


DOCKETED

Docket Number:	19-ERDD-01
Project Title:	Research Idea Exchange
TN #:	231300-3
Document Title:	High-resolution mapping of daily climate variables by aggregating multiple spatial data sets with the random forest algorithm
Description:	Peer-reviewed journal paper by Hashimoto et al, published in International Journal of Climatology
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

RESEARCH ARTICLE

High-resolution mapping of daily climate variables by aggregating multiple spatial data sets with the random forest algorithm over the conterminous United States

Hirofumi Hashimoto^{1,2}  | Weile Wang^{1,2} | Forrest S. Melton^{1,2} | Adam L. Moreno³ | Sangram Ganguly² | Andrew R. Michaelis^{1,2} | Ramakrishna R. Nemani²

¹School of Natural Sciences, California State University-Monterey Bay, Seaside, California

²NASA Earth Exchange, NASA Ames Research Center, Moffett Field, California

³Bay Area Environmental Research Institute, Petaluma, California

Correspondence

Hirofumi Hashimoto, School of Natural Sciences, California State University-Monterey Bay, Seaside, CA.

Email: hirofumi.hashimoto@gmail.com

High-resolution gridded climate data products are crucial to research and practical applications in climatology, hydrology, ecology, agriculture, and public health. Previous works to produce multiple data sets were limited by the availability of input data as well as computational techniques. With advances in machine learning and the availability of several daily satellite data sets providing unprecedented information at 1 km or higher spatial resolutions, it is now possible to improve upon earlier data sets in terms of representing spatial variability. We developed the NEX (NASA Earth Exchange) Gridded Daily Meteorology (NEX-GDM) model, which can estimate the spatial pattern of regional surface climate variables by aggregating several dozen two-dimensional data sets and ground weather station data. NEX-GDM does not require physical assumptions and can easily extend spatially and temporally. NEX-GDM employs the random forest algorithm for estimation, which allows us to find the best estimate from the spatially continuous data sets. We used the NEX-GDM model to produce historical 1-km daily spatial data for the conterminous United States from 1979 to 2017, including precipitation, minimum temperature, maximum temperature, dew point temperature, wind speed, and solar radiation. In this study, NEX-GDM ingested a total of 30 spatial variables from 13 different data sets, including satellite, reanalysis, radar, and topography data. Generally, the spatial patterns of precipitation and temperature produced were similar to previous data sets with the exception of mountain regions in the western United States. The analyses for each spatially continuous data set show that satellite and reanalysis led to better estimates and that the incorporation of satellite data allowed NEX-GDM to capture the spatial patterns associated with urban heat island effects. The NEX-GDM data is available to the community through the NEX data portal.

KEYWORDS

daily surface climate, machine learning, NEX-GDM, precipitation, random forest, solar radiation and wind speed, temperature

1 | INTRODUCTION

Long-term and high-resolution mapping of surface climate variables has been crucial for many applications that aim to

analyse regional land ecosystem response to climate patterns, including input for ecosystem modelling (Huntzinger *et al.*, 2012; Abatzoglou, 2013), validation for regional climate model output (Wang and Kotamarthi, 2014),

downscaling future climate projections (Pierce *et al.*, 2014), forecasting natural hazards (Nemani *et al.*, 2009), and many others. There have been previously produced high-resolution climate data sets that are publicly available and ready to use for ecosystem modelling; however, each data set has its own purpose, temporal and spatial resolution, time frame, and climate variables. For example, Livneh *et al.* (2013) produced a century-scale daily data set, but with relatively low resolution ($1/16^\circ$) for the purpose of supporting hydrometeorological modelling. Other data sets are limited to only precipitation and/or temperature (Vose *et al.*, 2014; Newman *et al.*, 2015; Oyler *et al.*, 2015a), which are not sufficient to drive ecosystem models. Currently, PRISM (Parameter-elevation Regressions on Independent Slopes Model) (Daly *et al.*, 1994) and Daymet (Thornton *et al.*, 1997) are the most commonly used for ecosystem modelling because those data sets include variables meeting the meteorological input requirements of the ecosystem models. The availability of a small number of data sets for forcing ecosystem models makes it difficult to assess the uncertainty of the statistics derived from ecosystem models (Wu *et al.*, 2017). Therefore, development of a new interpolation methodology that incorporates more observational input data sets is necessary for further climate and ecosystem studies, as this is important for improving the accuracy of these data sets. We developed the NASA Earth Exchange (NEX) Gridded Daily Meteorology (NEX-GDM) model to generate high-resolution climate variables by aggregating high volumes of spatial data through a machine learning technique. By using several high-volume data sets archived in the NEX facilities, NEX-GDM was able to create 1-km daily climate data sets for the conterminous United States that are continuous in both time and space.

Historically, due to the difficulty of high-resolution dynamic downscaling through regional climate models, spatial interpolation of weather station data has been the most reliable technique to account for spatial variability at the regional scale. Various spatial interpolation methodologies were developed to spatially interpolate weather station data; for example, Thiessen polygons (Thiessen, 1911), inverse-distance (Willmott *et al.*, 1985), splines (Hutchinson, 1995), and kriging (Jolly *et al.*, 2005). These methods are useful for small areas or regions with dense observation networks, but they have several issues when applied at the country or continental scale. For example, even though the influences of topographic effects, such as lapse rate, were well known, earlier interpolated gridded methods could not capture these known features. As a result, those methods cannot be applied to mountainous regions with few observations from sparse networks of weather stations. To account for the topographic effects, methodologies incorporating digital elevation models (DEM) were developed and applied at the continental

scale (Hutchinson, 1995; Thornton *et al.*, 1997). PRISM (Daly *et al.*, 2008) not only added elevation information, but also used topographic facet, coastal proximity, two-layer atmosphere, topographic position, and effective terrain height from DEM. Those methodologies primarily used the time-invariant information (i.e., elevation, or its derivatives). More recently, it has been recognized that adding time-variant information is effective to account for more spatial variability. Parmentier *et al.* (2015) have improved upon the interpolation schemes of maximum temperature by incorporating land surface temperature (LST) and Oyler *et al.* (2015a) used LST to create minimum and maximum temperature gridded data sets (TopoWx). However, none of these data sets incorporates time-variant spatial information for multiple climate variables other than temperature. Especially for running ecosystem models, it is imperative to have a several climate variables with the same spatial and temporal resolution. The difficulties in incorporating time-variant spatial information come from constraints including data storage, frequent missing-data, non-matching spatial resolution, insufficient availability, and lack of screening techniques that are universally applicable.

