that seek to match observations exactly at the station locations (e.g., Maurer et al. 2002; Hamlet and Lettenmaier 2005; Livneh et al. 2013). Some incorporate knowledge of physical processes into the interpolation method, essentially creating a simple model of temperature variations between station locations (e.g., Daly et al. 2008; Vose et al. 2014; Oyler et al. 2015a). A challenge with station data is that changes in station siting, instrumentation, and time of observation add nonclimatic artifacts to the data (Menne and Williams 2009). Some datasets correct for these inhomogeneities (e.g., Hamlet and Lettenmaier 2005; Vose et al. 2014; Oyler et al. 2015a), which makes

Corresponding author: Daniel Walton, waltond@ucla.edu

DOI: 10.1175/JCLI-D-17-0410.1

them better suited for long-term trend analysis. Some datasets also include uncertainty estimates or facilitate calculations of uncertainty (e.g., Oyler et al. 2015a; Newman et al. 2015).

Differences in interpolation algorithms can lead to large differences in climatologies (Simpson et al. 2005; Daly 2006; Stahl et al. 2006; Daly et al. 2008; Mizukami et al. 2014). For example, Daly et al. (2008) compared their dataset, PRISM, to Daymet (Thornton et al. 1997) and WorldClim (Hijmans et al. 2005) over the continental United States. PRISM determines temperatures using a local temperature–elevation relationship calibrated from nearby stations. Stations are given higher weights if they are closer to the target grid cell, and if they have similar coastal proximity or topographic position (among other factors). Daymet also uses stations to determine a local temperature–elevation relationship, but stations are weighted using a truncated Gaussian filter centered at the target grid cell. Meanwhile, WorldClim fits a thin-plate spline to station data to generate a temperature surface. Differences in climatology between these datasets were found to be largest over complex terrain and coastal areas of the western United States. January minimum temperatures (Tmin) in WorldClim and Daymet were found to be have cold biases of 3°–4°C in complex terrain, which Daly et al. concluded were due to failing to account for cold-air pooling. Meanwhile, along the central California coast, WorldClim and Daymet have biases in maximum temperature (Tmax) that likely result from poorly capturing the onshore marine layer, which complicates the relationship between temperature and elevation (Johnstone and Dawson 2010; Iacobellis and Cayan 2013). In contrast, PRISM accounts for coastal proximity and topographic position, which could explain why it outperforms the others in complex terrain and along the coast.

Oyler et al. (2015a) compared PRISM and Daymet to TopoWx. TopoWx is unique because it uses remotely sensed land skin temperature (LST) as an auxiliary predictor. Oyler et al. compared the datasets over the complex terrain of Nevada, where cold air pooling causes inversions in Tmin. TopoWx has the strongest inversions, PRISM has similar but slightly weaker inversions, and Daymet has comparatively smooth temperature variations without inversions. Oyler et al. found that elevation alone is weak predictor of Tmin, explaining only 6% of the variance in this region, while LST explained 77%. This could explain why Daymet—which uses elevation, but does not use LST or physically based station weights—does not capture inversions here.

Previous comparisons have found potential biases in station-based gridded datasets that use fixed lapse rates when accounting for elevation (Mizukami et al. 2014; Newman et al. 2015). Newman et al. (2015) compared their ensemble gridded dataset to that of Maurer et al. (2002; henceforth this dataset is referred to simply as "Maurer"), and noted that Maurer is consistently colder at high elevations. Newman et al. attribute this to the use of a fixed 6.5°C km^{-1} lapse rate in Maurer. Mizukami et al. (2014), also found Maurer to be relatively cold at high elevations.

Often the term "gridded data" is used to mean station-based gridded datasets only. However, there are multiple ways of generating historical data on a regular grid. A second approach is to run an atmospheric model that assimilates historical observations. Datasets constructed in this way are referred to as reanalysis. There are many global or continental-scale reanalysis products that assimilate observations [e.g., NARR, MERRA, NOAA-20CR, CERA-20C, and ERA-20C; for details, see Dee et al. (2016)]. However, the resolutions of these datasets—ranging from 0.3° to 5°—are too low for many applications. Thus, reanalysis is often downscaled to higher resolution (Cosgrove et al. 2003; Kanamitsu and Kanamaru 2007; Rasmussen et al. 2011; Stefanova et al. 2012; Xia et al. 2012; Abatzoglou 2013; Walton et al. 2015, 2017). One straightforward way to downscale reanalysis is with bilinear interpolation. For example, the temperature forcings in the North American Land Data Assimilation System version 2 dataset (NLDAS-2; Xia et al. 2012) are derived by interpolating North American Regional Reanalysis (NARR; Mesinger et al. 2006) to 1/8° resolution. Reanalysis can also be downscaled with a regional climate model, a process referred to as dynamical downscaling. Under this method, a regional climate model is forced at the lateral and ocean surface boundaries by reanalysis. For example, Kanamitsu and Kanamaru (2007) downscaled 200-km resolution NCEP–NCAR global reanalysis (Kalnay et al. 1996) to 10-km resolution over California with the Regional Spectral Model (Juang and Kanamitsu 1994). Similarly, Walton et al. (2015) downscaled the 32-km resolution NARR to 2-km resolution over the Los Angeles region with the Weather Research and Forecasting Model (WRF; Skamarock et al. 2008), and used a similar WRF setup to downscale NARR to 3-km resolution over the Sierra Nevada mountains (Walton et al. 2017).

Previous assessments of gridded datasets have been limited in a variety of ways. Some have only considered station-based datasets and excluded downscaled reanalysis (Daly et al. 2008; Newman et al. 2015; Oyler et al. 2015a). Many have compared only two or three datasets (Daly et al. 2008; Bishop and Beier 2013;

Mizukami et al. 2014; Newman et al. 2015; Oyler et al. 2015a). Behnke et al. (2016a) performed one of the most comprehensive evaluations to date, which considered eight datasets, including interpolated reanalysis, but datasets were only evaluated at station locations. Station-based datasets are constrained to match station data, so only evaluating them at station locations may give a misleading picture of their overall realism. Previous assessments of gridded datasets have excluded dynamically downscaled reanalysis. Dynamically downscaled reanalysis could have an advantage away from stations, to the extent that it realistically simulates physical processes that cause important spatial variations, such as onshore penetration of the marine layer in the coastal zone and cold-air pooling in complex terrain. Station-based datasets either struggle to capture these processes (e.g., Daymet, WorldClim, and Maurer) or attempt to model their effects through auxiliary predictors or physically based weights (e.g., TopoWx and PRISM).

One effect that has not been explored in previous assessments is snow–albedo feedback (SAF). Snow is highly reflective, and reductions in snow cover typically reveal surfaces that absorb more solar radiation, leading to warmer temperatures and further reductions in snow cover (Cubasch et al. 2001; Holland and Bitz 2003). Dynamically downscaling explicitly simulates SAF (Salathé et al. 2008; Letcher and Minder 2015; Walton et al. 2017), but it is unknown whether its effects are captured by station-based datasets. Low station density at high elevations could make it challenging to capture the narrow bands of amplified temperatures associated with SAF (Walton et al. 2017).

This study looks to answer the following questions about high-resolution temperature datasets:

1) How do temperature climatologies, variability, and trends in these datasets differ?
2) Can these differences be explained in terms of their methodological choices?
3) Which datasets are most realistic? While this question can be answered at station locations by comparing with observed data, it is challenging to answer away from stations where there are no observations to rely on. However, in some instances, there are physical arguments as to why some datasets are more realistic.
4) Does dynamically downscaled reanalysis—which explicitly simulates relevant processes (however imperfectly)—corroborate the spatial and temporal variations in station-based datasets? How convergent are these orthogonal approaches of creating gridded temperature data?
5) Are dynamical downscaled reanalysis and interpolated reanalysis equally realistic?

To answer these questions, this study compares eight high-resolution gridded datasets with a long running subset of the Global Historical Climatology Network–Daily (GHCND) stations (Menne et al. 2012a,b). The comparison is performed over California, which has coastal areas with maritime influence, complex terrain experiencing cold-air pooling, and high-elevation mountains with significant seasonal snow cover. The datasets used here are the following:

- PRISM (Daly et al. 2008),
- TopoWx (Oyler et al. 2015a),
- Daymet (Thornton et al. 1997),
- Livneh (Livneh et al. 2013; Maurer et al. 2002),
- Hamlet [an extension of Hamlet and Lettenmaier (2005)],
- Metdata (Abatzoglou 2013),
- NLDAS-2 (Xia et al. 2012), and
- NARR dynamically downscaled with WRF (Walton et al. 2017).

Together, these eight datasets represent the wide range of approaches to creating gridded temperature data discussed above. For a summary of their important features, see Table 1.

This paper is structured as follows. Section 2 provides detailed information about the eight gridded datasets. Section 3 covers the methodology used to assess their climatologies, variability, and trends. Results are given in section 4. Major findings are summarized and discussed in section 5.

2. Data

a. *GHCND station data*

California has 847 GHCND stations with some daily data during the 1981–2010 period (Fig. 1a). These include National Weather Service (NWS) Cooperative Observer Program (COOP) stations, Weather Bureau Army Navy (WBAN) stations, National Resource Conservation Service (NRCS) snow telemetry (SNOTEL) and snow course sites, and U.S. Forest Service and Bureau of Land Management (BLM) Remote Automatic Weather Stations (RAWS). Only a fraction of these stations have a sufficiently long record to reliably calculate climatologies and variability. So, we use a subset of 223 stations with at least 83% coverage during this period (Fig. 1b) as determined by Behnke et al. (2016b) and made available via the Dryad data package [see links in Behnke et al (2016b)].

b. *PRISM*

The Parameter–Elevation Relationships on Independent Slopes Model (PRISM; Daly et al. 1994,

TABLE 1. Details of gridded datasets used in this study.

	PRISM (AN81m)	TopoWx	Daymet	Livneh	Hamlet	Metdata	NLDAS-2	WRF
Category	Station based	Station based	Station based	Station based	Station based	Hybrid (uses station-based monthly data and reanalysis-based subdaily data)	Interpolated reanalysis	Dynamically downscaled reanalysis
Citation	Daly et al. (2008)	Oyler et al. (2015a)	Thornton et al. (1997)	Livneh et al. (2013)	Hamlet and Lettenmaier (2005)	Abatzoglou (2013)	Xia et al. (2012)	Walton et al. (2017)
Data available from	http://prism.oregonstate.edu	http://www.scrimhub.org/resources/topowx/	https://daymet.ornl.gov	https://ciresgroups.colorado.edu/livneh/data/daily-observational-hydrometeorology-data-set-conus-extent-canadian-extent-columbia-river-basin	Contact Mu Xiao muxiao@ucla.edu for extension; original available from http://www.hydro.washington.edu/Lettenmaier/Data/gridded/index_hamlet.html	http://climate.nkn.uidaho.edu/METDATA/	https://disc.gsfc.nasa.gov/SSW	http://research.atmos.ucla.edu/csrl/data/Gridded_Datasets/
Native resolution	2.5′ (∼4 km)	30 arc sec (∼800 m)	1 km	1/16° (∼6 km)	Extension: 1/16° (∼6 km)	1/24° (∼4 km)	1/8° (∼12 km)	9 km
Time period	1895–2016	1948–2016	1980–2016	1915–2011	1915–2016	1979–2016	1979–2016	1995–2015
Input data for temperature	COOP, WBAN, SNOTEL, RAWS, CDEC, Agrimet, and others	GHCND, SNOTEL, and RAWS	GHCND	COOP	COOP, Environment Canada, USHCN, and Historical Canadian Climate Database (HCCD)	NLDAS-2 for daily variability and PRISM for monthly means	NARR	NARR

Table 1. (*Continued*)

	PRISM (AN81m)	TopoWx	Daymet	Livneh	Hamlet	Metdata	NLDAS-2	WRF
Adjustments for temporal inhomogeneities	No	Yes, pairwise comparison algorithm of Menne and Williams (2009)	No	No	Yes, low-frequency variability adjusted to match USHCN, HCCD stations	No	No	No
Downscaling/ interpolation method	Elevation-regression model with stations weighted based on multiple physical factors	Regression kriging for climate normals with auxiliary predictors including lat, lon, elev, satellite LST; moving-window geographically weighted regression and inverse distance weighting for anomalies	Truncated Gaussian filter combined with elevation regression	SYMAP algorithm: inverse distance weighting with directional adjustment	SYMAP algorithm: inverse distance weighting with directional adjustment; daily data adjusted so that monthly climate normals match PRISM for 1971–2000	Bilinearly interpolated NLDAS-2 daily anomalies from monthly means added to PRISM monthly means	Bilinear interpolation of NARR to 1/8° spatial resolution and hourly temporal resolution	WRF coupled to Noah-MP
Lapse rate	Determined based on nearby stations	Determined based on nearby stations	Determined based on nearby stations	6.5°C km^{-1}	6.1°C km^{-1}, but climatologies adjusted to match PRISM	Variable	Variable	Variable

