An assessment of high-resolution gridded temperature datasets

Daniel Walton*

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

Alex Hall

Department of Atmospheric and Ocean Studies, University of California, Los Angeles, Los Angeles, CA

*Corresponding author email address: waltond@ucla.edu

Abstract

High-resolution gridded datasets are in high demand because they are spatially complete and include important fine-scale details. Here, eight high-resolution gridded temperature datasets are assessed by comparing with Global Historical Climatology Network – Daily (GHCND) station data. Previous assessments have focused on station-based datasets, which are generated by interpolating station data to a regular grid. Another way to generate spatially complete historical data is to downscaling reanalysis to higher resolution. 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. Not surprisingly, at stations, station-based datasets are found to agree closely with each other and with GHCND station data. However, away from the stations, spread among station-based datasets can exceed 6 °C. Some of these datasets are very likely biased away from stations, due to invalid assumptions about how temperatures vary with elevation. Meanwhile, reanalysis-based datasets have more freedom to differ from observations, and have systematic biases relative to station data. Dynamically downscaled reanalysis is less biased than interpolated reanalysis, and has more realistic variability and trends. Many station-based datasets have large unphysical trends and non-climatic variations because they do not correct inhomogeneities in station data. Station-based datasets could be improved through better quality control of station data and more realistic assumptions of how temperatures vary away from the stations. Reanalysis-based datasets are likely to improve from ongoing progress in global and regional climate modeling.

1. Introduction

High-resolution gridded temperature datasets are widely used because they are spatially complete and include fine-scale variations due to topography and other features. Such detail is important for many modeling applications in fields like 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 (e.g. Hidalgo et al. 2009, Pierce et a. 2014).

There are a variety of approaches for generating high-resolution gridded temperature data. One approach is to interpolate data from irregularly spaced stations to a regular grid. Datasets generated in this manner are termed *station-based* datasets. Some station-based datasets 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. 2015). A challenge with station data is that changes in station siting, instrumentation, and time of observation add non-climatic 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. 2015), which makes them better suited for long-term trend analysis. Some datasets include uncertainty or facilitate calculations of uncertainty. For instance, Newman et al. (2015) have generated an ensemble of possible historical sequences, which can be used to determine uncertainty by calculating the ensemble variance.

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, Thornton et al. 2012) and WorldClim (Hijmans et al. 2005) over the continental United States. PRISM determines temperatures on 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 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. (2015) 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 the inversions in Tmin. TopoWx had the strongest inversions, PRISM had 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 does not include any auxiliary predictors or use advanced station weights — is relatively smooth.

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 data to Maurer et al. (2002; henceforth "Maurer"), and noted that Maurer is consistently colder at high elevations. Newman et al. attribute this to the use of a fixed 6.5 K km^{-1} lapse rate in Maurer. Mizukami et al. (2014), also found Maurer to be relatively cold at high elevations.

A second approach to creating a gridded temperature dataset 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, ERA-20C; for details, see The Climate Data Guide: Atmospheric Reanalysis: Overview & Comparison Tables, available from https://climatedataguide.ucar.edu/climate-data/atmospheric-reanalysis-overview-comparison-tables). However, the resolutions of these datasets — ranging from 0.3 degrees to 5 degrees — 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, Walton et al. 2017). One straightforward way to downscale reanalysis is with interpolation. For example, the temperature forcings in the NLDAS-2 dataset (Xia et al. 2012) are derived by interpolating North American Regional Reanalysis (NARR; Mesinger et al. 2006) to 1/8 degree 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 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. 2015). 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. 2015). Behnke et al. (2016) 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, Maurer) or attempt to model their effects through auxiliary predictors or weights (e.g. TopoWx, PRISM). Interpolated reanalysis may also struggle in these areas since the native resolution of the original reanalysis is too low, and linear interpolation cannot recover these effects.

One effect that hasn't 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 it is 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 high-resolution spatially complete 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 station observations. Station data come from the Global Historical Climatology Network – Daily stations (GHCND, Menne et al. 2012a,b) as made available by Behnke et al. (2016b) via the Dryad data package (http://dx.doi.org/10.5061/dryad.7tv80). 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:

- PRISM (Daly et al. 2008)
- TopoWx (Oyler et al. 2015)
- 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)
- 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. Gridded datasets

a. WRF historical simulation

The first dataset is a dynamical downscaling of 32 km resolution NCEP North American Regional Reanalysis (NARR; Mesinger et al. 2006) for the 1981–2015 period using the Weather Research and Forecasting model v3 (WRF; 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 U.S. and northeastern Pacific Ocean, a 9 km domain covering California, and 3 km domain covering the Sierra Nevada. This study focuses on the 9 km domain covering California (Fig. 1a). A cubic spline fit to WRF 3-hourly output is used to calculate daily Tmax and Tmin.

b. PRISM

The Parameter-elevation Relationships on Independent Slopes Model (PRISM; Daly et al., 1994, Daly et al., 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. Multiple PRISM datasets are available, differing in temporal frequency (monthly or daily), resolution (2.5 min or 30 sec), and other factors. 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 2013-6-9 and 2014-6-9). This dataset uses stations from multiple networks to give the best estimate at any given time. Although station data are subjected to quality control procedures, no adjustments are made to ensure temporal homogeneity. PRISM incorporates data from ~10,000 stations spanning multiple networks, including COOP, RAWS, CDEC, Agrimet, NCRS, CIMIS and more. For complete details about the station networks included in these datasets, see the PRISM Climate Group webpage (http://prism.oregonstate.edu).

c. TopoWx

TopoWx or "Topography Weather" is gridded dataset of daily Tmin and Tmax based on station data and remotely-sensed land skin temperature (Oyler et al. 2015; data was downloaded from http://www.scrimhub.org/resources/topowx/). TopoWx covers the conterminous United States at 30 arcsec (~800 m resolution) for the period 1980–2015. TopoWx uses station data (Fig. 1b) from GCHND stations, National Resource Conservation Service (NRCS) snow telemetry (SNOTEL) and snow course sites, and US Forest Service and Bureau of Land Management (BLM) Remote Automatic Weather Stations (RAWS). The homogenization algorithm of Menne and Williams (2009) is used to correct for inhomogeneities caused by changes in observation practices, siting, and instrumentation. Missing values are filled by comparing with non-missing neighboring observations and applying spatial regression (Durre et al. 2010). Daily gridded Tmax and Tmin are created in a two-step process. The first step is to create gridded Tmax and Tmin climate normals for the 1981–2010 period using a regression-kriging framework. Predictors for the climate normals include latitude, longitude, and elevation, as well as 8-day average remotely-sensed land skin temperature (LST) from the Moderate Resolution Imaging

Spectroradiometer (MODIS) aboard the Aqua satellite (product MYD11A2; Dozier 1996, Wan 2008). The second step is to interpolate daily temperature anomalies. This step is performed with a combination of moving-window geographically weighted regression and inverse distance weighting. The interpolated daily anomalies for the 1948–2015 period are then added to the 1980–2010 normals to produce daily values.

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 THREDDS server (http://thredds.daac.ornl.gov/thredds/catalogs/ornldaac/Regional_and_Global_Data/DAYMET_COLLECTIONS/DAYMET_COLLECTIONS.html) on January 9, 2017. Daymet interpolates 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 (henceforth "Livneh") contains station-based meteorological variables and modelled hydrologic variables that covers the conterminous United States at 1/16° (~6 km) resolution for the period 1915–2011. Linveh data was downloaded from ftp.hydro.washington.edu/pub/bLivneh/CONUS/. 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 station data from National Weather Service (NWS) Cooperative Observer Program (COOP) weather stations over the conterminous United States. COOP station locations used by Livneh are shown in Fig. 1b. Only stations with at least 20 years of valid data were used. 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 "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 Hamlet temperatures are adjusted so 1971–2000 monthly normals match PRISM and low-frequency variability matches the quality controlled United States 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. NLDAS-2

The retrospective forcing dataset for the North American Land Data Assimilation System (NLDAS; Cosgrove et al. 2003, Mitchell et al. 2004) includes temperature data with 1-hour temporal resolution and 1/8° spatial resolution. It is intended to be used as forcing to the NLDAS project which runs land surface simulations using multiple land surface models. The most recent version of the project, NLDAS-2 (Xia et al. 2012), constructs its temperature forcing data by linearly interpolating 32 km, 3-hourly NARR 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 (a reanalysis product), but using linear interpolation instead of a regional climate model. Where interpolated NARR elevation differs with elevation of the 1/8° NLDAS-2 grid, temperatures are adjusted using a fixed lapse rate of 6.5 °C km^{-1}. Data was downloaded using the NASA Earthdata Simple Subset Wizard (https://disc.gsfc.nasa.gov/SSW/).

h. Metdata

Metdata (Abatzoglou 2013) is a hybrid dataset of meteorological forcings that combines the high temporal resolution (sub-daily) 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 and daily (and sub-daily) temporal resolution from http://metdata.northwestknowledge.net. To create Metdata, first NLDAS-2 is linearly interpolated to 4 km resolution. Next, interpolated NDLAS-2 sub-daily anomalies are calculated relative to monthly means. The final step is to add the sub-daily anomalies from NLDAS-2 onto PRISM monthly means. Thus, Metdata has the daily (and sub-daily) variability of NLDAS-2, but the climatologies and monthly variability of PRISM. Metdata is technically a hybrid dataset, but because monthly means are used in this assessment, and Metdata monthly climatologies and variability are derived from station-based PRISM, here it is grouped 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 bilinear interpolation. Only land areas are considered as some datasets don't 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. Stations data are not without error, but collectively they represent one of our best sources of temperature observations. Thus, if a gridded dataset has large differences with many GHCND stations, then the gridded dataset is probably biased. Meanwhile, differences with the average of the station-based gridded datasets do not necessarily indicate biases, 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.