A machine learning technique was used in NEX-GDM to overcome these shortcomings to incorporate various time-variant data sets. Although machine learning techniques have been extensively applied in land cover classification problems (Pal, 2005; Belgiu and Drăguț, 2016), their use is not well developed for the purpose of downscaling climate data sets. PERSIANN (Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks) is one application of a machine learning technique, which employed an artificial neural network (ANN) algorithm to estimate precipitation from satellite data (Hsu *et al.*, 1997). Several experimental studies have applied the random forest algorithm (Breiman, 2001) to develop a two-dimensional climate surface (Shi and Song, 2015; Higuchi *et al.*, 2016; Jing *et al.*, 2016). However, no study has applied the random forest to dozens of input data sets to create consistent climate data sets across many different climate variables. NEX-GDM uses several time-variant spatially continuous data sets, including reanalysis, satellite, and radar data sets. Therefore, our concept of this study is to combine multiple large spatially continuous data sets at high spatial and temporal resolutions through machine learning techniques. This can be realized by leveraging the super-computing power and mass storage of NEX.

NEX-GDM provides long-term, daily 1-km gridded data for precipitation, maximum temperature, minimum temperature, dew point temperature, wind speed, and solar radiation from 1979 to 2017 for the conterminous United States. The objectives of this study are to (a) create gridded data set using multiple time-variant spatially continuous

data sets for ecosystem modelling, (b) propose a straight-forward methodology to apply machine learning algorithms to interpolate ground observation data into spatial maps at the continental scale, and (c) test whether NEX-GDM accurately captures the spatial variability of those climate variables.

2 | METHODS

2.1 | Random forest algorithm

Random forest is a supervised machine learning technique that uses decision trees. Breiman (2001) developed the random forest algorithm by combining bagging and random variable selection within a binary decision tree. The rationale of random forest is that aggregation of a few hundred weak predictor decision trees can make a strong predictor. Random forest can be applied to both classification and regression problems.

The random forest regression algorithm makes a few hundred decision trees from samples randomly selected from the training data through a bootstrapping procedure. In the process of node-splitting into decision trees, the random forest algorithm randomly selects a subset of variables for each node; in this way, the random forest procedure avoids overfitting. The random forest regression algorithm produces an average of the values in the end nodes of the decision trees. When random forest is applied to classification, the final estimation is the majority class of end nodes. Each tree classifier can be expressed as

$$\text{Tree}_k = f(x, \Theta_k), \quad (1)$$

where x is the input vector, Θ_k is a randomly selected vector from training data sets, and n is the number of trees. The estimation of random tree is

$$\text{RF}(x) = \frac{1}{n} \sum_{k=1}^n \text{Tree}_k(x). \quad (2)$$

As the derivative of the random forest classification, the probability of a certain class can be calculated as the percentage of the decision trees that estimate the class to the total number of decision trees.

An advantage of using the random forest algorithm is the ability to calculate the contribution of each input variable used in training process, which is defined as “variable importance” (Breiman, 2001). Samples not selected in the random selection process, an out-of-bag (OOB) sample, are used to calculate the variable importance. The OOB error is the error of the prediction of applying the OOB samples. The variable importance is the difference of the OOB errors with and without permuting the target variables in the OOB samples. The variable importance of each variable is relative to that of the other variables, and so the values are always subject to change after adding other variables.

Variable importance benefits the random forest for selecting input variables in NEX-GDM, and can be useful information for the addition of new input data sets in the future. Here, we present the variable importance as percentages to make the sum of the importance of all the variables 100%. Generation of the random forests was performed using the OpenCV library (Bradski, 2000) for random forest applications.

2.2 | Application of the random forest algorithm to two-dimensional data

We developed the Aggregation and Interpolation of NEX Archives (AINA) model to apply the machine learning method to two-dimensional data. AINA estimated the spatial climate variables in daily time steps. The ground observation data were used as the response variables, while the spatial data products were used as explanatory variables for the random forest algorithm.

Even though random forest is computationally inexpensive compared to other complex machine learning techniques, such as deep learning, training random forest for all the pixels still takes too much time for high-resolution applications. It is impractical to train the model for each individual pixel and so it is necessary to reduce the number of trainings. However, setting the zone of each trained model and applying it inside the zone inevitably creates artificial patchy patterns on the boundaries where different trainings have been applied. To avoid such artificial patterns, AINA applied the random forest algorithms through the following steps:

1. Using a 20-km grid, AINA searched for the two hundred (for precipitation, maximum temperature, and minimum temperature) and one hundred (for dew point temperature, wind, and solar radiation) weather stations that were contained within or closest to each 20-km grid cell (Figure 1a). The number of weather stations was determined through visually checking from the density of stations need for each variable to provide adequate coverage at regional scales, and to ensure that the stations selected could explain the spatial variability.
2. We bilinearly interpolated each of the input data sets (reanalysis, satellites, radar, and topography data) to the location of each of the weather stations. After this step, each weather station has one sample from each of the input data sets for each date, providing multiple explanatory variables from the spatial data inputs and one response variable or observed value for each date at each weather station (Figure 1b).
3. AINA trained the random forest creating 100 decision trees for each 20 × 20 km grid cell. The training data consist of 100 or 200 samples (depending on the