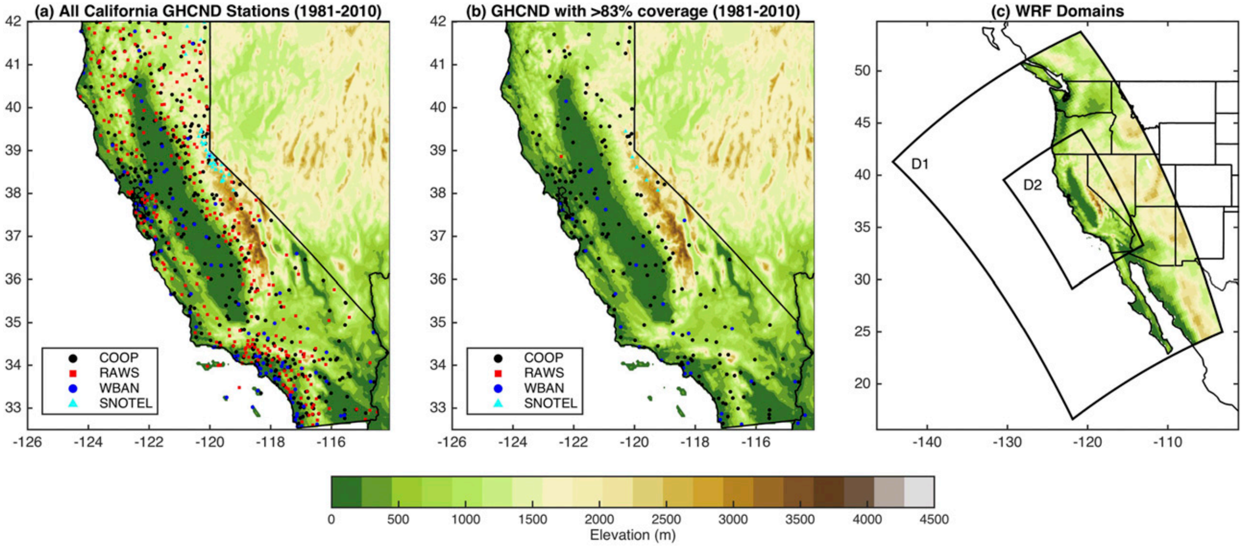

FIG. 1. (a) All California GHCND stations with at least some data during the 1981–2010 period. GHCND includes stations from the COOP, RAWS, WBAN, and SNOTEL networks. (b) GHCND stations with at least 83% coverage from 1981–2010. (c) Setup of 27-km and 9-km resolution one-way nested WRF domains.

2008) is a modeling system used to derive gridded temperature and precipitation data for the conterminous United States. At each grid cell, an elevation regression function is fit to station data using a moving window. Stations are weighted depending on multiple physical factors that reflect their similarity to the target grid cell. These factors include distance, cluster, elevation, coastal proximity, topographic facet, vertical layer, topographic position, and effective terrain height. Here we use the monthly dataset AN81m with 2.5-min (~4 km) resolution (PRISM Climate Group, Oregon State University; available from http://prism.oregonstate.edu; data created between 9 June 2013 and 9 June 2014). Although station data are subjected to quality control procedures, no adjustments are made to ensure temporal homogeneity in this PRISM dataset. PRISM incorporates data from ~10 000 stations spanning multiple networks, including COOP, RAWS, the California Data Exchange Center (CDEC), Agrimet, NRCS, the California Irrigation Management Information System (CIMIS), and more (see http://prism.oregonstate.edu for details). Many of these networks are part of GHCND (Fig. 1a).

c. TopoWx

TopoWx or "Topography Weather" is a gridded dataset of daily Tmin and Tmax based on station data and remotely sensed land skin temperature (Oyler et al. 2015a; data downloaded from http://www.scrimhub.org/resources/topowx/). TopoWx covers the conterminous United States at 30 arc sec (~800-m resolution) for the period 1948–2016. TopoWx uses station data from GCHND stations (Fig. 1a). TopoWx applies the homogenization algorithm of Menne and Williams (2009) to correct for changes in observation practices, siting, and instrumentation. Missing values are filled by comparing with nonmissing neighboring observations and applying spatial regression (Durre et al. 2010). Climate normals are computed using regression kriging, and daily anomalies are computed using moving window geographically weighted regression and inverse distance weighting. To help estimate climate normals in complex terrain and regions with low station density, TopoWx uses remotely sensed land skin temperature as an auxiliary predictor.

d. Daymet

Daymet (Thornton et al. 1997) is a dataset of daily meteorological variables on a 1 km × 1 km grid covering North America for the period 1980–2016. Version 3 (Thornton et al. 2016) is used here. Monthly summaries of daily Tmax and Tmin were downloaded from the Thematic Real-Time Environmental Distributed Data Services (THREDDS) server (http://thredds.daac.ornl.gov/thredds/catalogs/ornldaac/Regional_and_Global_Data/DAYMET_COLLECTIONS/DAYMET_COLLECTIONS.html) on 9 January 2017. Daymet fits a smooth curve to data from GHCND stations to a 1 km × 1 km grid using a weighted average of nearby stations. Weights are determined by a truncated Gaussian filter centered at the target grid cell. The radius of the Gaussian filter varies continuously throughout the domain to adjust for varying station density. Tmax and

Tmin values are adjusted for elevation using a linear temperature–elevation relationship.

e. Livneh

The Livneh et al. (2013) dataset (this dataset is hereafter simply called ‘‘Livneh’’) contains station-based meteorological variables and modeled hydrologic variables that covers the conterminous United States at 1/16° (~6 km) resolution for the period 1915–2011. (Livneh data are available from https://ciresgroups.colorado.edu/livneh/data/daily-obserational-hydrometeorology-dataset-conus-extent-canadian-extent-columbia-river-basin.) Livneh is an extension and upgrade to the Maurer et al. (2002) dataset, which used a similar methodology but spanned the shorter 1950–2000 period at a lower resolution of 1/8° (~12 km). Livneh temperatures are created by gridding data from COOP weather stations over the conterminous United States. Gridding is performed on station temperature data via the synergraphic mapping system (SYMAP; Shepard 1984). Under SYMAP, for a grid point, the temperature is calculated as a weighted average of the temperature at the four nearest stations. The weights are determined by a combination of inverse distance weighting and down-weighting stations that are close to other stations. For a full description of the gridding procedure, the reader is referred to Livneh et al. (2013) and Maurer et al. (2002).

f. Hamlet

The original Hamlet and Lettenmaier (2005) dataset spans 1915–2003 at 1/8° (~12 km) resolution (data available from http://www.hydro.washington.edu/Lettenmaier/Data/gridded/index_hamlet.html). It has now been extended to cover 1915–2015, its resolution has been increased to 1/16° (~6 km), and temperatures are now adjusted so that 1971–2000 climate normals match PRISM. This extension, henceforth simply ‘‘Hamlet,’’ was provided by Mu Xiao of UCLA. Hamlet generally follows the Maurer methodology of interpolating daily COOP station data using the SYMAP algorithm. The two major differences are that 1) Hamlet temperatures are adjusted so 1971–2000 monthly normals match PRISM and 2) low-frequency variability matches the quality-controlled U.S. Historical Climatology Network (USHCN; Menne et al. 2009) stations. The use of quality-controlled stations to determine low-frequency variability is intended to make the Hamlet dataset suitable for trend analysis and long-term hydrologic simulations. This extension appears to be similar to the extension created by Hamlet et al. (2010).

g. WRF

This dataset is a dynamical downscaling of 32-km resolution NCEP North American Regional Reanalysis (Mesinger et al. 2006) for the 1981–2015 period using version 3 of the Weather Research and Forecasting Model (Skamarock et al. 2008) performed by Walton et al. (2017). Under this setup, WRF is forced at the lateral and ocean surface boundaries by NARR. WRF is coupled to the Noah-MP land surface model (Niu et al. 2011). WRF is arranged in a one-way nested setup with a 27-km resolution domain covering the western United States and northeastern Pacific Ocean, a 9-km domain covering California, and a 3-km domain covering the Sierra Nevada. This study focuses on the 9-km domain covering California [Fig. 1c (indicated as D2)]. A cubic spline fit to WRF 3-hourly output is used to calculate daily Tmax and Tmin.

h. NLDAS-2

This dataset is the historical forcing for the North American Land Data Assimilation System (NLDAS; Cosgrove et al. 2003; Mitchell et al. 2004), which includes temperature data with 1-h temporal resolution and 1/8° spatial resolution. The most recent version of the project, NLDAS-2 (Xia et al. 2012), linearly interpolates 32 km, 3-hourly NARR temperature data in space and time to achieve 1/8°, 1-hourly data for the period 1979–2016. So, like the WRF simulation, NLDAS-2 is a downscaling of NARR, but using linear interpolation instead of a regional climate model. Data were downloaded using the NASA Earthdata Simple Subset Wizard (https://disc.gsfc.nasa.gov/SSW/).

i. Metdata

Metdata (Abatzoglou 2013) is a hybrid dataset of meteorological forcings that combines the subdaily temporal resolution of NLDAS-2, with the spatial climatologies and monthly variability of PRISM. Metdata is available for the 1979–2016 period at 4-km horizontal resolution from http://metdata.northwestknowledge.net. To create Metdata, NLDAS-2 subdaily anomalies (relative to monthly means) are interpolated to 4-km resolution and added to PRISM monthly means. Because this study analyzes monthly data and Metdata's monthly variability comes from PRISM, Metdata is grouped here with the station-based datasets.

3. Methods

a. Regridding to the WRF 9-km grid

To facilitate comparisons among the datasets, each dataset is regridded to the 9-km WRF grid. For TopoWx and Daymet, which have substantially higher resolution than WRF, regridding is performed using a moving window approach: averages are taken over all grid cells whose centers reside within the nearest WRF grid cell. For all other datasets, regridding is performed with bilinear interpolation. Only land areas are considered as some datasets do not have data over oceans or lakes. All analysis is performed over the 1981–2010 period. For comparisons with GHCND station data, the nearest grid cell in the regridded dataset is used. To adjust for elevation differences between GHCND station locations and the nearest WRF grid cell, a lapse rate of 6.5°C km^{-1} is used. This adjustment is only made for Tmax. No adjustment is made for Tmin, because Tmin differences were found to be only weakly correlated with elevation differences.

b. Climatologies

Annual climatologies are computed for each gridded dataset. Climatologies are displayed two ways: as differences relative to the GHCND station data, and as differences relative to the average of the station-based gridded datasets. Station data are not without error, but collectively they represent our best primary source of temperature data. Thus, if a gridded dataset has large differences with many stations, then the gridded dataset is probably biased. Meanwhile, differences with the station-based gridded dataset average are not used to detect biases, necessarily, but they do show how the gridded datasets compare to each other. Importantly, these differences are spatially complete—unlike the differences with GHCND stations data—so they reveal how the datasets compare to each other away from the stations.

c. Linear trends

Linear trends are computed at each grid cell using least squares linear regression on the full sequence of monthly anomalies (all 360 months in the 1981–2010 period). This is too short a period to draw inferences about overall historical trends in temperatures. Instead, this analysis is intended to highlight differences in trends between the datasets. Important differences are expected between datasets based on whether they account for inhomogeneities in the data. Linear trends are also computed for the GHCND station data, using all nonmissing monthly anomalies.

d. Variability

To compare temperature variability, the standard deviation of the full sequence of monthly temperature anomalies is computed for the period 1981–2010 at each grid cell. Variability is also computed for GHCND station data, using all nonmissing monthly anomalies. For a deeper investigation into spatial covariability, empirical orthogonal function (EOF) analysis is performed on the full sequence of monthly anomalies. EOFs (spatial patterns) represent the primary modes of spatial covariability within the domain. The corresponding principal components (PCs) are time series that represent how these patterns are scaled up and down in time. The three leading EOFs are compared, along with their principal components.

e. Snow–albedo feedback

To test for SAF, April temperature differences are computed between 2007, a warm year with low snow cover, and 2010, a cold year with high snow cover. April snow cover differences are computed for WRF and remotely sensed data from the Moderate-Resolution Imaging Spectroradiometer onboard the *Terra* satellite (MODIS/*Terra* Snow Cover Monthly L3 Global 0.05 CMG; Hall et al. 2006; data available from http://nsidc.org/data/MOD10CM). Comparing temperature and snow cover differences will allows us to determine whether WRF and the other datasets have similarly amplified temperature differences due to SAF in narrow bands where snow cover is lost.

f. Surface lapse rates

Coastal areas and complex terrain in California may be subject to inverted temperature profiles from penetration of the marine layer and cold-air pooling (Lundquist et al. 2008; Daly et al. 2010). If interpolation algorithms do not account for the complicated relationships between temperature and elevation, then they may produce errant temperature patterns. Here we examine surface lapse rates in three representative datasets: TopoWx, which uses satellite LST as an auxiliary predictor and has been shown to better capture Tmin in complex terrain (Oyler et al. 2015a); PRISM, which explicitly incorporates physical factors like coastal proximity in its regression weights; and Livneh, which uses a fixed lapse rate of 6.5°C km^{-1}. To calculate the surface lapse rate at each grid cell, linear regression is applied to temperature and elevation data from surrounding grid cells (defined as grid cells within two grid lengths in the x or y direction).