To quantify range in climatologies among the datasets, inter-dataset spread is calculated at each grid cell. Spread is calculated for different subgroups of datasets. This allows us to see how including different datasets changes the spread. The first subgroup is the PRISM relatives: PRISM, Hamlet, and Metdata. This group is expected to have a small spread since Hamlet is adjusted to match PRISM's climatology, and Metdata is constructed using PRISM's monthly mean values. The second subgroup is all station-based datasets. The third is all station-based datasets and WRF. The final group is all datasets (station-based, WRF, and NLDAS-2).

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 as some datasets perform adjustments for inhomogeneities in the data, while others do not. Linear trends are also computed for the GHCND station data, using all non-missing 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 GCHND station data, using all non-missing 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 between these years are also computed, but for WRF and remotely sensed data from the Moderate-resolution Imaging Spectroradiometer aboard 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. Local 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 these complicated relationships between temperature and elevation, then they may produce errant temperature patterns. For instance, some datasets use fixed positive lapse rates throughout the domain, which could be problematic in areas experiencing inverted temperatures. To diagnose the local relationship between temperature and elevation throughout the domain, we use TopoWx at its native 30 arcsec (800

m) resolution. TopoWx has the highest native resolution of the datasets considered here and uses satellite LST as an auxiliary predictor for climate normals. So, it is likely to provide the most accurate and detailed information for determining the relationship between temperature and elevation. A local lapse rate is inferred at each grid cell by applying linear regression to temperature and elevation data from surrounding grid cells (defined as grid cells within two grid lengths (~1600 m) in the *x* or *y* direction). Calculation of the local lapse rate should aid us in determining where fixed positive lapse rate assumptions are valid.

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. TDI is calculated as follows:

$$\mathrm{TDI}(x, y) = \sum_{i=1}^{n} \frac{z(x, y) - z_{\min}(i)}{z_{\max}(i) - z_{\min}(i)}$$

where $z(x, y)$ is the elevation at the grid cell or station of interest, and $z_{\max}(i)$ and $z_{\min}(i)$ are the maximum and minimum elevation within the *i*th spatial window. Here we use the TDI computed by Oyler et al. (2015) 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 multi-scale local minima and 5 being a multi-scale local maxima. 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 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 are within 1 °C of station data at nearly all GHCND stations (Fig. 2). The station-based datasets are also within 1 °C of the station-based gridded dataset average (henceforth, *the average*) for most of the domain. 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 along the coast by up to 3 °C. Comparing with GHCND data, it appears that they may be slightly too warm in some locations along the coast. Meanwhile, Livneh Tmax is colder than average in parts 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 are substantially cooler throughout the domain. On average, NLDAS-2 and WRF differ from the station-based gridded average by -1.4 °C and -1.1 °C, respectively. They are also consistently colder than GHCND data, so it's likely that they have a cold bias. Differences between WRF and the station-based dataset average are correlated with elevation (r = -0.67) and become increasingly negative by approximately 1.0 °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 the 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 topography, 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: e.g. 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 warmer than the average (domain-average difference of +2.9 °C). Thus, WRF is more realistic than NLDAS-2 for Tmin.

Inter-dataset 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). 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 over the 1981–2010 period differ substantially among the datasets (Fig. 5). There are clear differences in trends between those station-based datasets that do not correct for inhomogeneities and those that do. Daymet, Livneh, PRISM, and Metdata do not correct inhomogeneities and have large trends, exceeding 1 °C decade^{-1} in some locations. A comparison of time series at selected grid cells shows that these large trends are generally due to sudden jumps or shifts in temperature by up to 10 °C (Fig. 6). It is likely that these irregularities are inhomogeneities due to changes in station location, measurement techniques or other factors, and are not representative of actual conditions. Daymet and Livneh are the most strongly affected by inhomogeneities (Fig. 5). PRISM and Metdata are also affected, but to a lesser extent. In contrast, TopoWx and Hamlet correct for inhomogeneities and have smooth trends fields. However, they may be too smooth. For instance, along the coast, Tmax trends at GHCND stations are consistently negative. It is unlikely that they so many stations would agree on a modest negative trend if it were not actually the case. Yet, TopoWx and Hamlet have universally positive Tmax trends along the coast and throughout California. Thus, it is probably the case that TopoWx and Hamlet miss real trends in some locations.

c. Variability

All datasets have greater temperature variability at higher elevations (Fig. 7). 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 reduced over the ocean, these grid cells are likely to have lower variability relative to 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. NLDAS-2 has lower Tmin variability along a wider coastal strip than the other datasets, just like it does for Tmax. Meanwhile, Daymet and Livneh have isolated regions with Tmin variability as high as 3 °C, which are likely due to the same uncorrected 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, EOF 1 explains between 78% and 86% of the variance, depending on the dataset (Fig. 8). EOF 1 is characterized by positive loadings over all of California, with larger loadings at high elevations. PC 1 (the time series representing how EOF 1 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. 7). EOF 2 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 PC 2 time series is also high, although not as high as for PC 1. EOF 3 is another dipole mode, this time representing variability that is oppositely phased between coastal and inland locations (2–4% of the variance). The corresponding PC 3s agree less than PCs 1 or 2. Daymet's EOF 3 stands out for its irregular loading pattern, which could be related to uncorrected inhomogeneities. However, PC 3 explains only a small fraction of the variability.

For Tmin, EOFs and PCs differ somewhat more than Tmax (Fig. 9). EOF 1, characterized by all positive loadings, explains 63–81% of the variance, a wider range than for Tmax (77–86%). Daymet's EOF spatial patterns differ considerably from the others. They have much higher loadings in the same regions that have large, unphysical trends. Daymet's PC 1 and PC 2 time series show shifts from lower values in the 1980s to higher values in the 2000s that are not present in the other datasets, and are consistent with the inhomogeneities discussed earlier.

Inhomogeneities are also likely responsible for the unusual spatial pattern of Livneh's EOF 3. These results suggest that uncorrected station inhomogeneities can contribute non-negligible artifacts to a station-based dataset's variability.

PRISM, Metdata, and TopoWx appear to have the most plausible variability. Their main EOFs are free from artifacts and their PCs series don't 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. 2015). While excluding other 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 is free of inhomogeneity-related artifacts, which is an advantage. 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 fine-scale 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 snow albedo feedback (SAF) on temperature anomalies (Fig. 10). WRF simulates large differences in snow cover between April 2007 and April 2010, which are supported by the 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 show substantially enhanced temperature differences at grid cells where MODIS/Terra indicates snow cover loss. This disparity could be due to low station density at high elevations or overly simplistic relationships between temperature and elevation. An alternative possibility is that WRF's SAF strength is unrealistically high and actual temperature differences are not amplified as much as WRF suggests.

e. Local lapse rates

Tmax lapse rates are positive throughout the domain (Fig. 11a). They are 4–8 °C km^{-1} inland, but 2–4 °C km^{-1} for large portions of the coastal mountains. Thus, Livenh's use of a fixed 6.5 °C km^{-1} lapse rate could be suitable for Tmax inland, but not for portions of the coastal mountains. Meanwhile, a fixed 6.5 °C km^{-1} lapse rate is much less suitable for Tmin. Tmin lapse rates are generally near zero or even negative for most of California (Fig. 11b). Using a fixed lapse rate would have little effect if station density were high everywhere and all elevations were adequately sampled. However, this is not the case: station density is low in many parts of the domain and stations are often located near topographic minima (Fig. 12). Thus, many grid cells are far away from and at higher elevations than their nearest stations. At these grid cells, temperatures are determined by extrapolating up from the stations, and using a suitable lapse rate is most important. For example, in the coastal mountains of Northern California, station density is low and almost all stations are located near topographic minima (Fig. 13a). Tmin lapse rates in this area are typically near zero (Fig. 11b), which is substantially different from the fixed lapse rate of 6.5 °C km^{-1} used in Livneh. This explains why Livneh is cold relative to the station-based gridded datasets average here (Fig. 13b), and differences become increasingly negative with height by 2.9 °C km^{-1} (Fig. 13c).

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 (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 has not been included in previous assessments of gridded datasets.