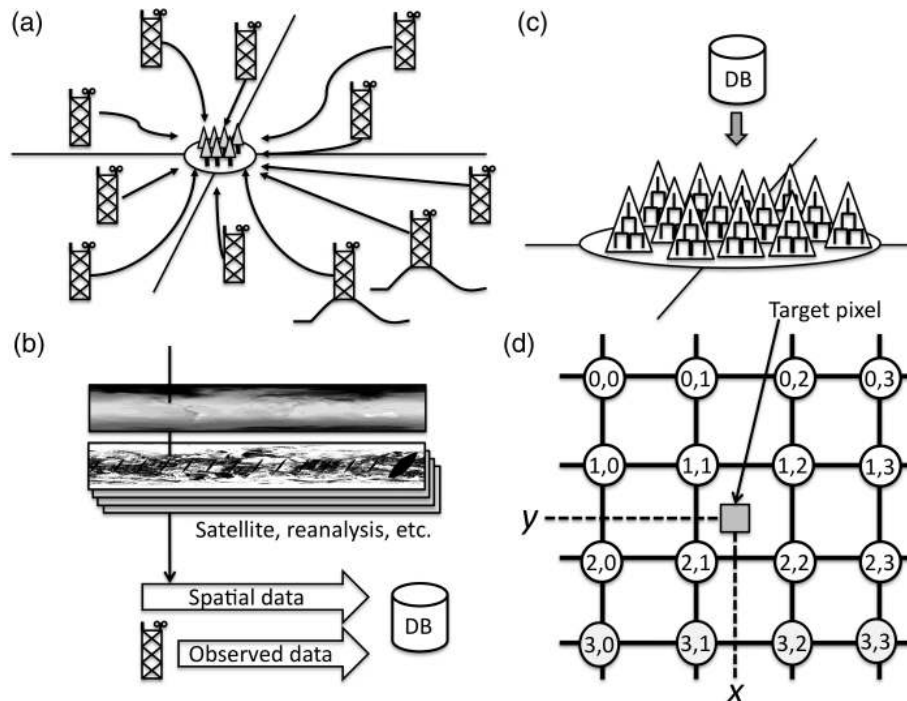


FIGURE 1 Conceptual diagram of NEX-GDM application of random forest algorithm to two-dimensional data. (a) Each random forest found the 200 closest weather stations. (b) The database was created by extracting single point data from the spatial data and observed data at the stations. (c) At each random forest point, we grew the trees using the database. (d) We estimated the value at the target pixel by weighted average of the surrounding 16 random forests. The numbers are notations of row and column of the x and y axis, respectively

variable) from the weather stations nearest to each grid cell (Figure 1c).

- The values of the target pixel (1 km resolution pixels in this study) were estimated from the 16 surrounding random forests. Then, the final estimates were calculated by weighing the 16 estimates (Figure 1d). The final estimate is given by

$$\text{Estimate} = \sum_{i=0}^3 \sum_{j=0}^3 w_{i,j} \text{RF}_{i,j}(x), \quad (3)$$

where i and j denotes i th row and j th column of the 16 surrounding random forests ($\text{RF}_{i,j}$) and $w_{i,j}$ is the weight for each random forest. We set the coefficients of bi-cubic linear-interpolation as $w_{i,j}$.

The random forest could have bias when using skewed training data (Zhang and Lu, 2011). Therefore, caution must be taken with applying random forest to spatial applications when the training data for a target pixel is skewed compared to the surrounding area. In AINA, using the nearest 10 weather station samples, the ratio-scale variables (i.e., precipitation, wind speed, and solar radiation) are adjusted using reduced major axis regression, while the interval scale variable (maximum temperature, minimum temperature, and dew point temperature) are adjusted from the difference of the mean.

The reduced major axis (RMA) regression is commonly used especially when X -axis is measured with error (Smith, 2009).

This new methodology allows us to create high resolution data without incurring an exponential cost for computation (i.e., $O(n^2)$), and apply AINA separately for each day and each variable. Some gridded climate data sets (e.g., TopoWx) calculate the daily value as summation of a well-described long-term mean value plus the interpolated deviation from surrounding weather stations. Meanwhile, other approaches (e.g., Daymet) directly interpolate the values from surrounding weather stations directly. The AINA approach is more closely related to daily-independent interpolation approaches such as Daymet.

2.3 | Specific treatment for each variable

For precipitation estimates, AINA applied the random forest algorithm twice. First, AINA estimated the probability of rainfall or snowfall events for each pixel. If the percentage of probability is greater than 50% (i.e., the majority of decision trees conclude that rainfall or snowfall occurred), we assumed there was rainfall or snowfall and then precipitation was estimated for the pixel again using random forest regression.

For dew point temperature, the available dew point temperature record is sparser than maximum temperature or minimum temperature records. If dew point temperature was

estimated independently from minimum temperature, unrealistic results can occur such as dew point temperatures that are far greater than minimum temperatures. To avoid such inconsistencies, AINA calculated the difference between dew point temperature and minimum temperature, and then interpolated the difference, following the same approach used by Daly *et al.* (2015). The final results for dew point temperature came from the summation of minimum temperature and the interpolated difference.

2.4 | Procedural incorporation of diverse spatial data

One of the objectives of this study is to incorporate multiple spatially continuous data sets, and thus AINA must be simple enough to allow for ingestion of diverse spatial data. In this regard, AINA has the following advantages compared to the other widely used interpolation schemes:

1. *Robust to erroneous data.* The spatial data, especially satellite data, tend to have erroneous data, which can include extremely high or low values. Estimated values in the random forest algorithm are limited in the maximum and minimum values of the observed data used for training the random forest. Also, the use of multiple independent spatially continuous data sets reduces the influence of an erroneous error in a single data set.
2. *Requires no physical assumptions.* AINA does not require any physical assumptions, while knowledge-based statistical models need specific data set as input of its physical assumption. This allows us AINA to incorporate many different kinds of data set without any explicit formula.
3. *AINA calculates the variable importance of each input variable.* The variable importance output can help decide which spatially continuous data sets should be included in the future. This feature differentiates AINA from other machine learning models that treat their input variables as a black box and do not have ability to evaluate the contribution of each spatially continuous data set input to the machine learning algorithm.