In addition, the topographic dissection index (TDI; Holden et al. 2011) is used to determine where stations are located relative to local topographic minima and maxima. Here we use the TDI computed by Oyler et al. (2015a) on the 800-m TopoWx grid, which uses five spatial windows ($n = 5$) with sizes 3, 6, 9, 12, and 15 km. With this setup, TDI values range from 0 to 5, with 0 being a multiscale local minimum and 5 being a multiscale local maximum. A station's TDI is taken to be the TDI at the grid cell closest to that station. Knowing a station's TDI tells us whether a station's nearby grid

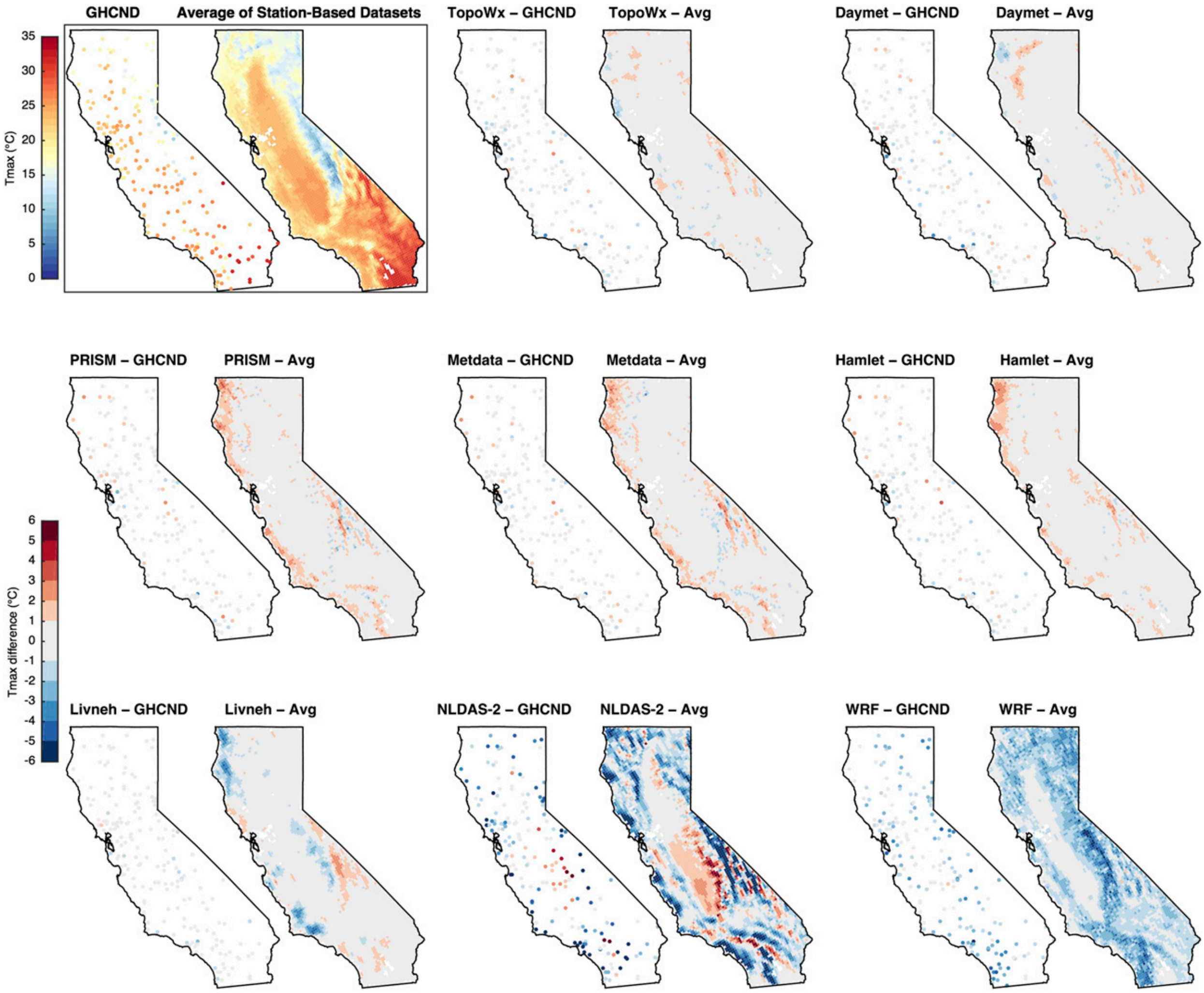

FIG. 2. Shown at top left is 1981–2010 Tmax annual-mean climatology (°C) at GHCND stations and averaged over the station-based datasets. The remainder of the panels shows differences in 1981–2010 annual-mean Tmax climatology with GCHND station data and with the station-based dataset average (°C). To adjust for the elevation differences between the GCHND stations and the nearest grid cell, a lapse rate of 6.5°C km^{-1} was used.

cells are generally above or below it, which is useful for understanding how lapse rates are applied.

4. Results

a. Climatologies

For Tmax, the station-based datasets match GHCND within 1°C at nearly all stations (Fig. 2). Most everywhere, the station-based datasets are similar (within 1°C of the station-based average). As expected, PRISM, Metdata, and Hamlet have nearly identical climatologies. This is no surprise because Metdata is built on PRISM monthly data, and Hamlet is adjusted to match PRISM normals for 1971–2000. These three datasets tend to have warmer than average Tmax values on the windward side of the coastal mountains by up to 3°C, likely because PRISM has inverted Tmax conditions along the coast (discussed further in section 4e). Meanwhile, Livneh Tmax is colder than average in the higher elevations of the coastal mountains by up to 4°C. Interestingly, comparing at the station locations, there is little indication that Livneh diverges from the other datasets; it is only revealed through a spatially complete comparison. This highlights the importance of comparing station-based datasets everywhere, not just at station locations. The reanalysis-based datasets (NLDAS-2 and WRF) are substantially cooler throughout the domain. On average, NLDAS-2 and WRF are colder than the station-based gridded average by 1.4° and 1.1°C,

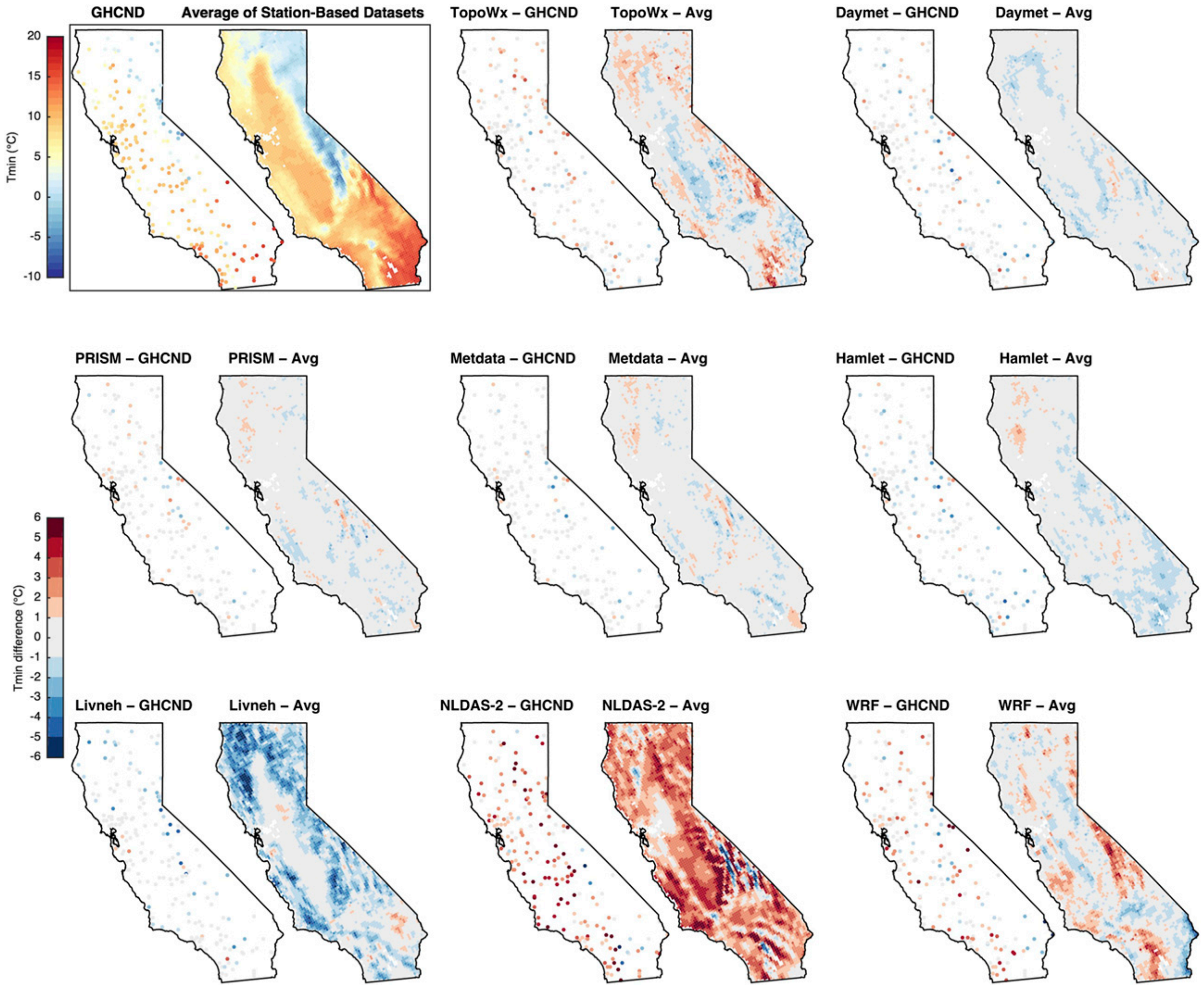

FIG. 3. As in Fig. 2, but for Tmin. Note that no elevation-based adjustments are made for Tmin.

respectively. They are also consistently colder than GHCND data (by 1.8° and 1.6°C, respectively), so it is highly likely that they have a cold bias. WRF's cold bias appears to be related to elevation ($r = -0.67$) and with a slope of approximately $-1.0°\mathrm{C\,km^{-1}}$ (based on least squares linear regression). NLDAS-2 shows dramatic differences with the other datasets along the edges of topographic features and along the coast, exceeding 6°C in some cases. Although both WRF and NLDAS-2 are derived by downscaling NARR, they have large differences in their climatologies, indicating that the choice of downscaling technique is important.