As expected, when evaluated at station locations, station-based datasets all have similar climatologies that closely match GHCND station data. Differences in station-based datasets are more pronounced away from stations, where interpolation algorithms have greater impact. For Tmax, the largest differences (up to 6 °C) occur in isolated parts of the coastal mountains. For Tmin, differences are large (2–6 °C) in complex terrain throughout the domain. The existence of large differences away from the stations despite close agreement at stations, underscores the need to compare station-based datasets 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 some systematic biases relative to GHCND station data. WRF has a cold bias at high elevations for Tmax, but not for Tmin. Meanwhile, NLDAS-2 has a dramatic cold bias (> 6 °C) along the edges of topographic contours for Tmax and a strong warm bias throughout the domain in Tmin.

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. In contrast, the other datasets allow for the relationship between temperature and elevation to vary throughout the domain. 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 is 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 a necessity when studying daily minimum temperatures.

Differences in trends are the result of choices made about which stations to include and whether to perform station homogenization. Daymet, Livneh, PRISM, and Metdata do not correct station data for inhomogeneities and subsequently have large non-climatic trends near affected stations. In contrast, TopoWx, Hamlet, NLDAS-2, and WRF have smoother trend fields. TopoWx applies a homogenization algorithm that removes inhomogeneities by comparing a station with its neighbors. This removes the obvious jumps in temperature due to non-climatic factors. However, our findings also suggest it may smooth out true, local trends, forcing the regional trend on each grid cell, a known side effect of homogenization algorithms (Pielke et al. 2007). Hamlet also accounts for inhomogeneities, but in a different way: by adjusting the low-frequency variability at all stations to match the small set of USHCN stations. This leads to overly smoothed variability in Hamlet. WRF and NLDAS-2 appear to have realistic trends. However, it is possible for reanalysis to suffer from non-climatic trends if inhomogeneities in the assimilated station data or changes in data coverage are not accounted for. Interpolated reanalysis, such as NLDAS-2, will directly inherit any problems. Dynamically downscaled reanalysis may be less affected since it generally receives input only at the lateral and ocean boundaries. Overall, users should be aware that certain gridded datasets have unphysical jumps or trends, while others may have overly smooth trends.

Most datasets have broad agreement in the spatial patterns and the timing of the leading modes. Daymet and Livneh are the main exceptions. Inhomogeneities strongly affect these datasets, causing prominent non-climatic artifacts in the spatial patterns and jumps in the associated time series. These datasets also have unrealistically high variability in some locations. NLDAS-2 is unusual because it has significantly reduced variability very near the coast.

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. Its spatial patterns of the leading modes are qualitatively similar to the most plausible station-based datasets. WRF's variability is clearly more realistic than some station-based datasets, such as Daymet, which is strongly affected by inhomogeneities. Its trends are most similar to TopoWx and Hamlet, the two datasets that correct for inhomogeneities. These results suggest that dynamical downscaled reanalysis can produce a spatially complete picture of the historical temperatures on par with the 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 of 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.

Large differences between gridded datasets indicate gridded dataset choice is a considerable source of uncertainty. Uncertainty in station-based datasets could be reduced with straightforward fixes, like making variable lapse rates and station homogenization standard practice. Further reductions could come from identifying best-performing interpolation algorithms. This could be done by running cross-validation tests on a standardized set of stations. Improving realism in downscaled reanalysis is less straightforward. Improvements are likely to come from ongoing progress in improving atmospheric models and regional climate models used to generate and downscale reanalysis.

It is important that users of gridded datasets are aware of their limitations. Often station-based gridded datasets are treated as ground truth, without acknowledging either the problems with station data or the assumptions needed to generate a spatial complete temperature field from point measurements. This often plays out in the context of climate model evaluation, where models are compared against a station-based gridded dataset, and any differences are attributed to model deficiencies. In fact, a model may not be wrong just because it differs from a single station-based dataset, especially if the station-based dataset has known problems. More generally, it is recommended that users employ two or more independent gridded datasets to test the sensitivity of their results to dataset choice. Note that selecting closely related datasets, like PRISM and Metdata, could dramatically underestimate this important source of uncertainty.

Acknowledgments

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. Clim.*, **33**, 121–131, doi: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. *Ecological Applications*, **26**(5), 1338–1351. http://dx.doi.org/10.1002/15-1061

Behnke. R., S. Vavrus, A. Allstadt, T. Albright, W. Thogmartin, and V. Radeloff, 2016b: Data from: Evaluation of downscaled, gridded climate data for the conterminous United States. Dryad Digital Repository, doi: http://dx.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(8): e70260. doi:10.1371/journal.pone.0070260.

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**(D22), 8842, doi:10.1029/2002JD003118.

Cubasch, U., and Coauthors, 2001: Projections of future climate change. Climate Change 2001: The Scientific Basis, J. W. Kimand J. Stone, Eds., Cambridge University Press, 525–582.

Daly, C., 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, doi:10.1175/ 1520-0450(1994)033,0140:ASTMFM.2.0.CO;2.

Daly, C., 2006: Guidelines for assessing the suitability of spatial climate data sets. *Int. J. Climatol.*, **26**, 707–721, doi:10.1002/joc.1322.

Daly C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, P. P. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031–2064, doi: 10.1002/joc.1688.

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

Dozier, J., 1996: A generalized split-window algorithm for retrieving land-surface temperature from space. *IEEE Trans. Geosci. Remote Sens.*, **34**, 892–905, doi:10.1109/36.508406.

Durre I., M. J. Menne, B. E. Gleason, T. G. Houston, R. S. Vose, 2010: Comprehensive automated quality assurance of daily surface observations. *J. Appl. Meteorol. Climatol.*, **49**, 1615–1633, doi: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, doi:10.5067/IPPLURB6RPCN.

Hamlet A.F. and Lettenmaier D.P., 2005: Production of temporally consistent gridded precipitation and temperature fields for the continental U.S., 2005: *J. Hydrometeorology* 6 (3), 330-336, doi: http://dx.doi.org/10.1175/JHM420.1.

Hamlet, A., and Coauthors, 2010: Chapter 3, “Historical Meteorological Driving Data”. From Final project report for the Columbia basin climate change scenarios project. Report, University of Washington, Seattle, WA. Accessed 2017-1-18. Available online at: 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, doi: 10.1175/2009JCLI2470.1.

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

Holden, Z. A., J. T. Abatzoglou, C. H. Luce, L. S. Baggett, 2011: Empirical downscaling of daily minimum air temperature at very fine resolutions in complex terrain. *Agric. For. Meteorol.*, **151**, 1066–1073, doi:10.1016/j.agrformet.2011.03.011.

Holland, M. M., and C. M. Bitz, 2003: Polar amplification of climate change in coupled models. *Clim. Dyn.*, **21**, 221–232.

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, doi: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. Nat. Acad. Sci.*, **107**(10), 4533–4538, doi:10.1073/pnas.0915062107.

Juang, H.-M. H., and M. Kanamitsu, 1994: The NMC Nested Regional Spectral Model. *Mon. Weath. Rev.*, **122**(1), 3–26, doi:10.1175/1520-0493(1994)122<0003:TNNRSM>2.0.CO;2.

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

Kalnay, E. and Coauthors,1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**(3), 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

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**(19), 7576–7595, doi: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, doi:10.1175/JCLI-D-12-00508.1.

Lundquist JD, Pepin N, Rochford C. 2008. Automated algorithm for mapping regions of cold-air pooling in complex terrain. *J. Geophys. Res.*, **113**, D22107, doi: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, doi:10.1175/ 1520-0442(2002)015,3237:ALTHBD.2.0.CO;2.

Menne, M. J., C. N. Williams, and R. S. Vose, 2009: The United States Historical Climatology Network Monthly Temperature Data - Version 2. *Bull. Amer. Meteor. Soc.*, **90**, 993-1007, doi: 10.1175/2008BAMS2613.1.

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

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

Menne, M. J., and Coauthors, 2012b: Global Historical Climatology Network - Daily (GHCN-Daily), Version 3. NOAA National Climatic Data Center. doi:10.7289/V5D21VHZ.

Menne, M. J., C. N. Williams, Jr., and R. S. Vose, 2015. United States Historical Climatology Network Daily Temperature, Precipitation, and Snow Data. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, doi: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, doi: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. Hydrometeorol.*, **15**, 474–488, doi: http://dx.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**(1), 39–49, doi:10.1175/BAMS-86-1-39.

Newman, A.J., and Coauthors 2015: Gridded ensemble precipitation and temperature estimates for the contiguous United States. *J. Hydrometeorol.*, **16**(6), 2481–2500, doi:10.1175/JHM-D-15-0026.1.

Niu, G.-Y., et al. (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, doi: 10.1029/2010JD015139.

Oyler, J. W., A. Ballantyne, K. Jencso, M. Sweet, and S. W. Running, 2015: 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, doi:10.1002/joc.4127.