2.5 | Application to the conterminous United States

By applying AINA to the conterminous United States, we created NEX-GDM to provide daily precipitation, maximum temperature, minimum temperature, dew point temperature, wind speed, and solar radiation with a spatial resolution of 1-km from 1979 to 2017 using the Lambert azimuthal equal area projection.

2.6 | Parameters in AINA

AINA requires only three parameters, while other statistical methodologies require several parameters to be optimized. The random forest algorithm needs (a) the number of samples, (b) the decision-tree depth, and (c) the number of decision trees. These parameters are in turn constrained by the methodology. (a) The number of samples must be a few hundred because a few hundred stations can cover a few hundred kilometres around the random forest point, which represents the regional climate pattern to be explained by the random forest. Taking into account that available stations are different for each climate variable (Figure 2), precipitation, maximum temperature, and minimum temperature used the data from the surrounding 200 stations, while dew point temperature, wind speed, and solar radiation ingested the data from the 100 surrounding stations. (b) Accordingly, to get sufficient samples in the decision trees' end-points, the depth of the decision trees was set to 6 for precipitation, maximum temperature, and minimum temperature, and to 5 for dew point temperature, wind speed, and solar radiation. (c) Usually, the more trees the random forest has, the less over-fitting occurs. Empirically, 100 decision trees have been shown to be sufficient (Oshiro *et al.*, 2012), and so we set 100 decision trees for each random forest. Each pixel's value is estimated from the 16 surrounding random forests. As a result, each pixel was calculated from more than 1,600 trees though the random forest, even though the random forests are weighted. Thus, 100 decision trees are sufficient for each random forest.

2.7 | Comparison with PRISM and Daymet

To show the improvement of our output from the existing data sets, we compared the output with the existing interpolated weather data from Daymet and PRISM. Both are publicly available and are the most frequently used data sets for the study of daily climate over the conterminous United States (e.g., Mourtzinis *et al.*, 2017). The spatial patterns of annual and daily data for precipitation and minimum temperature of NEX-GDM were compared with those of Daymet and PRISM as described in section 4.1.

AmeriFlux is a network of sites measuring ecosystem flux and meteorological data in North, Central, and South America (Baldocchi *et al.*, 2001), and hourly and half-hourly data were used for the validation of the NEX-GDM data. AmeriFlux data are independent from all the interpolated data sets. Another reason AmeriFlux was used is that the observation sites are well distributed to cover most of the land cover types in the United States (Yang *et al.*, 2008). The RMSE, bias, and correlation coefficients between the AmeriFlux data and the pixel values are calculated for each year against all three data sets. Those statistics were averaged from 2001 to 2015, when the data from at least 10 flux towers are available.