For Tmin, the station-based datasets agree closely with GHCND data (within 1°C) at most stations (Fig. 3). Differences are larger near strong terrain gradients, such as those along the western side of the Sierra Nevada. These discrepancies could be due to elevation mismatches between the stations and the WRF grid, as no elevation adjustments were made to Tmin (adjustments were made only for Tmax). TopoWx and Livneh are the station-based datasets that differ most from the average. Unlike the others, TopoWx uses satellite LST as a predictor for Tmin, which could explain why it differs. Livneh is clearly the most different and is colder than average by 2°–6°C in areas of complex terrain, such as the coastal mountains of Northern California. This likely is due to Livneh's use of a fixed lapse rate, which is examined in more detail in section 4e. WRF agrees closely with the station-based dataset average over most of the domain (domain-average difference of +0.3°C). It does differ in a few areas, such as along the eastern California border with Arizona, where it is 3°–4°C colder, and on the lee sides of the several mountain complexes, where it is 2°–5°C warmer. In contrast, NLDAS-2 has a strong warm bias throughout the domain when compared with GHCND data and is much

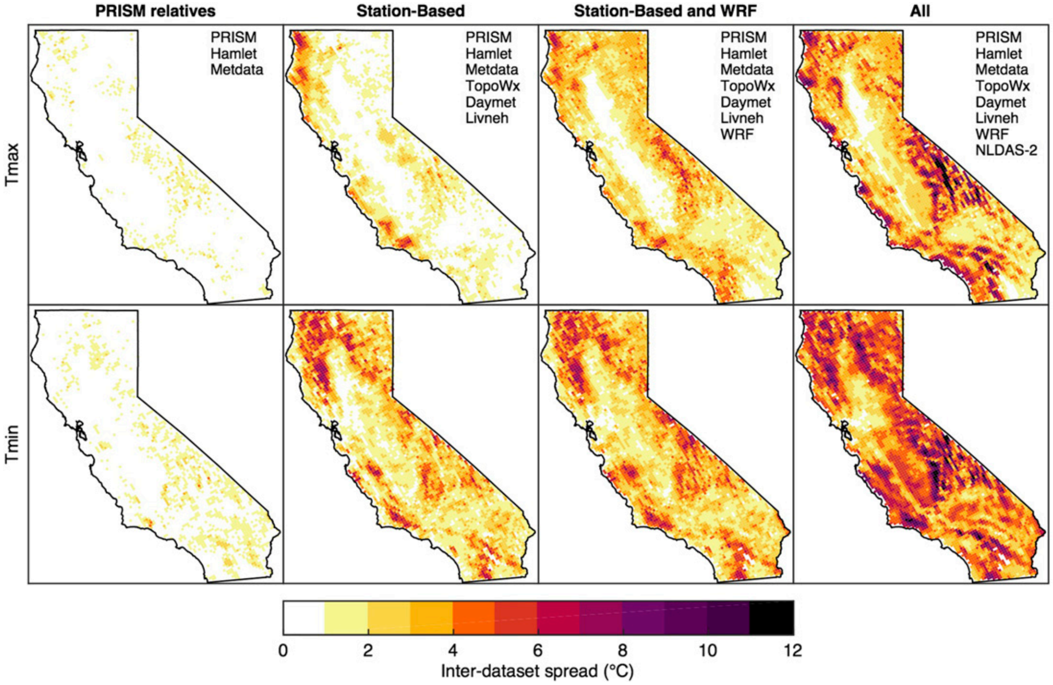

Fig. 4. Interdataset spread (°C) in climatological (top) Tmax and (bottom) Tmin calculated for four different groups. Datasets included in each group are listed in the top-right corner of each panel.

warmer than the average (domain-average difference of +2.9°C). Thus, WRF has a more realistic Tmin climatology than NLDAS-2.

Interdataset spread varies dramatically based on which datasets are considered (Fig. 4). The spread in Tmax among PRISM relatives (PRISM, Hamlet, and Metdata) is small (domain average of 0.5°C). This makes sense as Hamlet is adjusted to match PRISM's climatology, and Metdata is constructed using PRISM's monthly mean values. It becomes larger, especially in the coastal mountains, when all station-based datasets are included (domain average of 1.3°C). When WRF is included, the domain-average spread increases to 2.3°C, with greater spreads at high elevations. When NLDAS-2 is included, spreads increase further, to 3.5°C. A similar progression happens for Tmin: 0.8°C for PRISM relatives, 2.5°C for all station-based, 3.0°C for station-based and WRF, and 4.8°C for all datasets. When all datasets are included, certain locations have extreme spreads (up to 12°C), especially along strong topographic gradients, where NLDAS-2 differs sharply from the others.

b. *Trends*

Linear trends in Tmax and Tmin differ substantially among the datasets (Fig. 5). There are clear differences in trends between those that use homogenized and unhomogenized station data. Daymet, Livneh, PRISM, and Metdata use unhomogenized data and have large trends, exceeding 1°C decade^{-1} in some locations. In contrast, TopoWx and Hamlet correct for inhomogeneities and have smooth trend fields free of nonclimatic artifacts. The reanalysis-based datasets, WRF and NLDAS-2, also have smooth trends fields, although NLDAS-2 has a large trend (up to 1°C decade^{-1}) in central California that is inconsistent with the homogenized datasets, and likely unphysical.

Two primary types of inhomogeneities are present in unhomogenized gridded datasets. The first type is due to missing data or changes in data availability. For example, Fig. 6b shows a location where Daymet has large jumps (up to 10°C) corresponding to when the closest GHCND station (a RAWS station) station goes on and offline. The second type is due to inhomogeneities inherited from the station data. For example, Fig. 6c shows a case where Livneh has large inhomogeneities (3°–4°C) that appear to be inherited from the nearest COOP station. Livneh uses an inverse distance weighting scheme that causes it to very closely match station data near station locations, more closely than other station-based datasets, which could explain why its trend field looks so similar to the station trends (Fig. 5).

For some datasets, the inhomogeneities are systematic and can be seen in the California average. Figure 7 shows California-average monthly anomalies relative to TopoWx, a homogenized dataset likely to have more trustworthy trends. Unhomogenized datasets (Daymet,

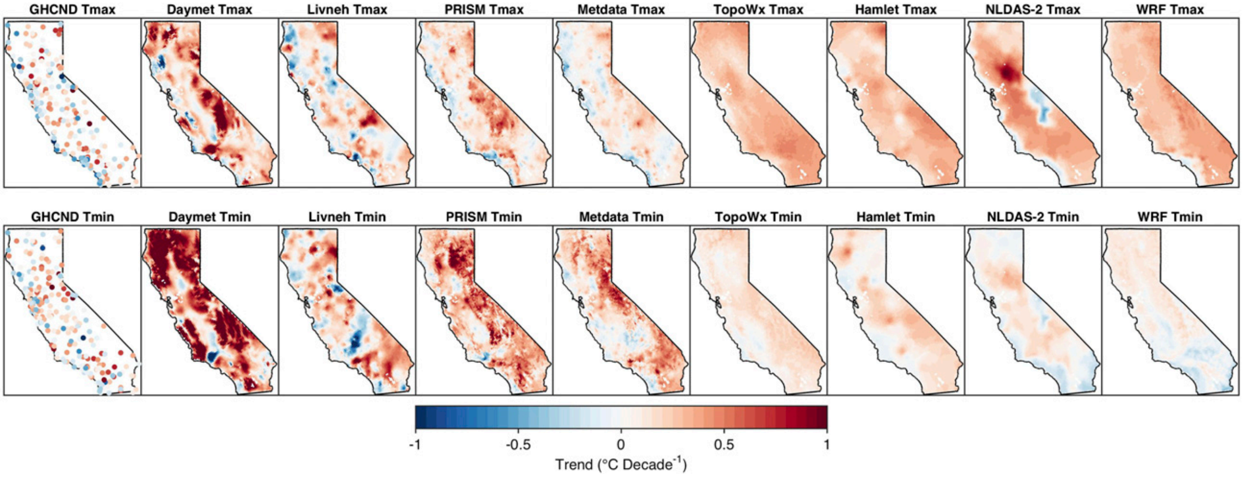

FIG. 5. Trend (°C decade^{-1}) in (top) Tmax and (bottom) Tmin based on linear regression of monthly anomalies for all months in 1981–2010 time period. For GHCND, only anomalies from nonmissing months are used.

Livneh, PRISM, and Metdata) have cold trends in Tmax relative to TopoWx (Fig. 7a). This could be due in part to the well-known transition in instrumentation from liquid-in-glass (LiG) thermometers to maximum–minimum temperature systems (MMTS), which had a cooling effect on Tmax values (Menne et al. 2009). In contrast, there are warm trends in Tmin for the nonhomogenized datasets relative to TopoWx (Fig. 7b). Although there are nonclimatic warm trends at SNOTEL stations (Oyler et al. 2015b), there is almost certainly another factor at play here, since there are so few SNOTEL stations in California and they are confined to

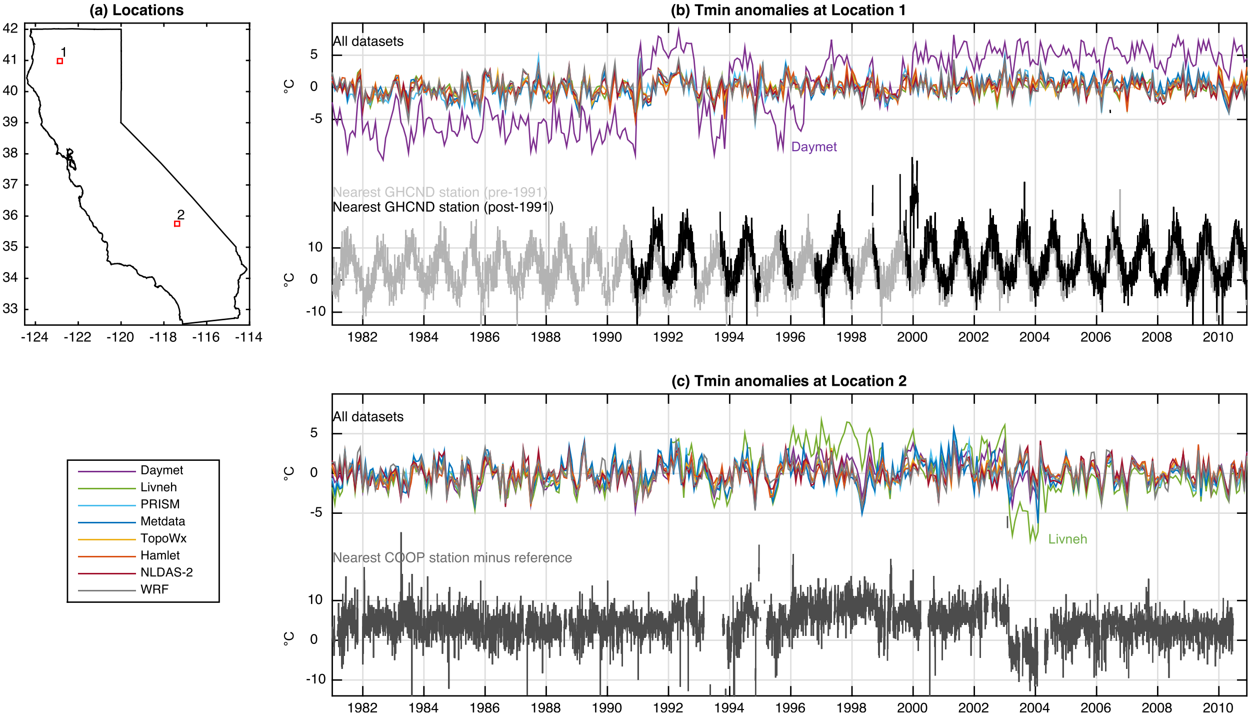

FIG. 6. Exploration of Tmin anomalies (°C) at selected grid cells where gridded datasets show inhomogeneities. (a) Locations of the two grid cells (40.9871°N, 122.8463°W and 35.7605°N, 117.3811°W). (b) Monthly Tmin anomalies at location 1 (colored lines) and raw daily Tmin values at two nearby GHCND stations. The first (gray line) is the closest station to location 1 prior to 1990, and the second (black line) is the closest from 1990 onward. (c) Monthly Tmin anomalies at location 2 (colored lines) and differences in daily Tmin between the nearest COOP station and a nearby reference station (dark gray line).