Pierce, D.W., D. R. Cayan, B. L. Thrasher, 2014: Statistical Downscaling Using Localized Constructed Analogs (LOCA). *J. Hydrometeorol.*, **22**, 2558:2585, doi:10.1175/JHM-D-14-0082.1.

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

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, doi:10.1175/2010JCLI3985.1.

Salathé, E. P., R. Steed, C. F. Mass, 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, doi: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**(2), 137–161, doi:10.14430/arctic407.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Technical Note, NCAR/TN-4751+STR.

Stahl, K., R. D. Moore, J. A. Floyer, M. G. Asplin, and I. G. McKendry, 2006: Comparison of approaches for spatial interpolation interpolation of daily air temperature in a large region with complex topography and highly variable station density. *Agric. For. Meteor.*, **139**, 224–236, doi: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. *Clim. Dyn.*, **38**(11-12), 2449-2466, doi: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 in Ecology and Evolution*, **6**, 412–423, doi: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. Hydrology*. **190**, 214–251, doi: 10.1016/S0022-1694(96)03128-9.

Thornton, P. E., 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, Tennessee, USA. http://dx.doi.org/10.3334/ORNLDAAC/1328 (accessed 19 November 2016).

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**(5), 1232–1251, doi:10.1175/JAMC-D-13-0248.1.

Wan Z. 2008. New refinements and validation of the MODIS land-surface temperature/emissivity products. *Remote Sens. Environ.*, **112**, 59–74, doi:10.1016/j.rse.2006.06.026.

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**(12), 4597–4617, doi:10.1175/JCLI-D-14-00196.1.

Walton, D. B., 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**(4), 1417-1438, doi:10.1175/JCLI-D-16-0168.1.

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, doi:10.1029/2011JD016048.

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

Dataset name	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 sub-daily data)	interpolated reanalysis	dynamically downscaled reanalysis
Citation	Daly et al. 2008	Oyler et al. 2015	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	ftp.hydro.washington.edu/pub/bLivn	Contact Mu Xiao muxiao@ucla.edu for extension; original available	http://climate.nkn.uidaho.edu/METDATA/	https://disc.gsfc.nasa.gov/SSW	http://research.atmos.ucla.edu/csrl/

				eh/CONUS/	from http://www.hydro.washington.edu/Lettenmaier/Data/gridded/index_hamlet.html			data/Gridded_Datasets/
Native Resolution	2.5 min (~4 km)	30 arcsec (~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,	GHCN-Daily, SNOTEL, RAWS	GHCN-Daily (Menne et al., 2012)	COOP	COOP, Env. Canada, USHCN, HCCD	NLDAS-2 for daily variability, PRISM for monthly means	NARR	NARR

	Agrimet, others							
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	Regression kriging for climate normals with auxiliary predictors including lat, lon,	Truncated Gaussian filter combined with	SYMAP algorithm: inverse distance weighting with	SYMAP algorithm: inverse distance weighting with directional adjustment; daily data adjusted so	Bi-linearly interpolatied NLDAS-2 daily anomalies from monthly	Bi-linear interpolation of NARR to 1/8° spatial resolution	WRF coupled to Noah-MP

	based on multiple physical factors	elev, satellite LST; moving-window geographically weighted regression and inverse distance weighting for anomalies	elevation-regression	directional adjustment	that monthly climate normals match PRISM for 1971-2000	means added to PRISM monthly means	and hourly temporal resolution	
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

Figure Captions

Fig 1. (a) Setup of 27 km resolution and 9 km resolution nested WRF domains. (b) Locations of COOP stations used by Livneh, and GHCND, RAWS, and SNOTEL stations used by TopoWx.

Fig 2. (top left) 1981–2010 Tmax annual-mean climatology at GHCND stations and averaged over the station-based datasets (units: °C). (others) Differences in 1981–2010 annual-mean Tmax climatology with GCHND station data and with the station-based dataset average (units: °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.

Fig 3. (top left) 1981–2010 Tmin annual-mean climatology at GHCND stations and averaged over the station-based datasets (units: °C). (others) Differences in 1981–2010 annual-mean Tmax climatology with GCHND station data and with the station-based dataset average (units: °C). Note: no elevation-based adjustments are made for Tmin.

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

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

Fig 6. Tmin anomalies (°C) at three grid cells where some datasets show inhomogeneities. Locations of the three grid cells are shown on the left.

Fig 7. Standard deviation (°C) of monthly Tmax and Tmin anomalies for all months in the period 1981-2010. For GHCND, only anomalies from non-missing months are used.

Fig 8. Three largest Tmax EOFs and their associated PCs for each dataset for the period 1981-2010. Percentages of explained variance are included in the upper right corner of each panel.

Fig 9. Three largest Tmin EOFs and their associated PCs for each dataset for the period 1981-2010. Percentages of explained variance are included in the upper right corner of each panel.

Fig 10. Differences in (a) MODIS snow covered fraction (b) WRF SCF, and (c-j) daily average temperatures (°C) for each dataset, computed as April 2007 minus April 2010.

Fig 11. Local lapse rate (°C km^{-1}) calculated at each grid cell as the negative slope determined by linearly regressing TopoWx climatological (a) Tmax and (b) Tmin values onto elevation for all grid cells whose *x* and *y* coordinates are each within 1 km. 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 calculations.

Fig 12. Topographic dissection index (TDI) at each California COOP station used by Livneh. Warm colors indicate stations located near topographic maxima. Cold colors indicate locations near topographic minima.

Fig 13. (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, Holden et al. 2011a) is plotted at each COOP station (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) versus elevation at all grid cells within the coastal region shown in (a) and (b). Slope computed using least-squares linear regression.

Figures

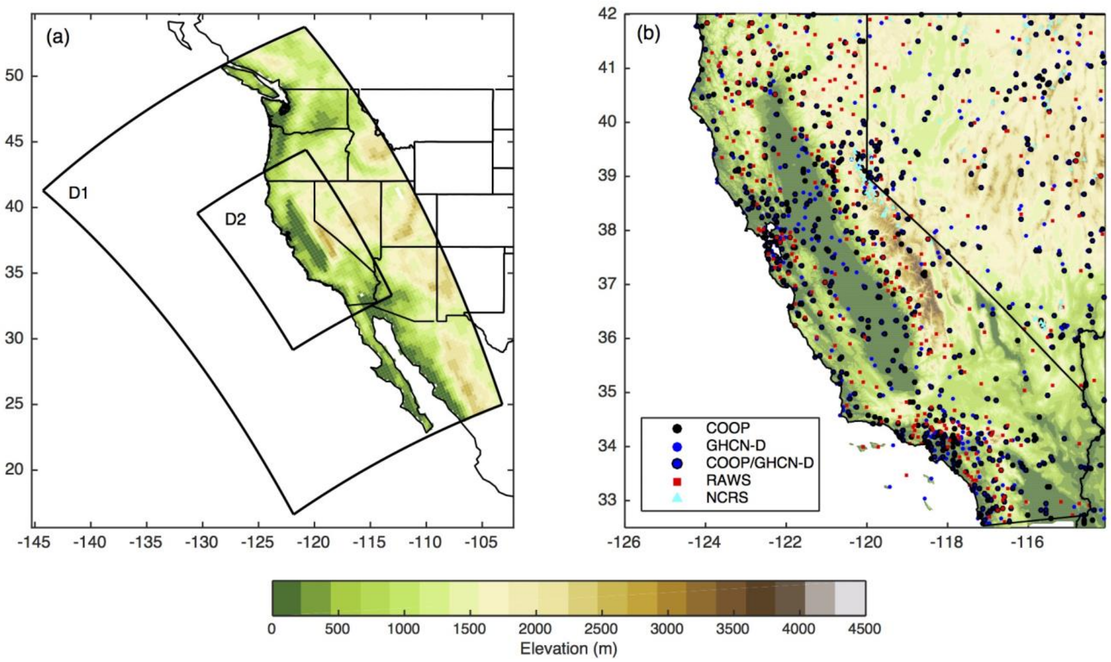


Fig 1. (a) Setup of 27 km resolution and 9 km resolution nested WRF domains. (b) Locations of COOP stations used by Livneh, and GHCND, RAWS, and SNOTEL stations used by TopoWx.