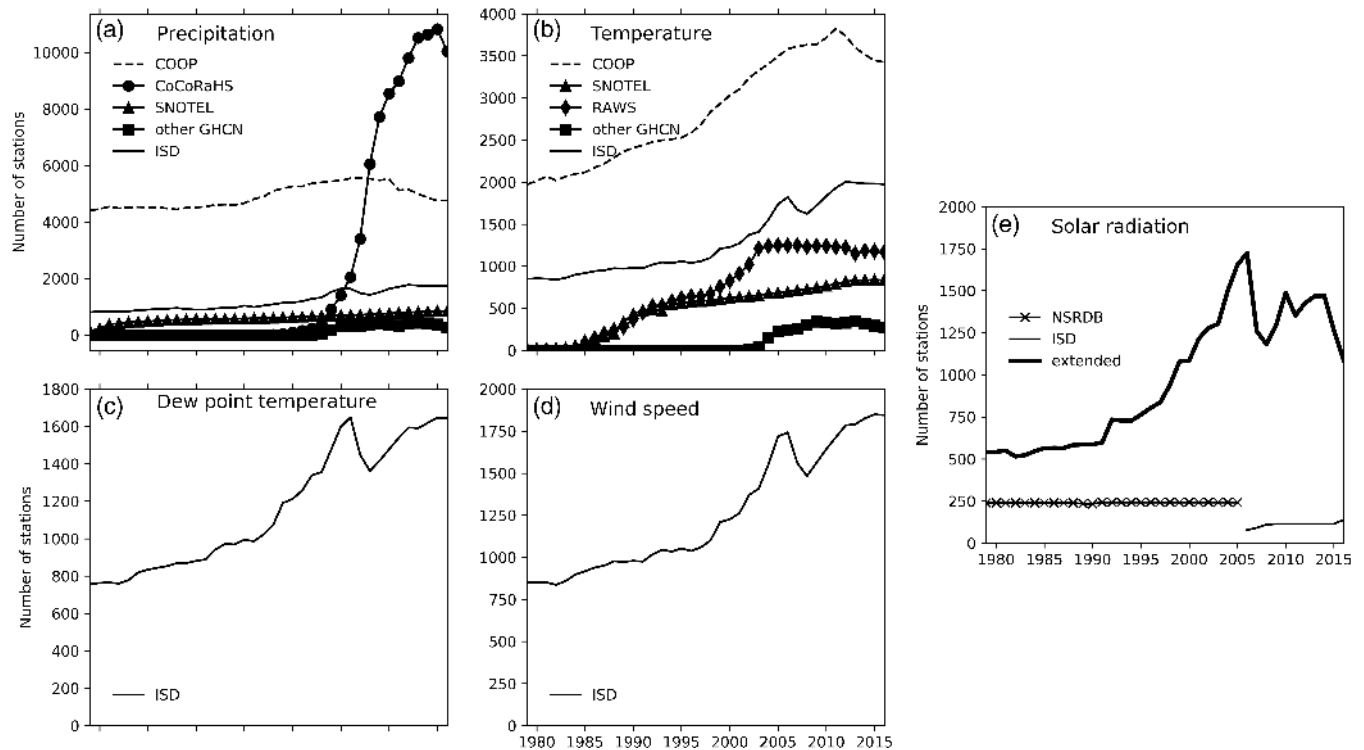


FIGURE 2 Number of ground observation stations of each source data set for (a) precipitation, (b) temperature, (c) dew point temperature, (d) wind speed, and (e) solar radiation. The thick line of “extended” in the plot (e) was derived from the National Solar Radiation Database (NSRDB) and ISD to increase the input

3 | DATA SETS

3.1 | Weather station data

We used the Global Historical Climatology Network-Daily database (GHCN-D; Menne *et al.*, 2012), Integrated Surface Database (ISD; Smith *et al.*, 2011), and National Solar Radiation Database (NSRDB; National Renewable Energy Laboratory, 1992; 2007) as the ground weather station data for training purposes.

3.1.1 | Global Historical Climatology Network-Daily database

GHCN-D is a collection of daily data sets of ground weather observation worldwide. We used GHCN-D for precipitation, maximum temperature, and minimum temperature in the United States. The majority of GHCN-D precipitation data in the United States originates with reports from the Cooperative Observer Program (COOP) network (NCDC, 1981) and the Community Collaborative Rain, Hail and Snow Network (CoCoRaHS; Reges *et al.*, 2016) (Figure 2a), while the majority of temperature data is from COOP (Figure 2b). The number of available stations has increased over time for all of the variables.

COOP relies on weather station records measured primarily by volunteers throughout the United States. Although observation quality is an important issue for interpolating weather station data, COOP records are well known to have varying quality (Davey and Pielke, 2005).

The poor quality of the data set can directly influence the surrounding pixels through spatial interpolation. To filter or correct the erroneous record, various efforts have been made for long-term climate analysis or spatial studies (Menne and Williams, 2009). However, the stricter the filtering algorithm, the greater the interpolation error caused by the resultant scarcity of observations. For example, only 6% of COOP stations passed the strict bias test for precipitation (Daly *et al.*, 2007), leaving little data with which to work. AINA is inherently robust to erroneous data as it is supported by the random forest algorithm, which can smooth out anomalous data by decision tree processes therefore we did not apply filtering process to COOP data. However, this smoothing feature of AINA also has the potential to affect the frequency and magnitude of extreme events in the final outputs.

The observation time in a day is also another important issue for daily data, because our algorithm requires matching the observation time for each weather station variable with the observation times for the various spatially continuous data sets. Unlike other governmental observation data, COOP has no rules regarding observation time, and the reporting time varies depending on stations, though the peak tends to be 0700 or 1700 LST. COOP provides the reporting time information in an hourly time step. To match the spatially continuous data sets with the COOP and CoCoRaHS observation times, NEX-GDM defined each day as 12Z–12Z UTC, which is 0400 LST. Pacific standard time (PST) and 0700 LST eastern standard time (EST). This definition is the

same definition used for PRISM (Daly *et al.*, 1994). Data with reporting times that were earlier than this range were moved to the previous day.

GHCN-D includes the sub-data sets created by the National Center for Environmental Information (NCEI) (formerly National Climatic Data Center, NCDC), but the daily data are not summarized from 12Z to 12Z UTC. The hourly data from the same NCEI stations are available as ISD. Therefore, we created daily data from the ISD data after first excluding all of the NCEI sub-data sets to avoid redundancy with the ISD data. Those excluded data sets from the GHCN-D data were the First-Order Summary of the Day, Automated Surface Observing System (ASOS) data, Global Surface Summary of the Day, and ISD sub-data set. For regions close to the US–Canada border, a sub-data set from Environment Canada was used even though the network is outside of the United States.

Occurrence of rainfall is necessary to estimate the probability of rainfall, and trace precipitation that has no numeric value assigned to the precipitation event must be set to be greater than zero. For this study, we used 0.05 mm/day as the value for trace precipitation, which is half of the minimum reporting value for precipitation, instead of setting trace precipitation to a null or zero value as in the original observational data (Note that we did not follow the conventional definition of the term, “wet day” or “rain day.” The user should set the threshold (i.e., 0.1 mm/day) in the analysis of the wet days to compare the NEX-GDM with other data sets.

3.1.2 | ISD version 2

ISD is a collection of weather reports from the entire world at an hourly time step (Smith *et al.*, 2011). We extracted the data only for the conterminous United States and converted them into a daily summary for each variable. The same procedure used with the Global Surface Summary of the Day (GSOD) data set was used to convert the ISD hourly data into daily data. To match with the definition of a day used in the GHCN-D, the daily data was summarized from hourly data between 12Z UTC for each day to 12Z UTC of the next calendar day.

The transition from manual observation to ASOS obstructs the analysis of long-term trends by creating inconsistency. Unlike other discontinuity issues that occur randomly (such as relocation or shifting measurement height), the ASOS transition happened mainly in a specific period of time and caused the same artificial directional trend. NEX-GDM is able to minimize random errors, given its low dependency on any single station, but even NEX-GDM cannot ignore such a change. Therefore, even for NEX-GDM it is necessary to correct the artificial trend. However, the day of transition is not described for all stations. Instead, only a partially completed record of ASOS transition dates is available (NCDC, 2002). Thus, before

the ISD data was used for input to NEX-GDM, the data discontinuity for maximum temperature, minimum temperature, dew point temperature, and wind speed was corrected as follows:

1. The transition from manual to ASOS was identified for the target station using visibility, in which ASOS reported the maximum of 10 miles (NOAA, 1998). The available ASOS transition information (NCDC, 2002) showed that the distribution of ASOS visibility has characteristics of a mean greater than 8 miles with skewness less than -1.5 .
2. We calculated daily time series of the mean of the surrounding stations having coordinates within $\pm 3^\circ$ of latitude and longitude from the target station.
3. Two regressions were calculated between the target station and the mean of the surrounding stations: for the year before the transition and for the year after the transition.
4. Using those two regressions, the time series of the target station before the transition was adjusted to the time series after the transition.