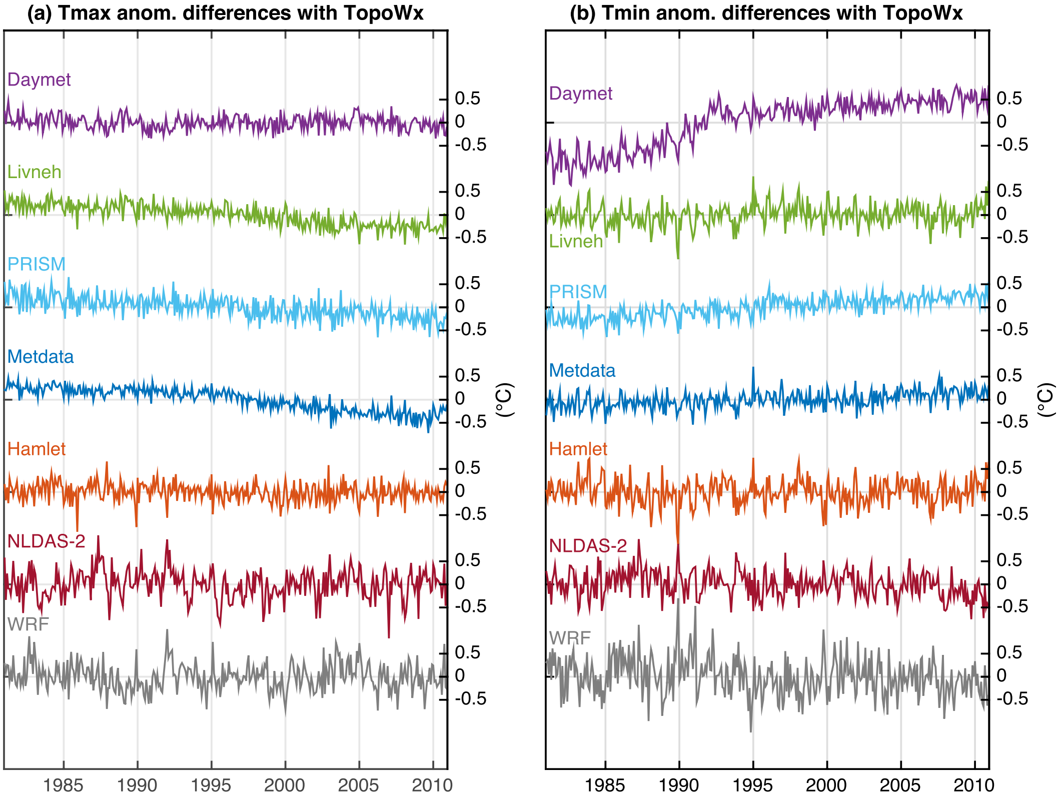

FIG. 7. California-average (a) Tmax and (b) Tmin monthly anomaly differences with TopoWx, for the period 1981–2010.

small areas of the domain (Fig. 1). A more influential factor may be the LiG to MMTS transition, which had a warming effect on Tmin (Menne et al. 2009). Daymet has the largest relative trend in California-average Tmin, with an increase of over 1°C between 1985 and 1992 alone (Fig. 7b). Based on our analysis, Daymet grid cells at high elevations experience similar issues to those of location 1 in Fig. 6b, namely that changes in data availability cause large jumps, particularly when a nearby station comes online. We suspect that the introduction of RAWS stations (320 stations, starting in 1985) could explain Daymet's large trend (Fig. 8).

c. *Variability*

All datasets have greater temperature variability at higher elevations (Fig. 9). In most datasets, Tmax variability peaks in the high elevations of Sierra Nevada, in the range of 2°–3°C. At lower elevations, Tmax variability is in the range of 1°–2°C. NLDAS-2 has much lower Tmax variability (0.5°–1°C) along a wider coastal strip than GHCND or any of the other gridded datasets. Because NLDAS-2 differs so consistently from GHCND along the coast, it is almost definitely biased there. Grid cells in this coastal strip likely reside between land and ocean grid cells in NARR. Thus, when linear interpolation is applied, grid cells in this strip have temperatures with intermediate properties that are mixture between land and ocean. Since temperature variability is lower over the ocean, these grid cells are likely to have lower variability than their inland counterparts.

Tmin variability is lower than Tmax variability in all datasets. For most datasets, Tmin variability is generally in the 1–1.5°C range at low elevations and in the 1.5°–2°C at higher elevations. TopoWx and Hamlet have the least Tmin variability, probably in part because the apply homogenization algorithms that remove nonclimatic jumps. NLDAS-2 has lower Tmin variability along the coast, just like it does for Tmax. Meanwhile, Daymet and Livneh have Tmin variability as high as 3°C, which is likely due to the inhomogeneities that lead to large trends at these locations.

Generally, the datasets have very similar spatial patterns (EOFs) and nearly identical time series (PCs) for the major modes of variability. For Tmax, EOF1 explains between 78% and 86% of the variance, depending on the dataset (Fig. 10). EOF1 is characterized by positive loadings over all of California, with larger loadings

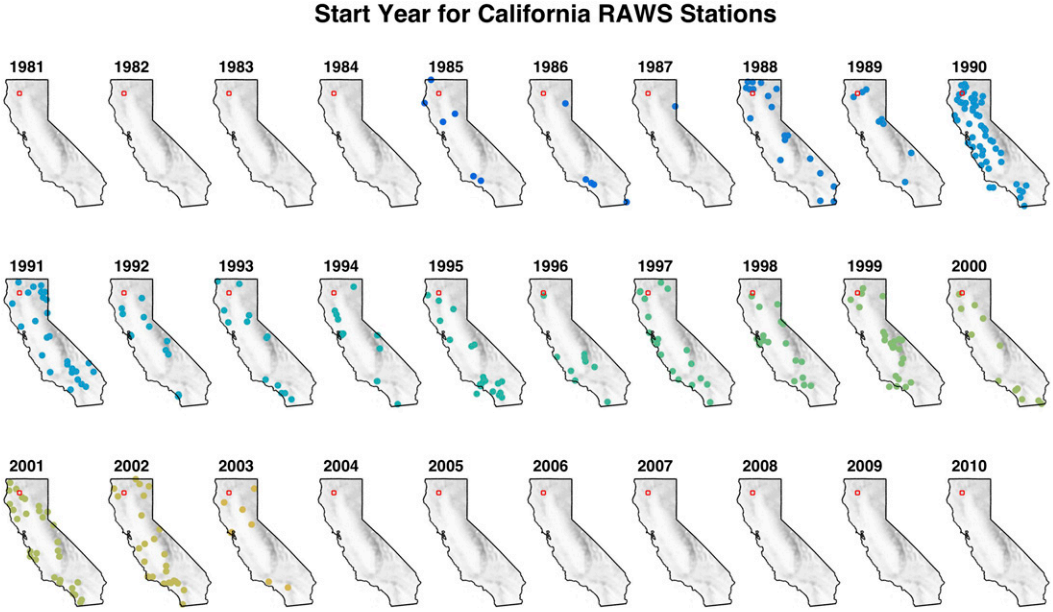

FIG. 8. Year that each California RAWS station came online (colored dots). The red square marks location 1 from Fig. 6, the grid cell where Daymet has a jump in 1990.

at high elevations. PC1 (the time series representing how EOF1 is scaled up or down each month) is nearly identical for each dataset. One notable difference is that Livneh, Hamlet, and NLDAS-2 have EOFs that do not follow topographic contours as closely as the other datasets. NLDAS-2 is also has much weaker loadings along the coast, consistent with smaller variability found there (cf. Fig. 9). EOF2 explains 6%–8% of the variance and has a very consistent dipole pattern with positive loadings in Northern California and negative loadings in Southern California. Agreement among PC2 time series is also high, although not as high as for PC1. EOF3 is another dipole mode, this time representing variability that is oppositely phased between coastal and inland locations (2%–4% of the variance). The corresponding PC3s agree less than PC1s or PC2s. Daymet's EOF3 stands out for its irregular loading pattern, which again is likely related to the inhomogeneities discussed above.

For Tmin, EOFs and PCs differ somewhat more than Tmax (Fig. 11). For example, EOF1, characterized by all positive loadings, explains 63%–81% of the variance, a wider range than for Tmax (77%–86%). Daymet's EOF

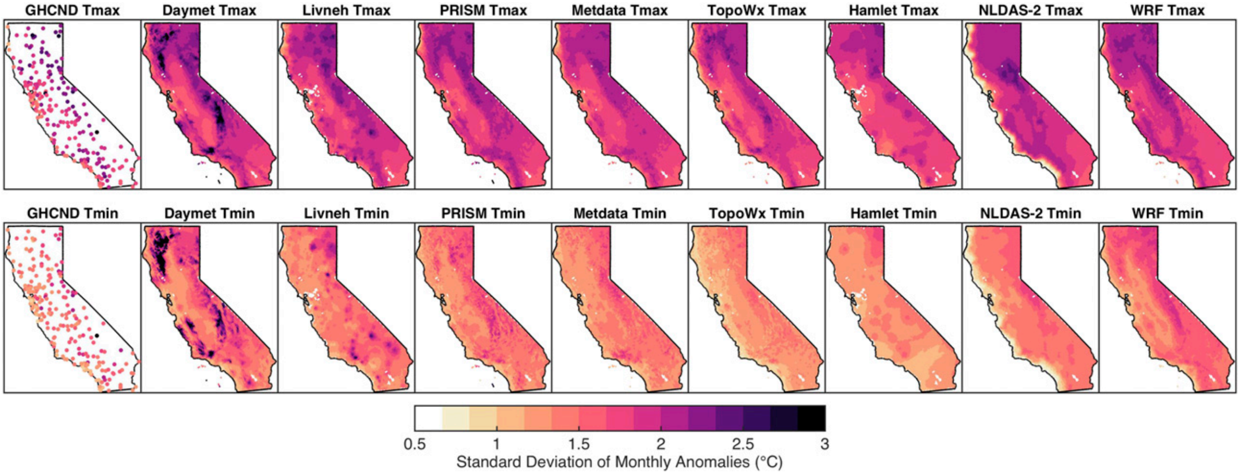

FIG. 9. Standard deviation (°C) of monthly Tmax and Tmin anomalies for the period 1981–2010. For GHCND, only anomalies from nonmissing months are used.

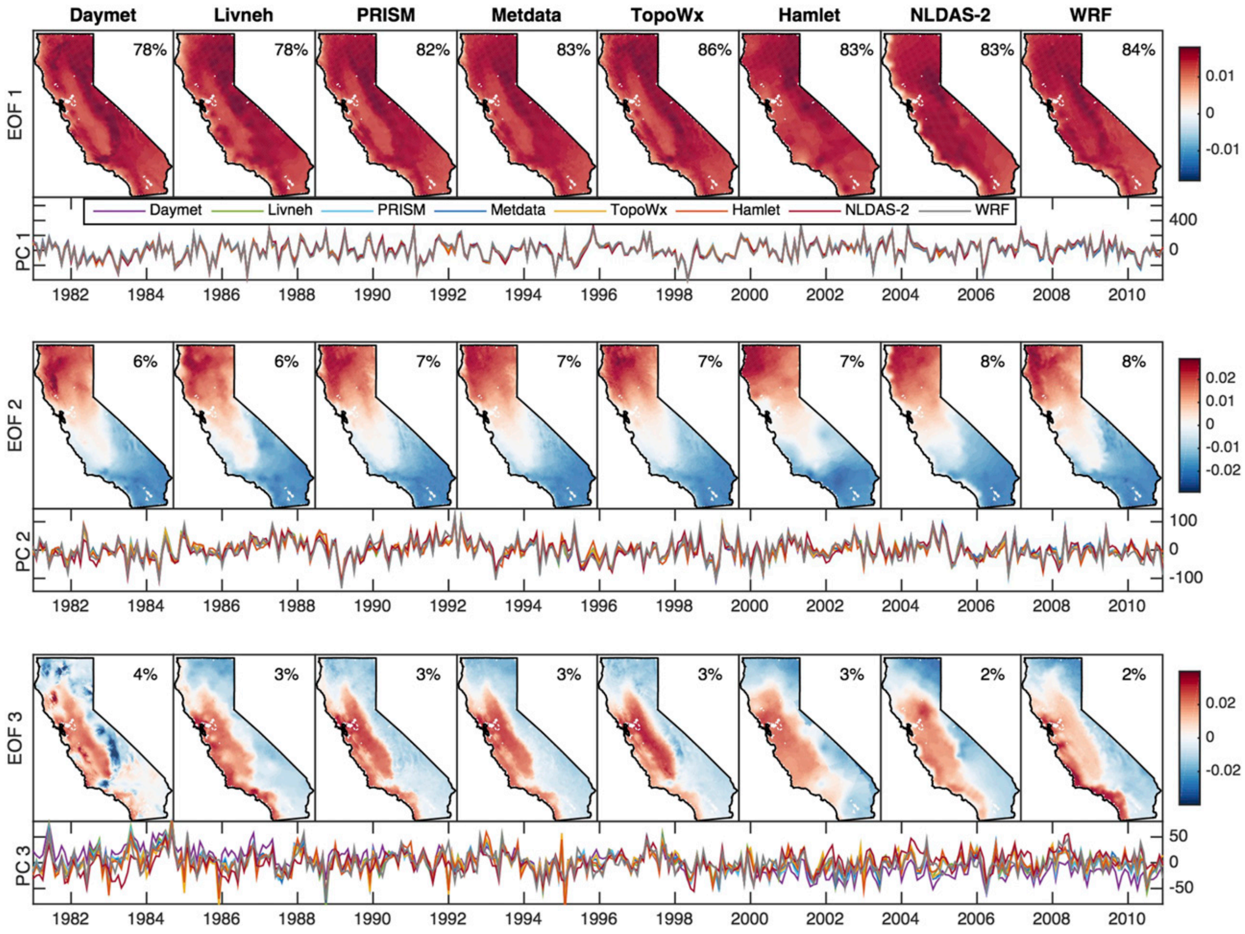

FIG. 10. Three largest Tmax EOFs and their associated PCs for each dataset for the period 1981–2010. Percentages of explained variance are included in the top-right corner of each panel.

spatial patterns differ considerably from the others. They have much higher loadings in the same regions that have large, unphysical trends. Inhomogeneities are also likely responsible for the unusual spatial pattern of Livneh's EOF3. These results suggest that nonclimatic variations can make significant contributions a station-based dataset's variability, not just its long-term trends.