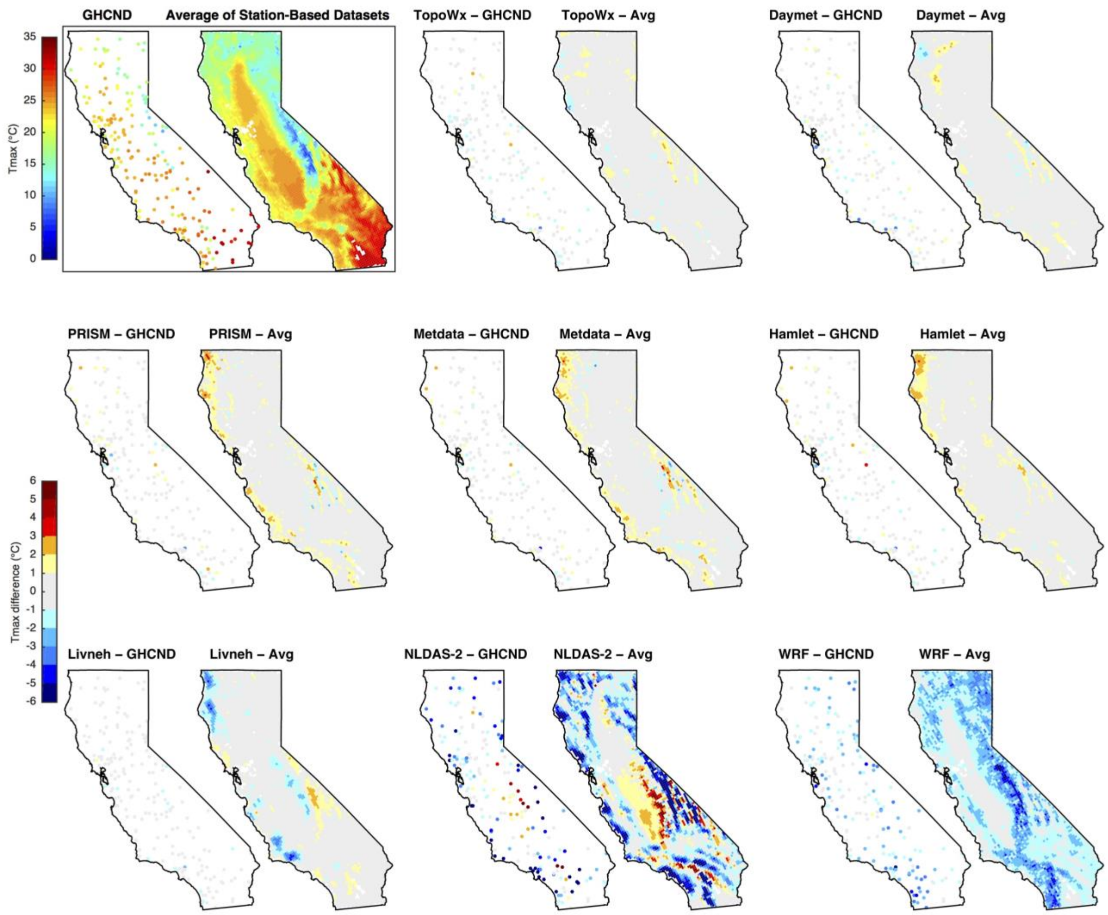


Fig 2. (top left) 1981–2010 Tmax annual-mean climatology at GHCND stations and averaged over the station-based datasets (units: °C). (others) Differences in 1981–2010 annual-mean Tmax climatology with GCHND station data and with the station-based dataset average (units: °C). To adjust for the elevation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

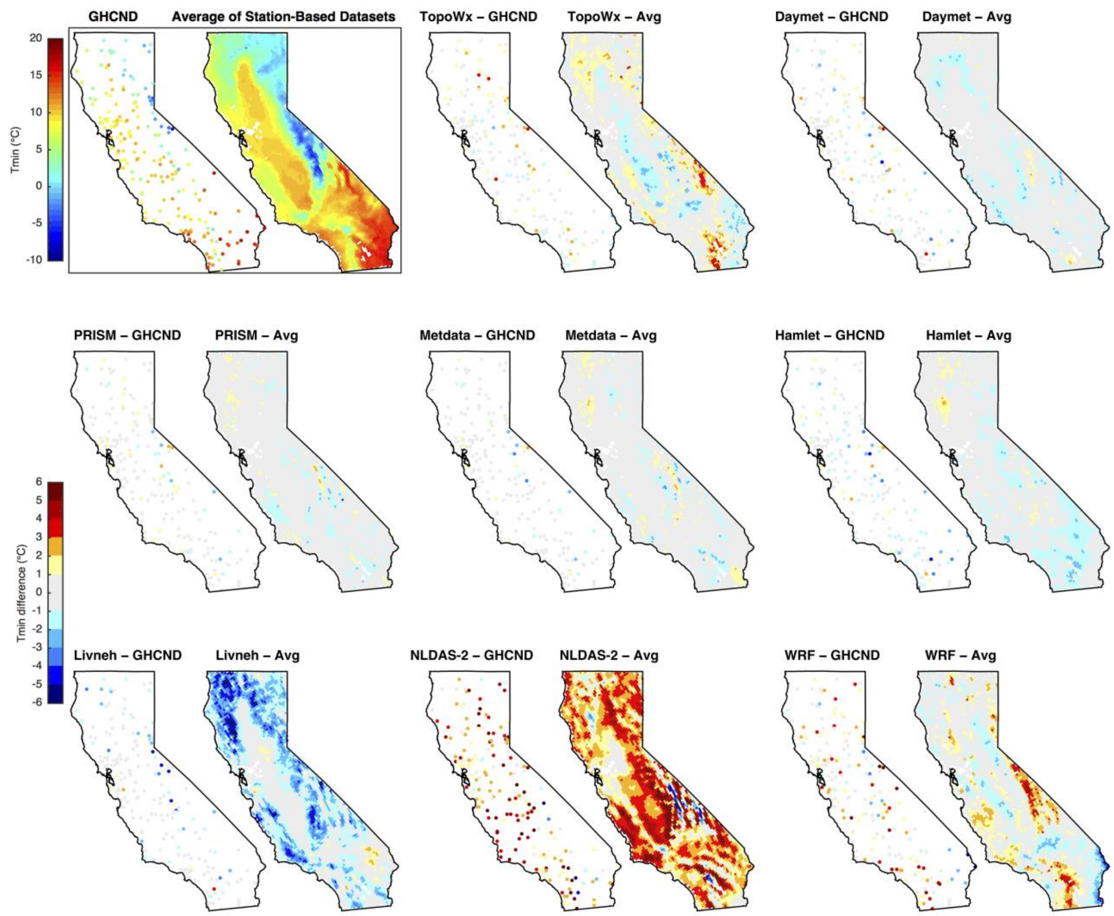


Fig 3. (top left) 1981–2010 Tmin annual-mean climatology at GHCND stations and averaged over the station-based datasets (units: °C). (others) Differences in 1981–2010 annual-mean Tmax climatology with GCHND station data and with the station-based dataset average (units: °C). Note: no elevation-based adjustments are made for Tmin.

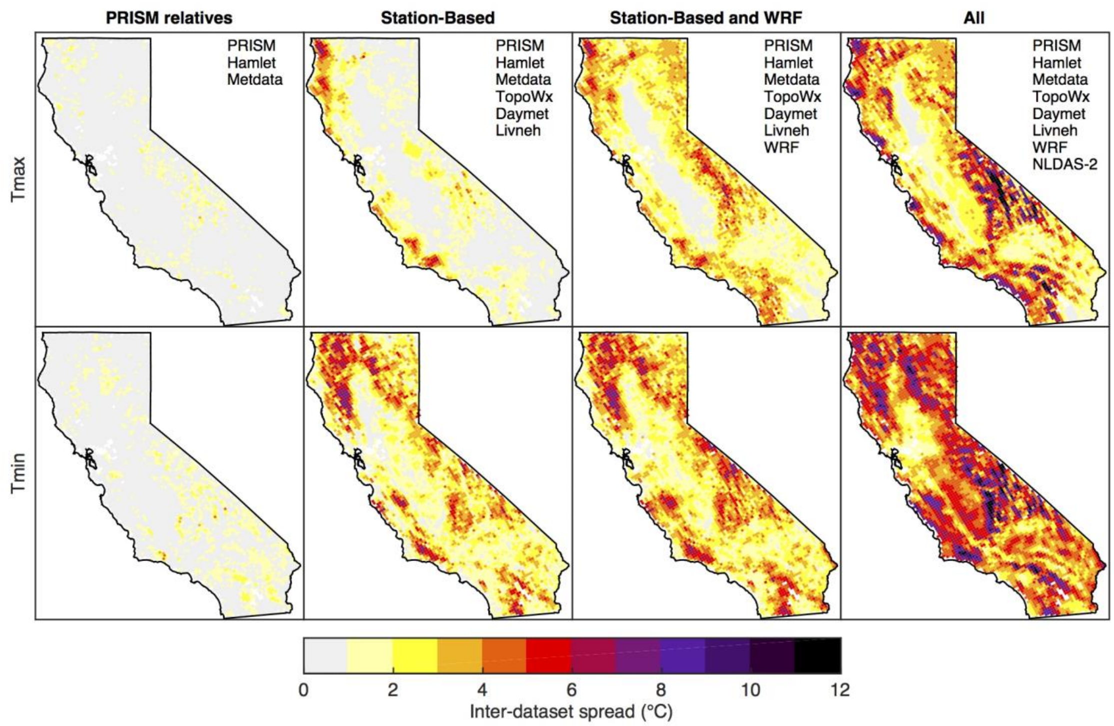


Fig 4. Inter-dataset spread (°C) in climatological Tmax (top row) and Tmin (bottom row) calculated for four different groups. Datasets included in each group are listed in the upper right corner of each panel.