Figure 3 is an example of the correction, where the abrupt drop of wind speed in 1995 was corrected, and showed a similar trend with the mean of the surrounding stations after the correction was applied.

3.1.3 | National Solar Radiation Database

ISD data does not cover solar radiation data prior to 2005. We used the NSRDB as the source of solar radiation data from 1979 to 2005 (Figure 2e). The database includes both observed and modelled data, but we only used the observed data in our analysis.

In the conterminous United States, far fewer stations collect data for solar radiation compared to other climate variables. Therefore, we interpolated the observed solar radiation data of ISD or NSRDB to the other ISD stations that did not measure solar radiation using the same AINA procedure. We added the ISD-measured climate data as explanatory variables to spatially continuous data set: mean temperature, dew point temperature, sea level pressure, pressure, wind speed, maximum temperature, minimum temperature, precipitation, and diurnal temperature range. The number of selected stations for each grid cell was set to 50, and the depth of the random forest was 5. Once again, the outputs were used as training data for creating the spatial maps of solar radiation. As a result, the total number of available stations for the subsequent AINA process was 500 to 1,800 (Figure 2e).

3.2 | Spatially continuous data sets

For creating NEX-GDM, we used a total of 30 spatial variables from 13 different data sets (Table 1 and Appendix S1, Supporting Information). If the data set was hourly or sub-

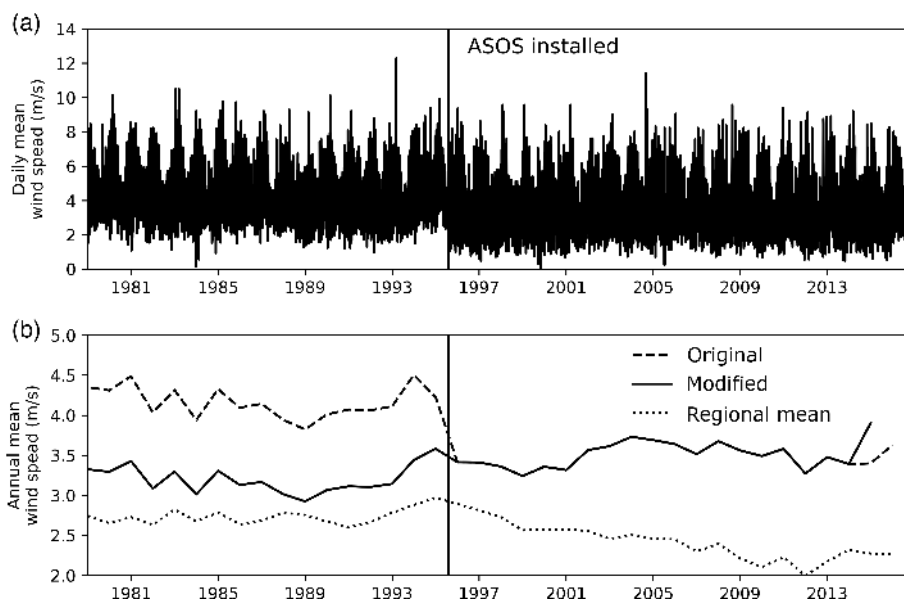


FIGURE 3 Example of wind speed at Atlanta International Airport (33.65°N; 84.42°W). (a) Time series of daily mean wind speed. (b) Annual mean wind speed after applying modification to the wind speed. The vertical line indicates the time of ASOS installation (August 1, 1995). The “regional mean” is the time series of the mean of surrounding stations, and was used for the modification process

TABLE 1 Summary of spatially continuous data sets used as input

Data name	Variables used for the input	Spatial resolution	Time step	Start year	End year	Reference
Reanalysis						
NCEP/NCAR reanalysis I	Temperature Precipitation Humidity u-Component wind speed v-Component wind speed Shortwave radiation	T62	6 hourly	1979	Present	Kalnay <i>et al.</i> (1996)
NCEP/DOE reanalysis II	Same as NCEP1	T62	6 hourly	1979	Present	Kanamitsu <i>et al.</i> (2002)
CFSR	Same as NCEP1	T382	6 hourly	1979	2009	Saha <i>et al.</i> (2010)
MERRA-2	Same as NCEP1	1/2 × 2/3 degree	Hourly	1980	Present	Gelaro <i>et al.</i> (2017)
NARR	Same as NCEP1	32.5 km	3 hourly	1979	Present	Mesinger <i>et al.</i> (2006)
Satellite						
GridSat-B1 v02r01	Infrared window water vapor	0.07 degree	3 hourly	1980	2015	Knapp <i>et al.</i> (2011)
TRMM 3B42 v7	Precipitation	0.25 degree	3 hourly	1998	2014	Huffman <i>et al.</i> (2007)
TERRA MODIS collection 6	Reflectance Brightness temperature Land surface temperature NDVI	0.05 degree	Daily Daily 8-day 16-day	2000	Present	Vermote and Vermeulen (2015a) Wan (1999) Huete <i>et al.</i> (1999)
AQUA MODIS collection 6	Same as TERRA MODIS	0.05 degree	Daily	2002	Present	
LTDR	Reflectance Brightness temperature	1 km	Daily Daily and 8-day	1982	2013	Pedely <i>et al.</i> (2007)
GIMMS 3G	NDVI	5 min	Half monthly	1982	2015	Pinzon and Tucker (2014)
Radar						
NCEP National Stage II analyses	Precipitation	4 km	Hourly	1996	Present	Lin and Mitchell (2005)
Static field						
GTOPO 30	Elevation Slope Aspect	30 second	-	-	-	Gesch and Larson (1996)
GSHHG	Distance from coast	-	-	-	-	Wessel and Smith (1996)

daily, the daily value was calculated from averaged data from 12Z UTC of each day to 12Z UTC of the next calendar day.