PRISM, Metdata, and TopoWx appear to have the most plausible variability. Their main EOFs are free from artifacts and their PC time series do not have noticeable jumps or trends. Hamlet also has these qualities, but its EOFs are much smoother in space and appear to miss topographic effects. Hamlet's overly smooth EOFs are a side effect of the way it avoids inhomogeneities. Low-frequency variability is adjusted to match interpolated values from stations in the U.S. Historical Climatology Network, a small network of long-running stations with continuous temperature records (Menne et al. 2009). While excluding short-term stations may help produce more realistic long-term trends, it has the side effect of lowering the effective resolution for low-frequency variability, resulting in overly smooth EOFs. Meanwhile, WRF does not rely directly on station data and appears free of inhomogeneity-related artifacts. Overall, WRF EOF spatial patterns are broadly similar to PRISM, TopoWx, and Metdata, but the smaller-scale details are different. WRF also has somewhat smoother Tmin spatial patterns, and does not have finescale variations (<10 km) in complex terrain that the others do, likely because of its lower resolution.

d. *Effect of snow cover*

WRF disagrees considerably with the other datasets over the influence of SAF on temperature anomalies (Fig. 12). WRF simulates large differences in snow cover between April 2007 and April 2010, which are corroborated by MODIS/*Terra* satellite data. WRF temperature differences between these years can reach 7°C at grid cells where snow cover is lost, versus 1°–4°C in the rest of the domain. Meanwhile, the other datasets do not

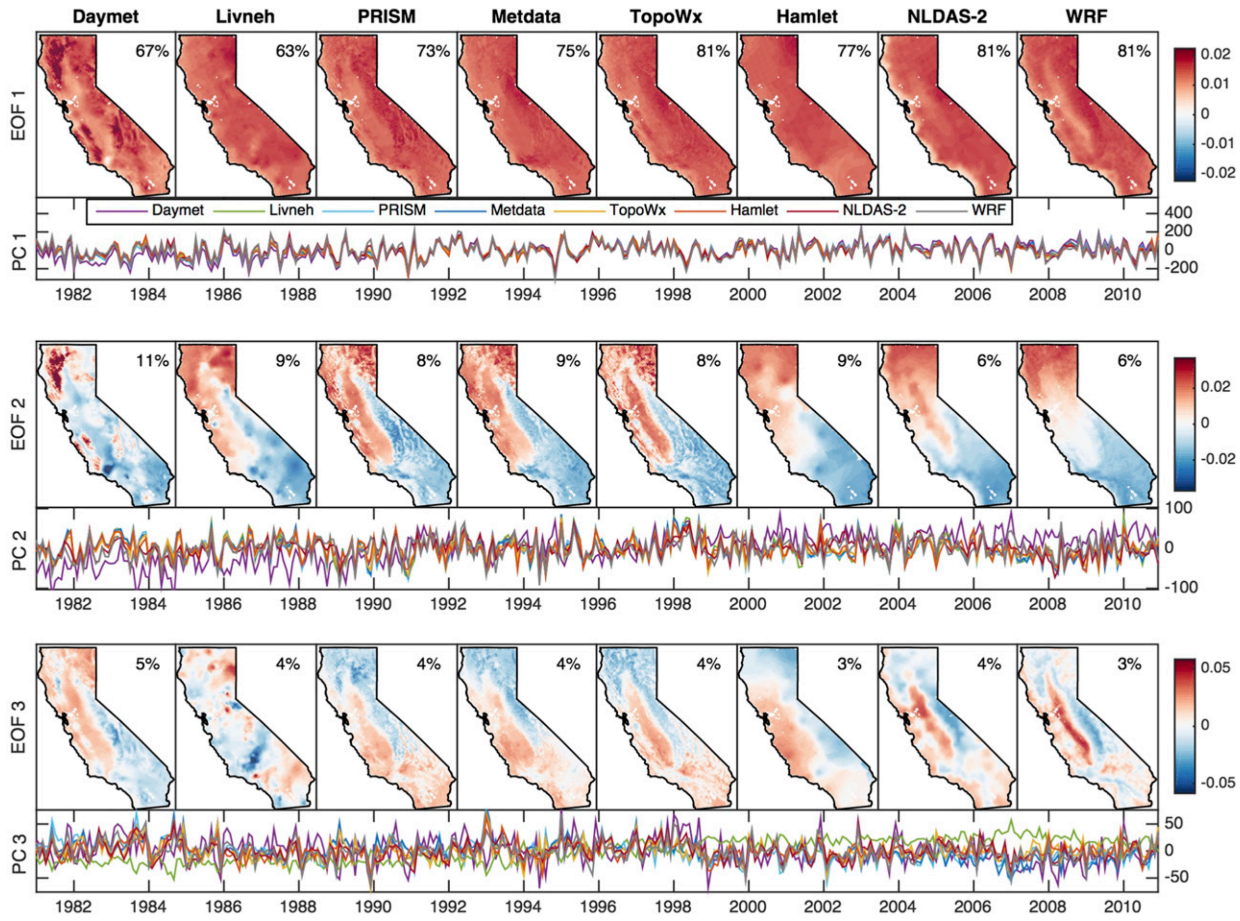

FIG. 11. As in Fig. 10, but for Tmin.

show substantially enhanced temperature differences at grid cells where snow cover is lost. It seems likely that low station density at high elevations would limit station-based datasets from capturing this effect. Overly simplistic relationships between temperature and elevation could also be problematic as they would not be able to capture enhanced warming within a narrow elevation band. It may also be the case that WRF's SAF strength is unrealistically high and actual temperature differences are not amplified as much as WRF suggests. In a study of the Alps, Winter et al. (2017) found that the ETH Zurich COSMO regional climate model (ETHZ-CLM) produced springtime SAF strength values in the 0°–5°C range with a mode of 2.5°C, while observational estimates using station observations suggest that the SAF strength is only 0.4°C. Estimates of WRF's springtime SAF strength for the Sierra Nevada are in the 1°–4°C range (Walton et al. 2017), which is similar to ETHZ-CLM and greater than the 0.4°C observational estimate. Thus, it appears that station-based datasets are missing a real effect—the enhanced warming from snow–albedo feedback—but the effect may be weaker than WRF suggests.

e. Surface lapse rates

TopoWx and PRISM agree that Tmax inland lapse rates are 4°–8°C km^{-1} (Figs. 13a,b). Thus, Livneh's fixed lapse rate of 6.5°C km^{-1} is generally appropriate for Tmax for most inland areas (Fig. 13c). Immediately adjacent to the coast, PRISM differs considerably from the others, showing strongly inverted conditions, with lapse rates reaching −10°C km^{-1}. PRISM explicitly accounts for the suppression of Tmax in low-lying coastal areas due to the penetration of marine air by using a coastal proximity factor and an inversion layer factor, which could explain why it captures this well-known effect (Daly et al. 2008). As for Tmin, both TopoWx and PRISM have lapse rates near zero or even negative for most of California (Figs. 13d,e), likely reflecting nighttime radiation inversions and cold-air pooling. Thus, using a 6.5°C km^{-1} lapse rate for Tmin is unsuitable for large swaths of California (Fig. 13f).

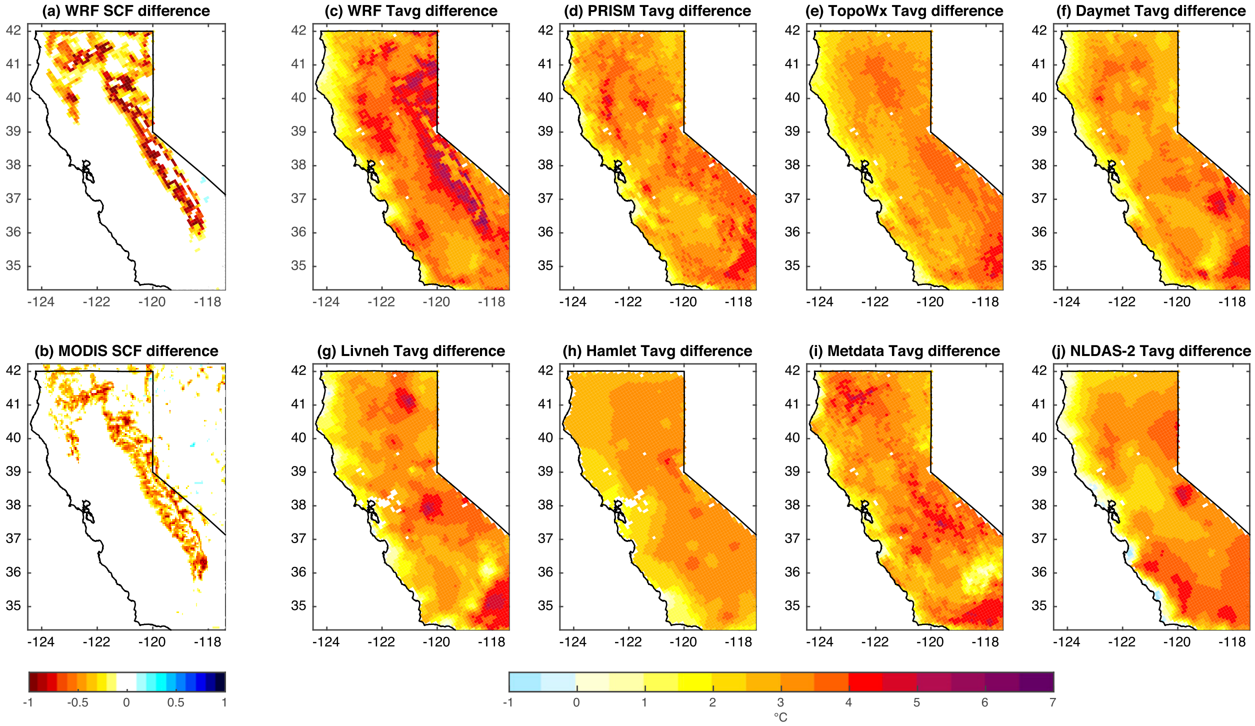
(a) WRF SCF difference
(c) WRF Tavg difference
(d) PRISM Tavg difference
(e) TopoWx Tavg difference
(f) Daymet Tavg difference
(b) MODIS SCF difference
(g) Livneh Tavg difference
(h) Hamlet Tavg difference
(i) Metdata Tavg difference
(j) NLDAS-2 Tavg difference
42
41
40
39
38
37
36
35
-124
-122
-120
-118
-1
-0.5
0
0.5
1
-1
0
1
2
3
4
5
6
7
°C

FIG. 12. Differences in (a) MODIS and (b) WRF snow cover fraction (SCF), and (c)–(j) daily average temperatures (°C) for each dataset, computed as April 2007 minus April 2010.

In general, the lapse rate used by a gridded dataset would have little impact if station density were high everywhere and all elevations were adequately sampled. However, that is not the case in many areas. For example, in the coastal mountains of Northern California, station density is low and almost all stations are located near topographic minima (Fig. 14a). The most credible data suggest that Tmin lapse rates are <2°C km^{-1} for much of this region (Figs. 13d,e), which is substantially different from the fixed lapse rate of 6.5°C km^{-1} used in Livneh. This explains why Livneh is relatively cold compared to the station-based gridded datasets average here (Fig. 14b), with differences becoming increasingly negative with height by 2.9°C km^{-1} (Fig. 14c).

5. Summary and discussion

This study assesses temperature climatologies, trends, and variability in eight high-resolution gridded datasets over California. Each dataset gives a different spatially complete picture of historical temperatures. Five are station-based datasets created by interpolating station data to a regular grid (PRISM, TopoWx, Daymet, Livneh, and Hamlet). Two are created by downscaling reanalysis data (NLDAS-2 and WRF). Finally, one dataset, Metdata, combines monthly means from station-based PRISM with daily variability from reanalysis-based NLDAS-2. This study seeks to identify differences in these datasets, trace these differences back to the datasets' methodologies, and determine which are the most realistic by comparing with station data. In our analysis, particular attention is paid to how the WRF simulation compares with the others, as dynamically downscaled reanalysis data have not been included in previous assessments of gridded datasets.