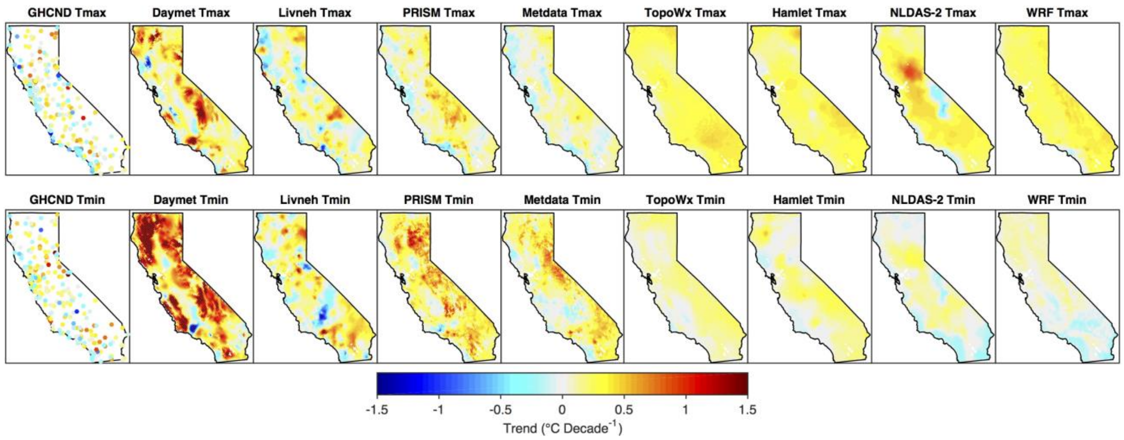


Fig 5. Trend (°C decade^{-1}) in Tmax (top row) and Tmin (bottom row) based on linear regression of monthly anomalies for all months in 1981–2010 time period. For GHCND, only anomalies from non-missing months are used.

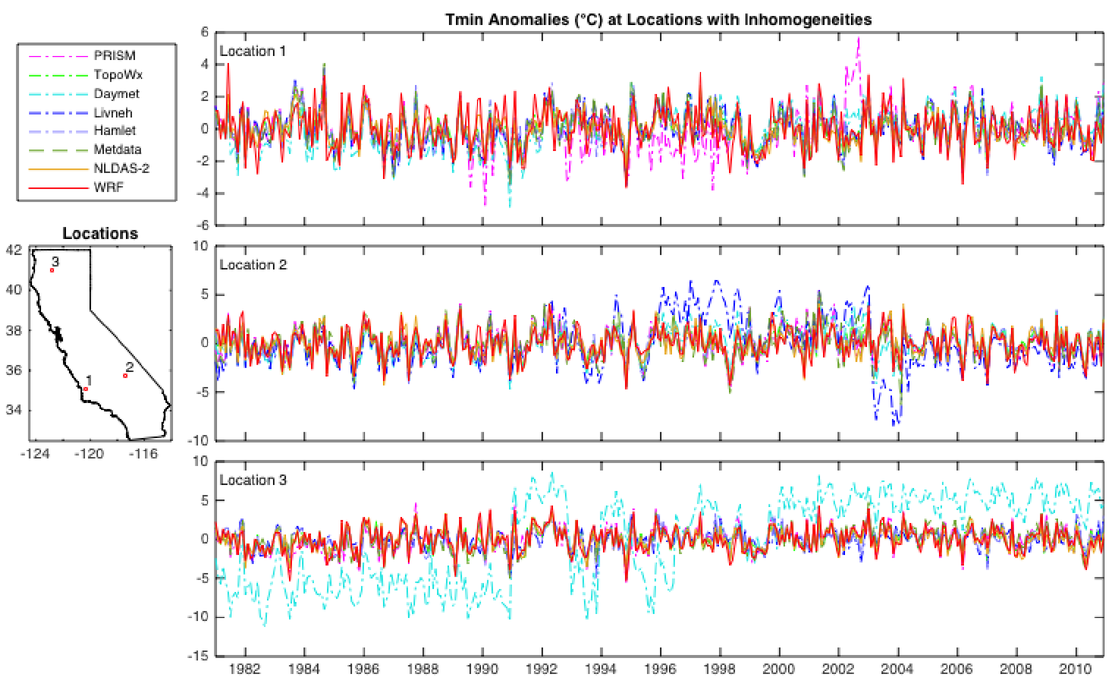


Fig 6. Tmin anomalies (°C) at three grid cells where some datasets show inhomogeneities. Locations of the three grid cells are shown on the left.

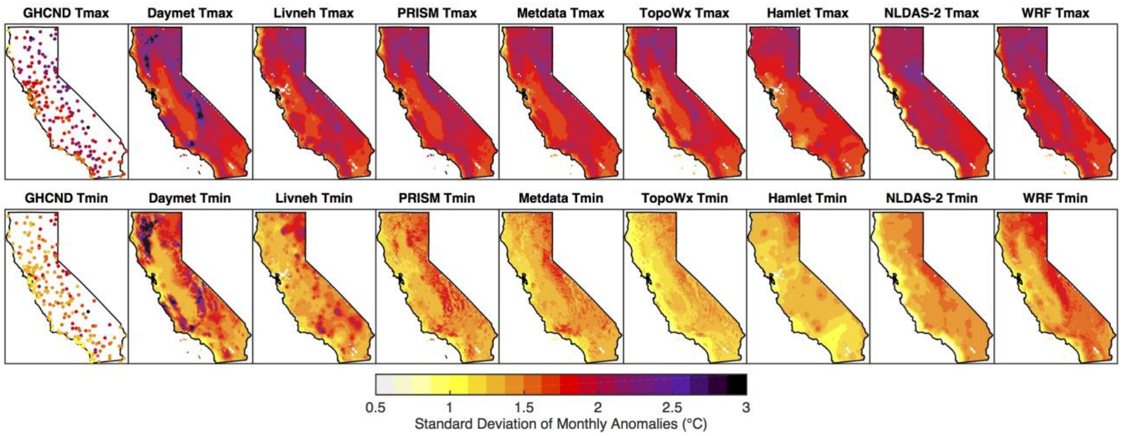


Fig 7. Standard deviation (°C) of monthly Tmax and Tmin anomalies for all months in the period 1981-2010. For GHCND, only anomalies from non-missing months are used.

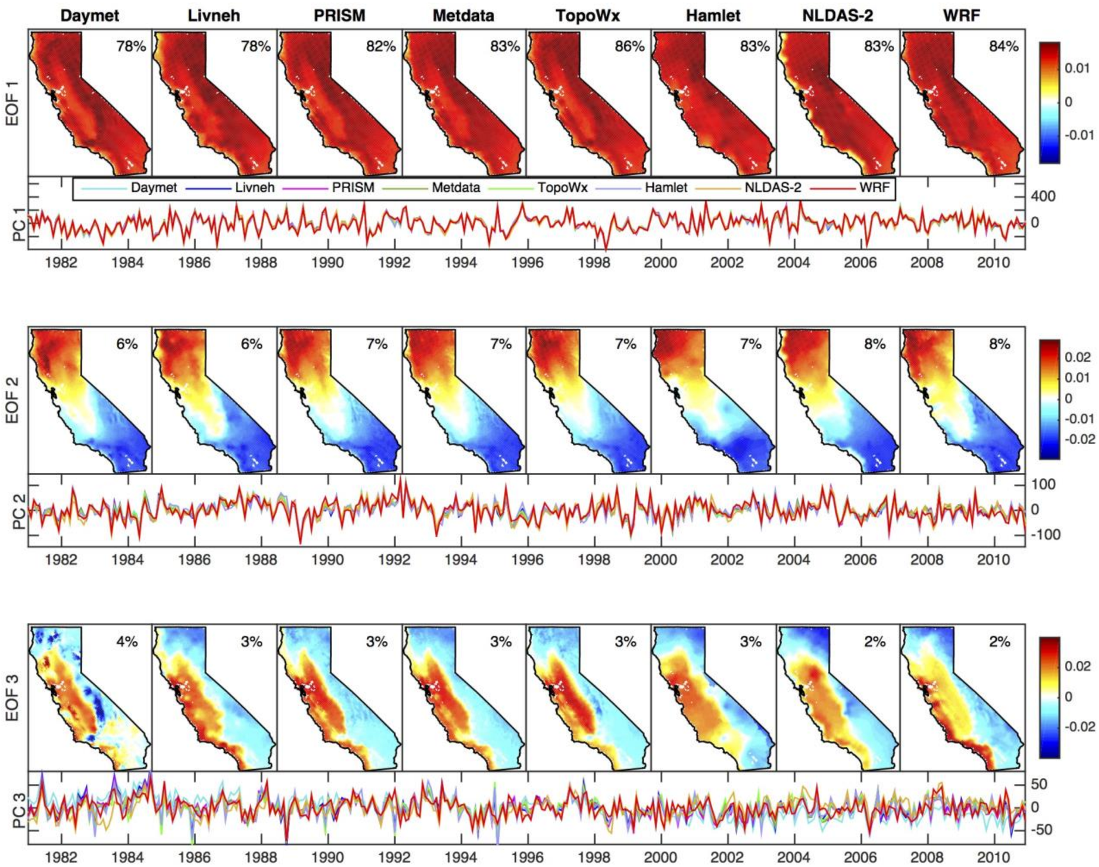


Fig 8. Three largest Tmax EOFs and their associated PCs for each dataset for the period 1981-2010. Percentages of explained variance are included in the upper right corner of each panel.