3.3 | Data sets used for validation

3.3.1 | PRISM and Daymet

We used PRISM (version D1, except for precipitation for which we used version D2) and Daymet (version 3) data for comparison purposes. PRISM has several versions for each climate variable. The version we used was the 4-km daily PRISM AN81d data set from 1981. Daymet provides 1-km data starting from 1980. After 2002, PRISM blended their original interpolation method with Stage IV only for precipitation data. PRISM defined the day as 12Z–12Z UTC, the same as NEX-GDM, while Daymet used the local calendar day as the reporting time from each station. However, the majority of the ground observation data is the same among those three data sets, and so it can be assumed that the day starts from the morning of the local time for all the data sets.

3.3.2 | Weather data from AmeriFlux and EARTH NETWORK

AmeriFlux is a US flux tower network whose main purpose is measuring the half-hourly flux of net ecosystem exchange and evapotranspiration for ecosystem research (Baldocchi *et al.*, 2001). Half-hourly weather data are also collected at flux towers. The ecosystem flux and weather data are measured over the canopy of the plants; therefore, the weather data of AmeriFlux must be biased relative to other weather station data, whose measuring height above the surface is 10 m for wind and 2 m for other weather variables. However, the data is still valuable for use in evaluating if the interpolation method can capture seasonal trends at flux tower sites. To make

daily data, times of half-hourly data were converted to UTC, and then the half-hourly data were summarized into daily data from 12Z to 12Z UTC. No gap filling was applied.

EARTH NETWORK data (courtesy of Earth Networks, Inc.) is a collection of weather observation data measured by private entities such as schools. The data set is completely independent from other governmental data sets. The density of weather observation points is much higher than the other weather station data sets used. We retrieved the maximum and minimum temperature data around New York City for 2011 for the purpose of testing whether NEX-GDM is able to capture spatial details on a small scale.

4 | RESULTS

4.1 | Comparison with PRISM and Daymet

4.1.1 | Comparison with long-term means

We compared the spatial patterns in the climatology of annual precipitation from NEX-GDM with PRISM and Daymet over the period from 1981–2010 for the purpose of describing the difference (Figure 4b,c for the western conterminous United States; Figure S1 for the entire conterminous United States). The difference of each pixel in the eastern conterminous United States overall was less than 10% for the comparisons with both PRISM and Daymet. Meanwhile, a difference of more than 30% occurred in the western United States, where orographic effects were dominant. To provide insights into the observed differences in the precipitation data sets, we also compared NEX-GDM with the USHCN version 2.5 data set, of which the GHCN-D were well filtered for erroneous data and adjusted for systematic

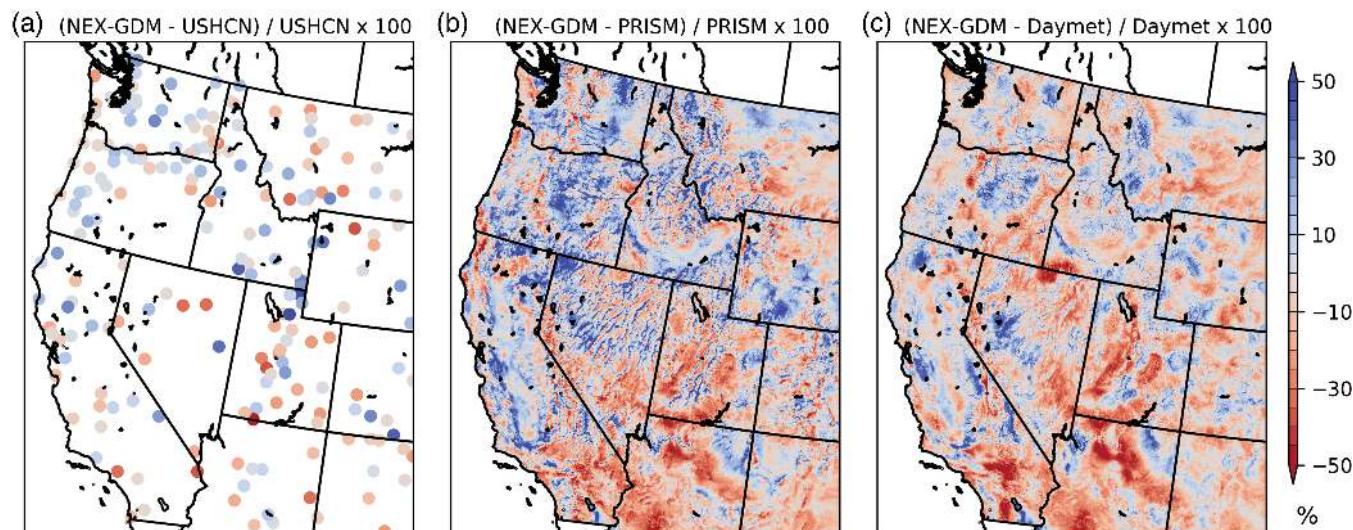


FIGURE 4 Difference of NEX-GDM in climatology (1981–2010) of annual precipitation against (a) USHCN, (b) PRISM, and (c) Daymet [Colour figure can be viewed at wileyonlinelibrary.com]