As expected, when evaluated at station locations, station-based datasets have similar climatologies that closely match GHCND station data. However, matching GHCND station data is an imperfect measure of accuracy. This metric favors datasets with interpolation algorithms that constrain the interpolated data to match exactly with the original data, and does not reflect how well these datasets do away from stations. Furthermore, gridded datasets that correct for inhomogeneities or other station data issues could be penalized, as they would no longer match the original data as closely. Thus, it is crucial to examine how these datasets behave away from stations. Differences are more pronounced (up to 6°C) away from stations, in complex terrain, and near the coast. The existence of large differences away from the stations despite close agreement at stations underscores the need to compare station-based datasets

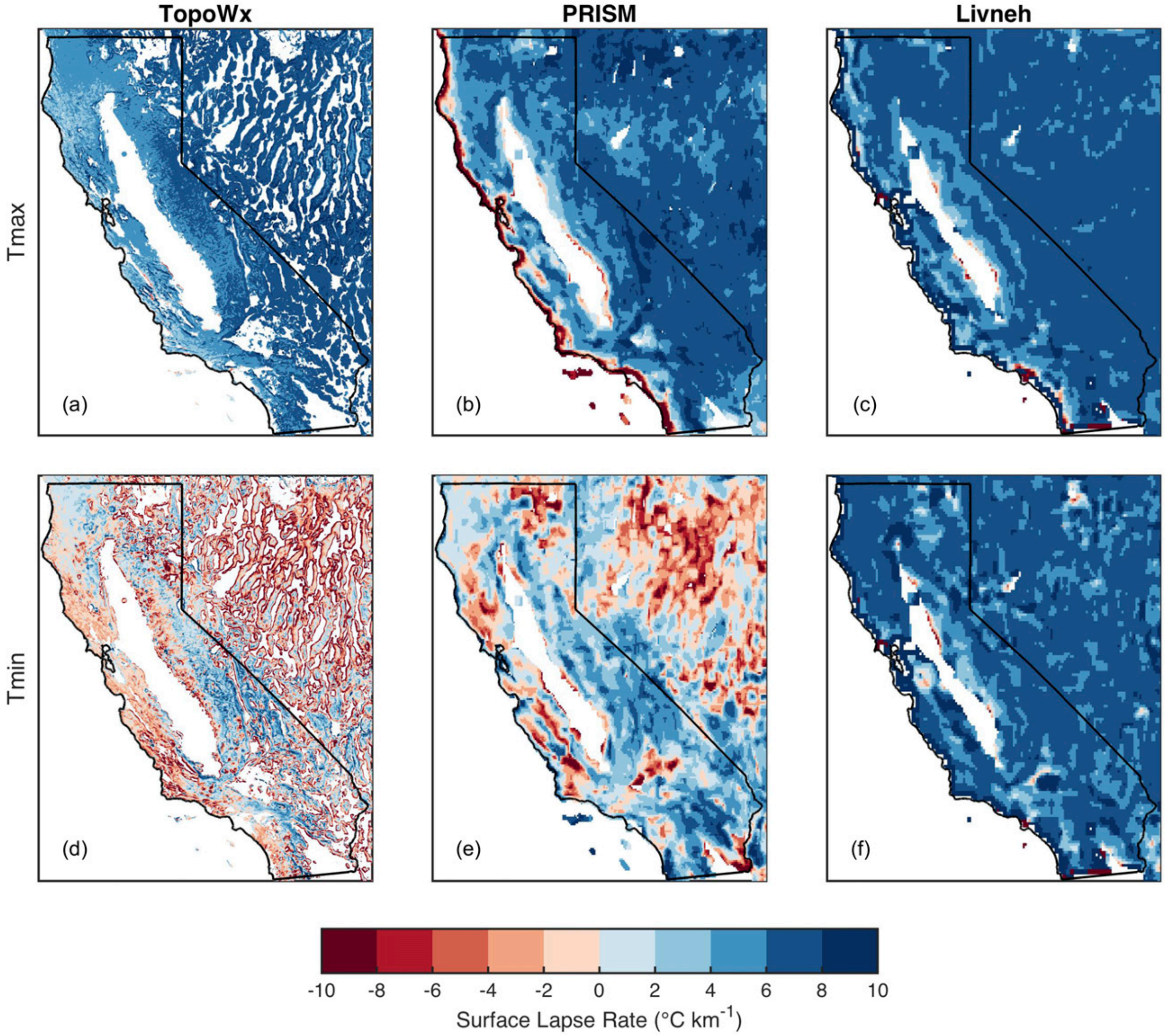

FIG. 13. Surface lapse rate (°C km^{-1}) calculated as the negative slope determined by linearly regressing climatological (a)–(c) Tmax and (d)–(f) Tmin onto elevation for all nearby grid cells within two grid lengths. Cool colors indicate decreasing temperature with height. Warm colors indicate increasing temperature with height (i.e., inverted conditions). Grid cells whose neighbors range in elevation by less than 100 m are excluded from the calculation.

everywhere, not just at station locations. Meanwhile, the reanalysis-based datasets, WRF and NLDAS-2, are not directly constrained to match station observations and have systematic biases relative to GHCND station data, making them less suitable for assessing absolute errors in temperature.

There are clearly large differences in climatology between these datasets, but away from the station locations it is difficult to know definitively which dataset is most realistic. It is possible, in some cases, to demonstrate that a dataset relies on a problematic assumption. For example, Livneh uses a fixed lapse rate of 6.5°C km^{-1} to adjust for elevation. Tmin lapse rates were found to be negative or near zero for much of the domain (inverted or neutral conditions). Thus, a fixed positive lapse rate of 6.5°C km^{-1} is not suitable for Tmin and explains why Livneh Tmin so cold at high elevations. This finding is consistent with Mizukami et al. (2014) and Newman et al. (2015), who found that datasets with fixed positive lapse rates have cold biases at high elevations. Based on these results, using a gridded dataset that accurately captures variable lapse rates is especially important when studying daily minimum temperatures in complex terrain.

Differences in trends are the result of choices made about station homogenization and how to handle missing

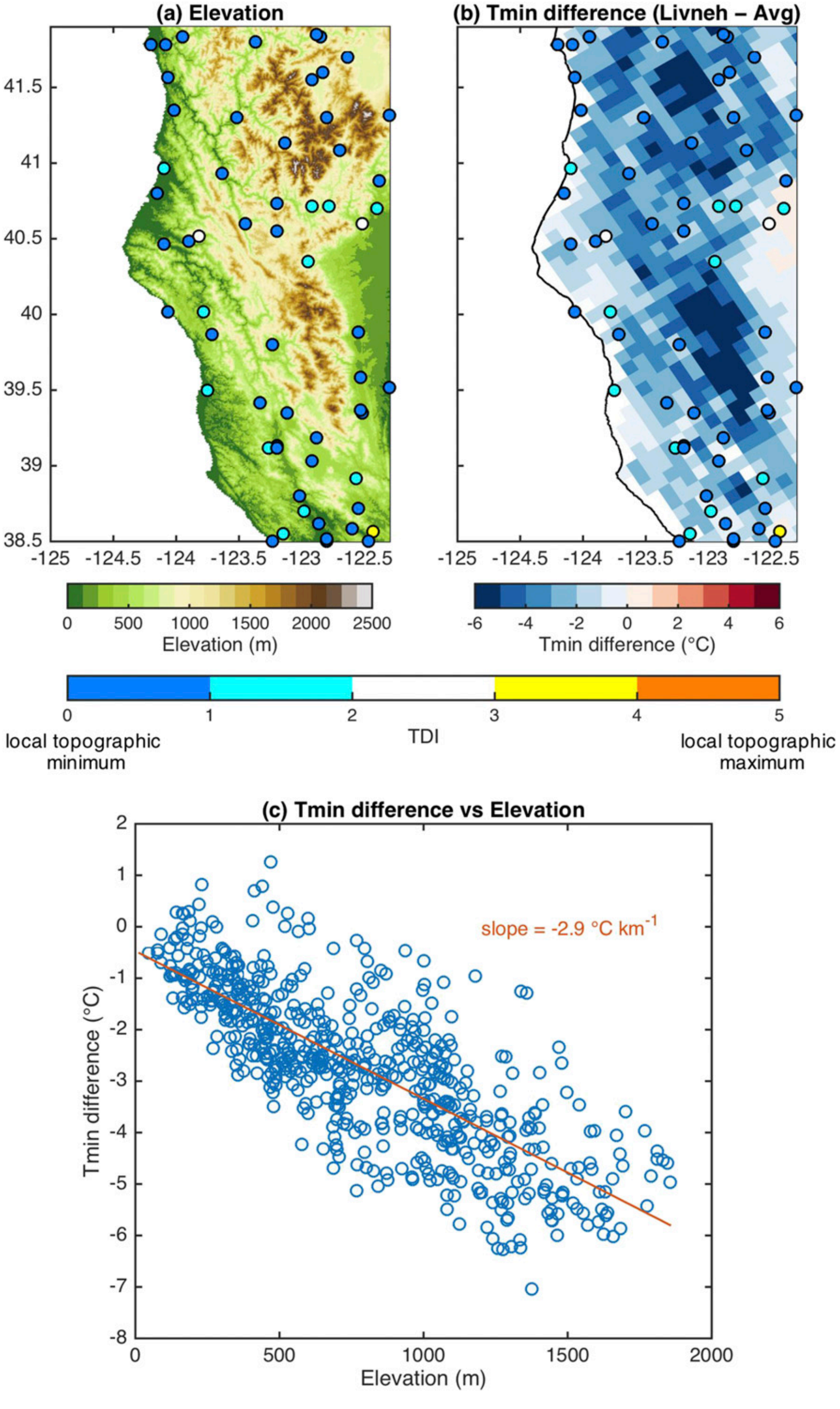

FIG. 14. (a) Elevation (m) in the coastal mountains of Northern California. (b) Tmin climatology difference between Livneh and station-based gridded dataset average. The topographic dissection index (TDI) is plotted at each COOP station (colored circles). Warm colors indicate station is near topographic maxima. Cold colors indicate station is near topographic minima. (c) Tmin difference (Livneh minus station-based gridded dataset average) vs elevation at all grid cells within the coastal region shown in (a) and (b). Slope computed using least squares linear regression.

data. Daymet, Livneh, PRISM, and Metdata do not homogenize station data and subsequently have large nonclimatic trends. Daymet appears especially sensitive to changes in station data availability, and the introduction of the RAWS stations could explain its large systematic trends. In contrast, TopoWx and Hamlet use homogenization procedures and have smooth trend fields. There are legitimate concerns that homogenization could smooth out true local trends, forcing the regional trend on each grid cell (Pielke et al. 2007). Based on our results, this seems to be a problem with Hamlet, but not with TopoWx. Meanwhile, WRF and NLDAS-2 have smooth trend fields like TopoWx and Hamlet, but the details are different enough to cast doubt on their accuracy. If accurately capturing trends is important, then using a homogenized dataset is necessary.

Most datasets have broad agreement in the spatial patterns and the timing of the leading modes. Daymet and Livneh are the main exceptions, with prominent nonclimatic artifacts in the spatial patterns and jumps in the associated time series. NLDAS-2 is also noteworthy, but for unrealistically low variability very near the coast, which makes it not recommended for coastal applications.

While the WRF simulation has important disagreements with station-based datasets, it still broadly similar in most aspects considered here. WRF's most glaring issue is a cold bias in Tmax at high elevations. But, for Tmin, it is within the range of station-based datasets. WRF's temporal variability is highly correlated with the most plausible station-based datasets, and its spatial patterns of the leading modes are qualitatively similar to the most plausible station-based datasets. In fact, WRF's variability is more realistic than some unhomogenized station-based datasets, such as Daymet, which has large jumps due to missing data. These results suggest that dynamically downscaled reanalysis can produce a spatially complete picture of the historical temperatures on par with station-based datasets in many aspects. In fact, it could potentially be a valuable, complementary perspective to station-based dataset in snow-covered areas, as it explicitly simulates the snow cover anomalies on temperature. However, further research is needed to determine if WRF's snow–albedo feedback strength is realistic.

Although WRF and NLDAS-2 are both downscalings of NARR, NLDAS-2 is less realistic in most aspects considered here. NLDAS-2 has large biases in both Tmax and Tmin. NLDAS-2 has less realistic variability especially very near the coast, which could be due to interpolation between grid cells across the land–sea interface. Thus, at least in this case, dynamical downscaling is found to add value over linear interpolation in downscaling historical reanalysis.

Often station-based gridded datasets are treated as ground truth, without acknowledging problems with station data or assumptions needed to generate a spatial complete temperature field from point measurements. Indeed, the large differences between gridded datasets seen here indicate that gridded dataset choice is a considerable source of uncertainty. It is important that users of gridded datasets are aware of their limitations and select datasets appropriate for the task at hand. If capturing trends is important, then homogenization is necessary. For capturing climatologies, reanalysis-based datasets may not be suitable because of their systematic biases; station-based datasets that capture variable lapse rates along the coast and in complex terrain are a better choice. Many station-based datasets could be improved with straightforward fixes, like making variable lapse rates and station homogenization standard practice. Reanalysis-based gridded datasets are likely to improve from ongoing progress in regional and global climate modeling and data assimilation.