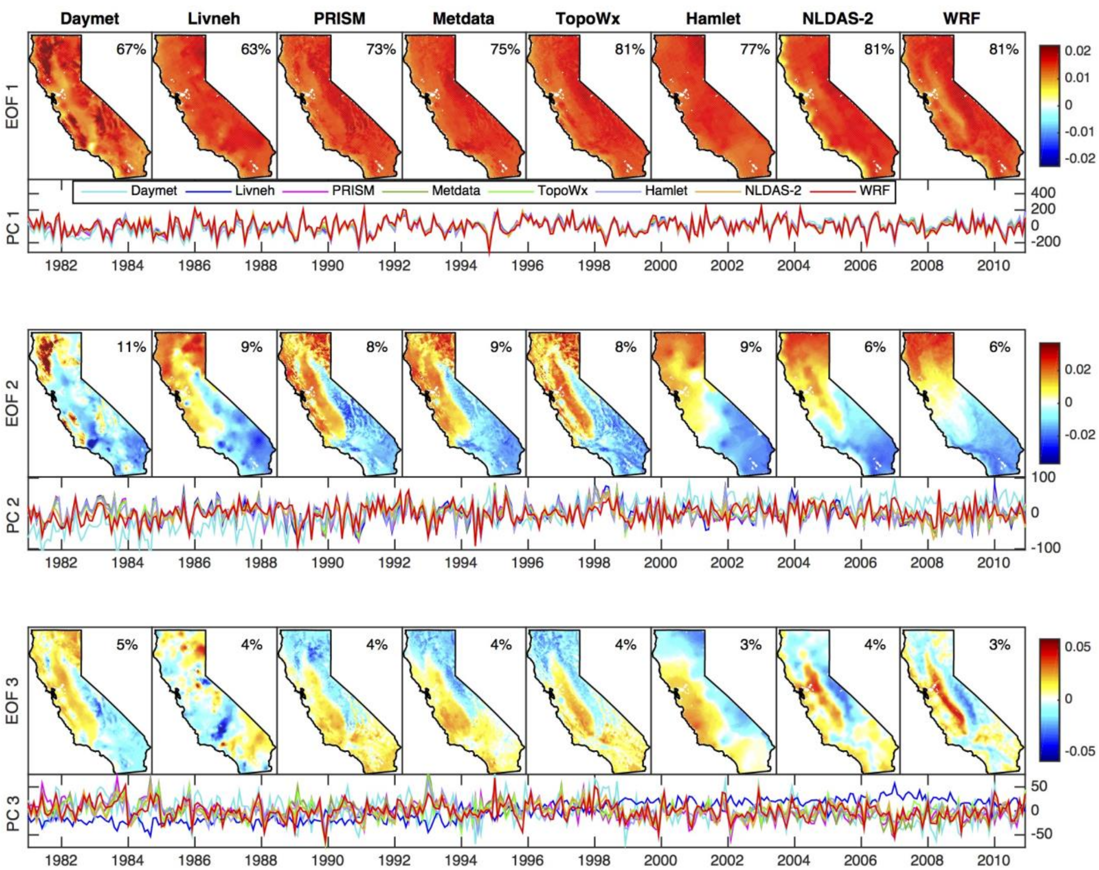


Fig 9. Three largest Tmin EOFs and their associated PCs for each dataset for the period 1981-2010. Percentages of explained variance are included in the upper right corner of each panel.

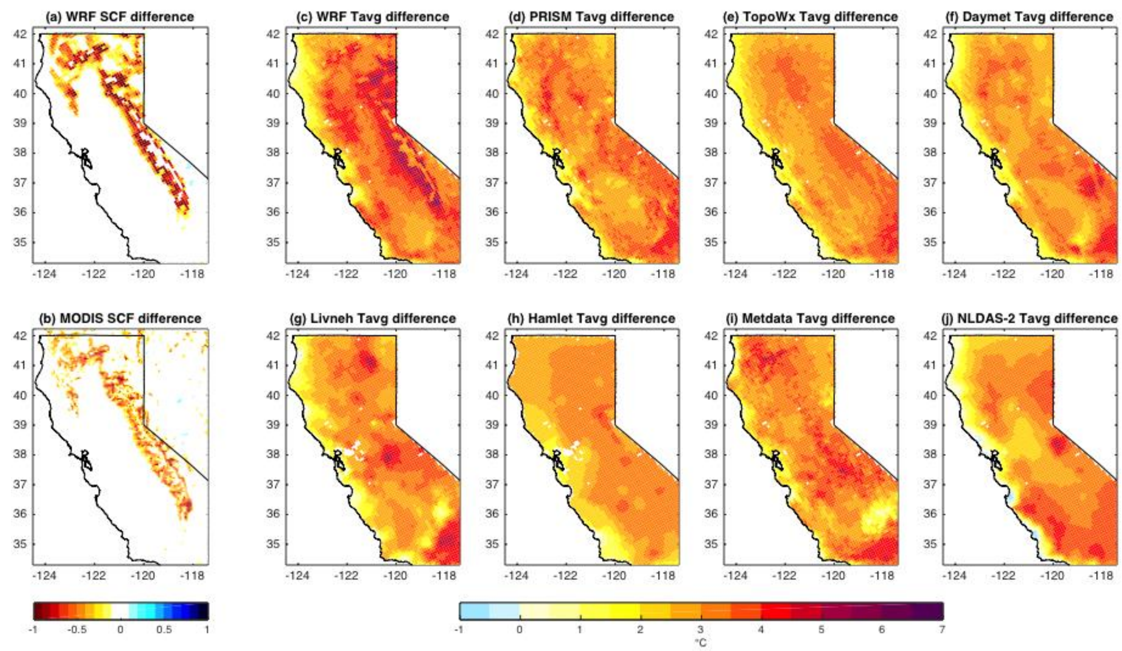


Fig 10. Differences in (a) MODIS snow covered fraction (b) WRF SCF, and (c-j) daily average temperatures (°C) for each dataset, computed as April 2007 minus April 2010.

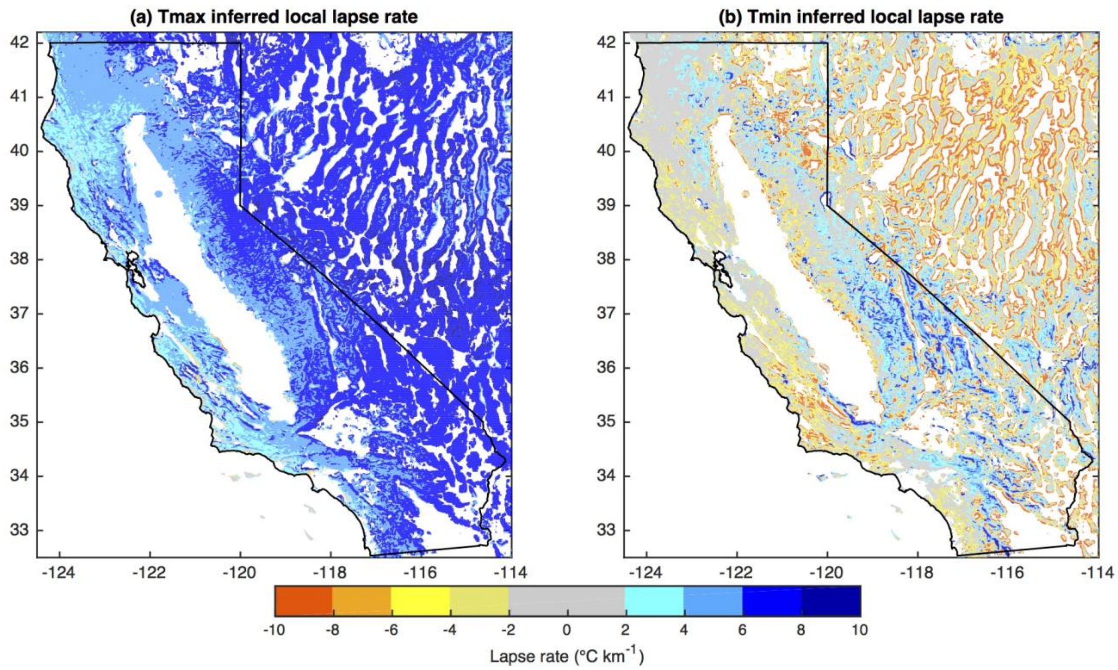


Fig 11. Local lapse rate (°C km^{-1}) calculated at each grid cell as the negative slope determined by linearly regressing TopoWx climatological (a) Tmax and (b) Tmin values onto elevation for all grid cells whose *x* and *y* coordinates are each within 1 km. Cool colors indicate decreasing temperature 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 calculations.