Acknowledgments. The authors thank three anonymous reviewers for their constructive feedback. The authors thank David Pierce for generating WRF Tmax and Tmin from 3-hourly output. The authors thank Mu Xiao for providing the extension of the Hamlet dataset. For information about obtaining this extension, direct inquiries to muxiao@ucla.edu. Funding for this work was provided by the U.S. Department of Energy (Grant DE-SC0014061; ''Developing Metrics to Evaluate the Skill and Credibility of Downscaling''). The authors are not aware of any conflicts of interest.

REFERENCES

Abatzoglou, J. T., 2013: Development of gridded surface meteorological data for ecological applications and modelling. *Int. J. Climatol.*, **33**, 121–131, https://doi.org/10.1002/joc.3413.

Behnke, R., S. Vavrus, A. Allstadt, T. Albright, W. Thogmartin, and V. Radeloff, 2016a: Evaluation of downscaled, gridded climate data for the conterminous United States. *Ecol. Appl.*, **26**, 1338–1351, https://doi.org/10.1002/15-1061.

——, ——, ——, ——, ——, and ——, 2016b: Data from: Evaluation of downscaled, gridded climate data for the conterminous United States. Dryad Digital Repository, https://doi.org/10.5061/dryad.7tv80.

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.

Caldwell, P., H.-N. 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.

Cosgrove, B. A., and Coauthors, 2003: Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *J. Geophys. Res.*, **108**, 8842, https://doi.org/10.1029/2002JD003118.

Cubasch, U., and Coauthors, 2001: Projections of future climate change. *Climate Change 2001: The Scientific Basis*, J. T. Houghton et al., Eds., Cambridge University Press, 525–582.

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.

——, R. P. Neilson, and D. L. Phillips, 1994: A statistical–topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.*, **33**, 140–158, https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

——, 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. *Int. J. Climatol.*, **28**, 2031–2064, https://doi.org/10.1002/joc.1688.

——, D. R. Conklin, and M. H. Unsworth, 2010: Local atmospheric decoupling in complex terrain alters climate change impacts. *Int. J. Climatol.*, **30**, 1857–1864, doi:10.1002/joc.2007.

Dee, D., and Coauthors, Eds., 2016: The Climate Data Guide: Atmospheric reanalysis: Overview & comparison tables. UCAR/NCAR, https://climatedataguide.ucar.edu/climate-data/atmospheric-reanalysis-overview-comparison-tables.

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.

Hall, D. K., V. V. Salomonson, and G. A. Riggs, 2006: MODIS/*Terra* snow cover monthly L3 global 0.05 deg CMG, version 5 (April 2000–December 2006 subset). National Snow and Ice Data Center, accessed 31 July 2015, https://doi.org/10.5067/IPPLURB6RPCN.

Hamlet, A. F., and D. P. Lettenmaier, 2005: Production of temporally consistent gridded precipitation and temperature fields for the continental United States. *J. Hydrometeor.*, **6**, 330–336, https://doi.org/10.1175/JHM420.1.

——, and Coauthors, 2010: Chapter 3: Historical meteorological driving data. Final project report for the Columbia basin climate change scenarios project. University of Washington, accessed 18 January 2017, http://www.hydro.washington.edu/2860/report/.

Hidalgo, H. G., and Coauthors, 2009: Detection and attribution of streamflow timing changes to climate change in the western United States. *J. Climate*, **22**, 3838–3855, https://doi.org/10.1175/2009JCLI2470.1.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis, 2005: Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.*, **25**, 1965–1978, https://doi.org/10.1002/joc.1276.

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. Meteor.*, **151**, 1066–1073, https://doi.org/10.1016/j.agrformet.2011.03.011.

Holland, M. M., and C. M. Bitz, 2003: Polar amplification of climate change in coupled models. *Climate Dyn.*, **21**, 221–232, https://doi.org/10.1007/s00382-003-0332-6.

Iacobellis, S. F., and D. R. Cayan, 2013: The variability of California summertime marine stratus: Impacts on surface air temperatures. *J. Geophys. Res. Atmos.*, **118**, 9105–9122, https://doi.org/10.1002/jgrd.50652.

Johnstone, J. A., and T. E. Dawson, 2010: Climatic context and ecological implications of summer fog decline in the coast redwood region. *Proc. Natl. Acad. Sci. USA*, **107**, 4533–4538, https://doi.org/10.1073/pnas.0915062107.

Juang, H.-M. H., and M. Kanamitsu, 1994: The NMC Nested Regional Spectral Model. *Mon. Wea. Rev.*, **122**, 3–26, https://doi.org/10.1175/1520-0493(1994)122<0003:TNNRSM>2.0.CO;2.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kanamitsu, M., and H. Kanamaru, 2007: Fifty-seven-year California Reanalysis Downscaling at 10 km (CaRD10). Part I: System detail and validation with observations. *J. Climate*, **20**, 5553–5571, https://doi.org/10.1175/2007JCLI1482.1.

Letcher, T. W., and J. R. Minder, 2015: Characterization of the simulated regional snow albedo feedback using a regional climate model over complex terrain. *J. Climate*, **28**, 7576–7595, https://doi.org/10.1175/JCLI-D-15-0166.1.

Livneh, B., E. A. Rosenberg, C. Lin, B. Nijssen, V. Mishra, K. M. Andreadis, E. P. Maurer, and D. P. Lettenmaier, 2013: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions. *J. Climate*, **26**, 9384–9392, https://doi.org/10.1175/JCLI-D-12-00508.1.

Lundquist, J. D., N. Pepin, and C. Rochford, 2008: Automated algorithm for mapping regions of cold-air pooling in complex terrain. *J. Geophys. Res.*, **113**, D22107, https://doi.org/10.1029/2008JD009879.

Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen, 2002: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *J. Climate*, **15**, 3237–3251, https://doi.org/10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2.

Menne, M. J., and C. N. Williams, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate*, **22**, 1700–1717, https://doi.org/10.1175/2008JCLI2263.1.

——, ——, and R. S. Vose, 2009: The United States Historical Climatology Network monthly temperature data, version 2. *Bull. Amer. Meteor. Soc.*, **90**, 993–1007, https://doi.org/10.1175/2008BAMS2613.1.

——, I. Durre, R. S. Vose, B. E. Gleason, and T. G. Houston, 2012a: An overview of the Global Historical Climatology Network–Daily database. *J. Atmos. Oceanic Technol.*, **29**, 897–910, https://doi.org/10.1175/JTECH-D-11-00103.1.

——, and Coauthors, 2012b: Global Historical Climatology Network–Daily (GHCN-Daily), version 3. NOAA National Climatic Data Center. https://doi.org/10.7289/V5D21VHZ.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, https://doi.org/10.1175/BAMS-87-3-343.

Mitchell, K. E., and Coauthors, 2004: The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J. Geophys. Res.*, **109**, D07S90, https://doi.org/10.1029/2003JD003823.

Mizukami, N., M. P. Clark, A. G. Slater, L. D. Brekke, M. M. Elsner, J. R. Arnold, and S. Gangopadhyay, 2014: Hydrologic implications of different large-scale meteorological model forcing datasets in mountainous regions. *J. Hydrometeor.*, **15**, 474–488, https://doi.org/10.1175/JHM-D-13-036.1.

Mote, P. W., A. F. Hamlet, M. P. Clark, and D. P. Lettenmaier, 2005: Declining mountain snowpack in western North America. *Bull. Amer. Meteor. Soc.*, **86**, 39–49, https://doi.org/10.1175/BAMS-86-1-39.

Newman, A. J., and Coauthors, 2015: Gridded ensemble precipitation and temperature estimates for the contiguous

United States. *J. Hydrometeor.*, **16**, 2481–2500, https://doi.org/10.1175/JHM-D-15-0026.1.

Niu, G.-Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, https://doi.org/10.1029/2010JD015139.

Oyler, J. W., A. Ballantyne, K. Jencso, M. Sweet, and S. W. Running, 2015a: Creating a topoclimatic daily air temperature dataset for the conterminous United States using homogenized station data and remotely sensed land skin temperature. *Int. J. Climatol.*, **35**, 2258–2279, https://doi.org/10.1002/joc.4127.

——, S. Z. Dobrowski, A. P. Ballantyne, A. E. Klene, and S. W. Running, 2015b: Artificial amplification of warming trends across the mountains of the western United States. *Geophys. Res. Lett.*, **42**, 153–161, https://doi.org/10.1002/2014GL062803.

Pielke, R., and Coauthors, 2007: Unresolved issues with the assessment of multidecadal global land surface temperature trends. *J. Geophys. Res.*, **112**, D24S08, https://doi.org/10.1029/2006JD008229.

Pierce, D. W., D. R. Cayan, and B. L. Thrasher, 2014: Statistical downscaling using localized constructed analogs (LOCA). *J. Hydrometeor.*, **15**, 2558–2585, https://doi.org/10.1175/JHM-D-14-0082.1.

Rasmussen, R., and Coauthors, 2011: High-resolution coupled climate runoff simulations of seasonal snowfall over Colorado: A process study of current and warmer climate. *J. Climate*, **24**, 3015–3048, https://doi.org/10.1175/2010JCLI3985.1.

Salathé, E. P., R. Steed, C. F. Mass, and P. H. Zahn, 2008: A high-resolution climate model for the U.S. Pacific Northwest: Mesoscale feedbacks and local responses to climate change. *J. Climate*, **21**, 5708–5726, https://doi.org/10.1175/2008JCLI2090.1.

Shepard, D. S., 1984: Computer mapping: The SYMAP interpolation algorithm. *Spatial Statistics and Models*, G. L. Gaile and C. J. Willmott, Eds., D. Reidel, 133–145.

Simpson, J. J., G. L. Hufford, C. Daly, J. S. Berg, and M. D. Fleming, 2005: Comparing maps of mean monthly surface temperature and precipitation for Alaska and adjacent areas of Canada produced by two different methods. *Arctic*, **58**, 137–161, https://doi.org/10.14430/arctic407.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

Stahl, K., R. D. Moore, J. A. Floyer, M. G. Asplin, and I. G. 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. Meteor.*, **139**, 224–236, https://doi.org/10.1016/j.agrformet.2006.07.004.

Stefanova, L., V. Misra, S. Chan, M. Griffin, J. J. O'Brien, and T. J. Smith III, 2012: A proxy for high-resolution regional reanalysis for the Southeast United States: Assessment of precipitation variability in dynamically downscaled reanalyses. *Climate Dyn.*, **38**, 2449–2466, https://doi.org/10.1007/s00382-011-1230-y.

Stoklosa, J., C. Daly, S. Foster, M. Ashcroft, and D. Warton, 2015: A climate of uncertainty: Accounting for error and spatial variability in climate variables for species distribution models. *Methods Ecol. Evol.*, **6**, 412–423, https://doi.org/10.1111/2041-210X.12217.

Thornton, P. E., S. W. Running, and M. A. White, 1997: Generating surfaces of daily meteorological variables over large regions of complex terrain. *J. Hydrol.*, **190**, 214–251, https://doi.org/10.1016/S0022-1694(96)03128-9.

——, M. M. Thornton, B. W. Mayer, Y. Wei, R. Devarakonda, R. S. Vose, and R. B. Cook, 2016: Daymet: Daily surface weather data on a 1-km grid for North America, version 3, 1980–2012. ORNL DAAC, Oak Ridge, accessed 19 November 2016, https://doi.org/10.3334/ORNLDAAC/1328.

Vose, R. S., S. Applequist, M. Squires, I. Durre, M. Menne, C. N. Williams Jr., C. Fenimore, K. Gleason, and D. Arndt, 2014: Improved historical temperature and precipitation time series for U.S. climate divisions. *J. Appl. Meteor. Climatol.*, **53**, 1232–1251, https://doi.org/10.1175/JAMC-D-13-0248.1.

Walton, D. B., F. Sun, A. Hall, and S. Capps, 2015: A hybrid dynamical–statistical downscaling technique. Part I: Development and validation of the technique. *J. Climate*, **28**, 4597–4617, https://doi.org/10.1175/JCLI-D-14-00196.1.

——, A. Hall, N. Berg, M. Schwartz, and F. Sun, 2017: Incorporating snow albedo feedback into downscaled temperature and snow cover projections for California's Sierra Nevada. *J. Climate*, **30**, 1417–1438, https://doi.org/10.1175/JCLI-D-16-0168.1.

Winter, K. J.-P. M., S. Kotlarski, S. C. Scherrer, and C. Schär, 2017: The Alpine snow-albedo feedback in regional climate models. *Climate Dyn.*, **48**, 1109–1124, https://doi.org/10.1007/s00382-016-3130-7.

Xia, Y., and Coauthors, 2012: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *J. Geophys. Res.*, **117**, D03109, https://doi.org/10.1029/2011JD016048.