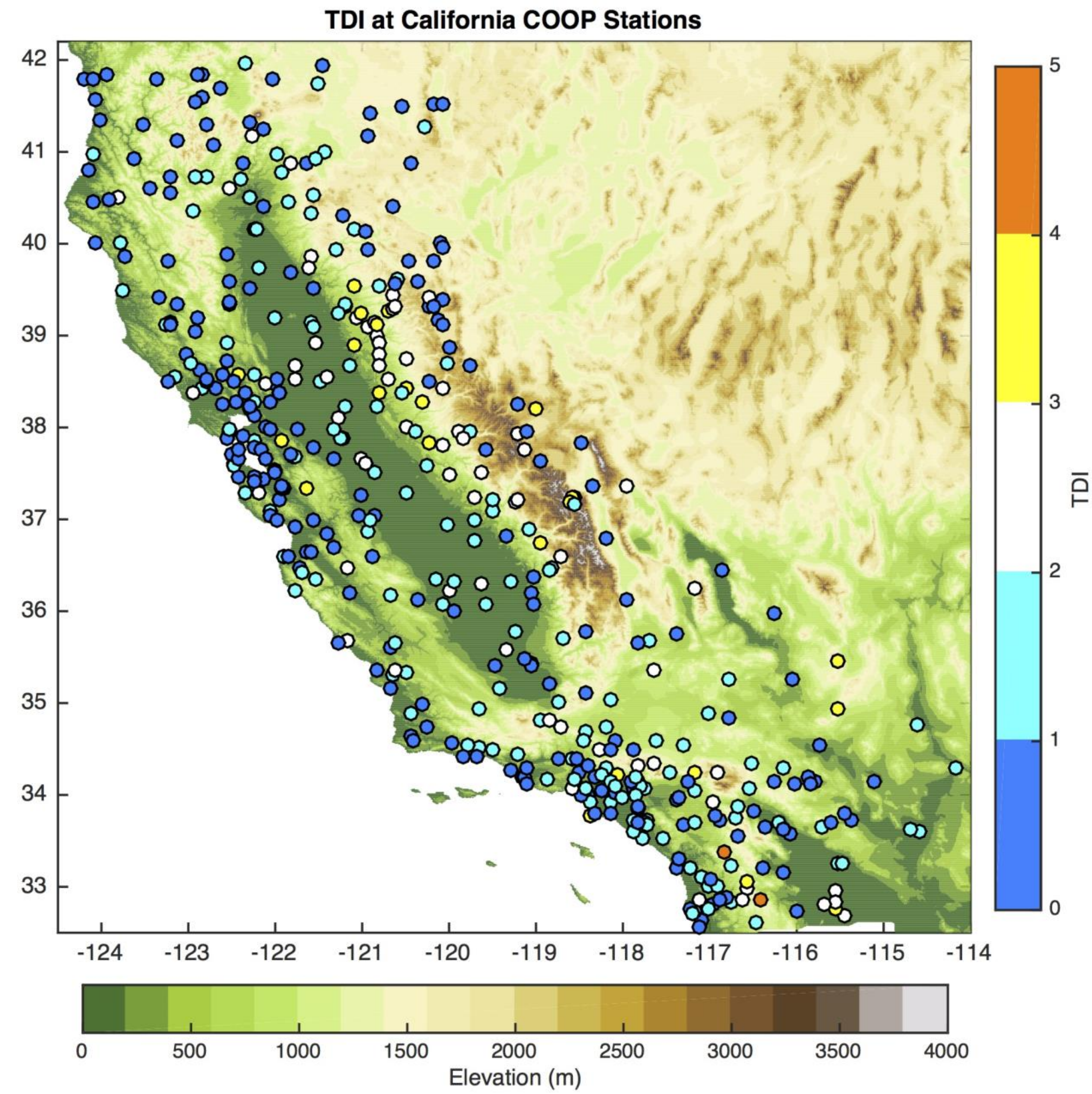
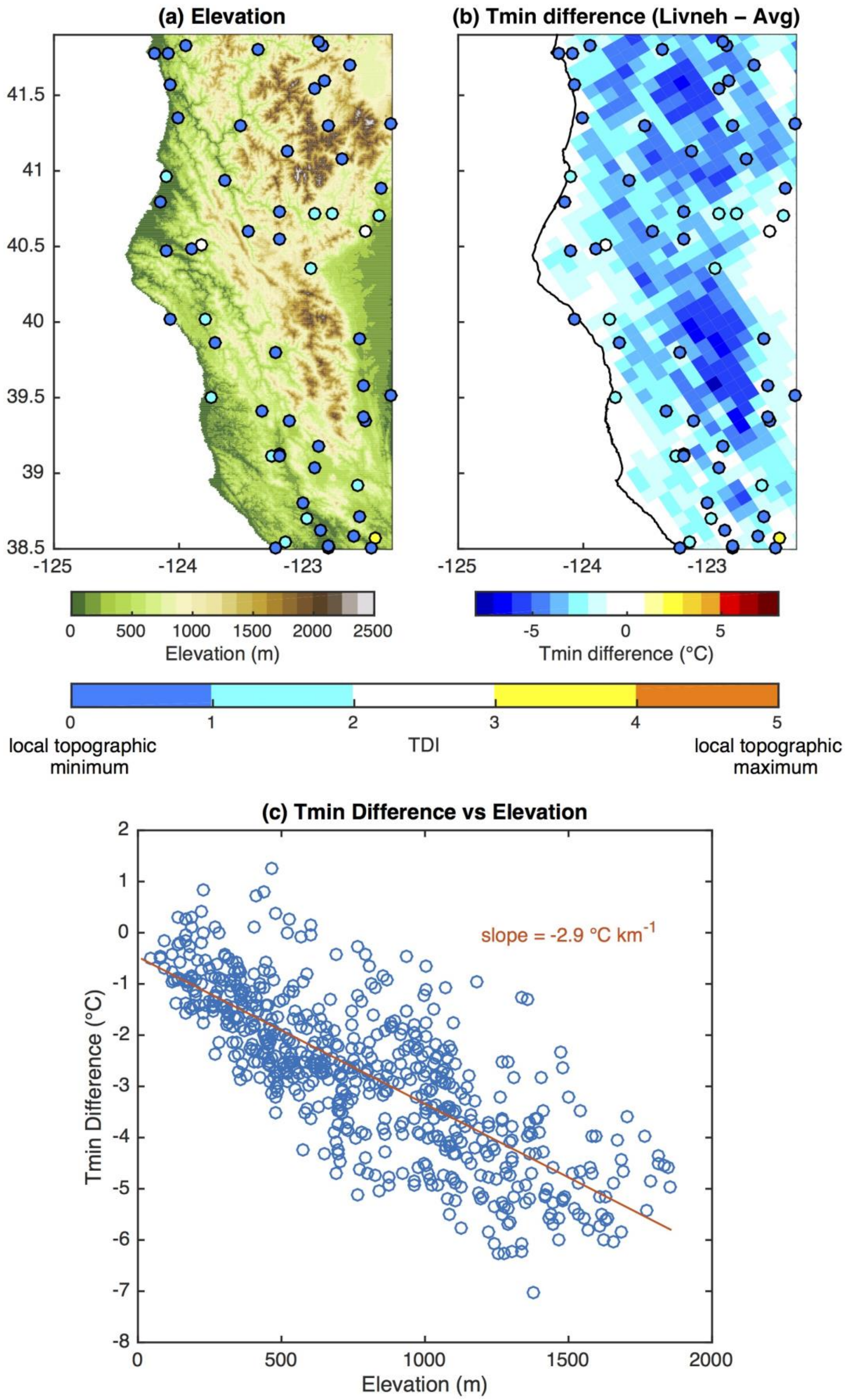


Fig 12. Topographic dissection index (TDI) at each California COOP station used by Livneh. Warm colors indicate stations located near topographic maxima. Cold colors indicate locations near topographic minima.



(a) Elevation
(b) Tmin difference (Livneh – Avg)
41.5
41
40.5
40
39.5
39
38.5
-125
-124
-123
0 500 1000 1500 2000 2500
Elevation (m)
-5 0 5
Tmin difference (°C)
0 1 2 3 4 5
local topographic minimum
TDI
local topographic maximum
(c) Tmin Difference vs Elevation
slope = -2.9 °C km-1
Tmin Difference (°C)
Elevation (m)
2 1 0 -1 -2 -3 -4 -5 -6 -7 -8
0 500 1000 1500 2000

Fig 13. (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, Holden et al. 2011a) is plotted at each COOP station (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) versus elevation at all grid cells within the coastal region shown in (a) and (b). Slope computed using least-squares linear regression.