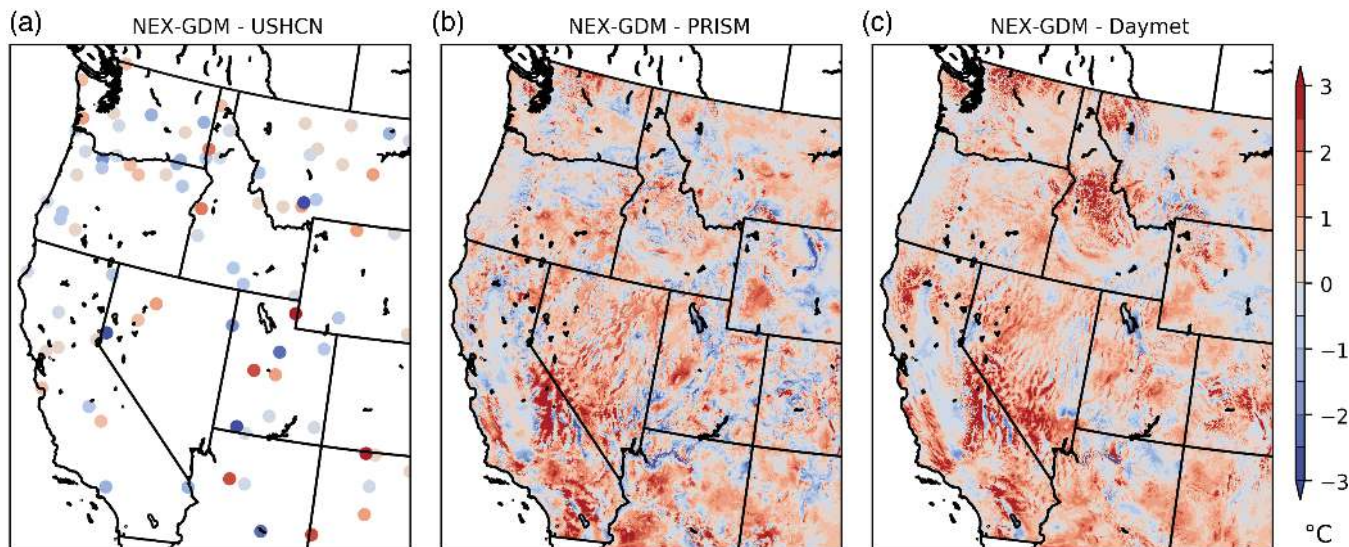


FIGURE 5 Difference of NEX-GDM in climatology (1981–2010) of annual mean of minimum temperature against (a) USHCN, (b) PRISM, and (c) Daymet [Colour figure can be viewed at wileyonlinelibrary.com]

bias (Menne *et al.*, 2009) (Figure 4a). The dry region in Arizona and Utah shows the overestimation of precipitation for the NEX-GDM data relative to PRISM and Daymet, especially in the higher elevation regions. NEX-GDM also underestimates over the Cascade Mountains in Washington, and over-estimates over the valley between the Coastal Range and the Cascade Mountains, relative to PRISM and Daymet. Difficulties in accounting for topo-climatic effect can be explained in part by the fact that AINA does not include an explicit formula to adjust or regress the precipitation data against the elevation.

The comparison of the minimum temperature climatology over the same period from 1981 to 2010 also shows the biggest differences in the southwestern United States (Figure 5 for the western conterminous United States; Figure S2 for the entire conterminous United States). Clearly, the NEX-GDM overestimated minimum temperature for high elevation regions compared to PRISM and Daymet. Again, this can be explained by lack of an explicit formula for lapse rate correction in AINA. In addition, the random forest algorithm cannot extrapolate beyond the maximum or the minimum values from the training samples, resulting in over-estimation of temperatures in regions that are higher in elevation than the highest weather station.

4.1.2 | Day-to-day comparison

To see the differences in greater detail, we compared daily precipitation from the three products on July 3, 1994, when Tropical Storm Alberto hit the southeastern United States (Figure 6). The magnitudes of precipitation are generally comparable among the three data sets. However, the spatial patterns of precipitation are quite different. PRISM precipitation reflects the heterogeneous spatial pattern from the station values. Most of these observations are likely accurate, but in the absence of an explanatory grid such as radar data (which

begins in 2002) that would place them into a more realistic spatial context, they appear in a bullseye pattern. The spatial pattern of Daymet precipitation is smoother and depicts distinct contour lines because Daymet incorporates only the distance from surrounding stations and elevation data with lapse rate in its estimation of precipitation. NEX-GDM precipitation does not show any clear dots or contour lines compared to the other two data sets. Thus, the impact of a single station on NEX-GDM is intermediate between PRISM and Daymet.

The same relationship can be found in the minimum temperature data. Figure 7 is the example on July 22, 2006 when a North American heat wave hit California. NEX-GDM and PRISM show more spatial variability in California's Central Valley than Daymet. Also, the NEX-GDM product shows a smaller lapse rate in the Sierra Nevada Mountains than PRISM and Daymet.

One of the main purposes of NEX-GDM is to provide daily surface climate data to run ecosystem models that require gridded data over conterminous United States. In support of this goal, we also calculated the error statistics for each month for the AmeriFlux sites. The RMSE, bias, and R^2 of NEX-GDM and PRISM are almost always comparable, and better than Daymet when compared to AmeriFlux data (Table 2). Using machine-learning techniques without any extra knowledge, NEX-GDM is able to capture spatial patterns comparable to knowledge-based interpolation schemes. Those statistics do not conclusively determine which data set is the best because each data set has a different purpose of use and the accuracy depends on the location (Behnke *et al.*, 2016). It is noteworthy that PRISM is a 4-km data set and we cannot directly compare the statistics with other data sets, even though it shows the highest accuracy. However, the statistics suggest that NEX-GDM can be useful as an alternative data set using completely different methods.

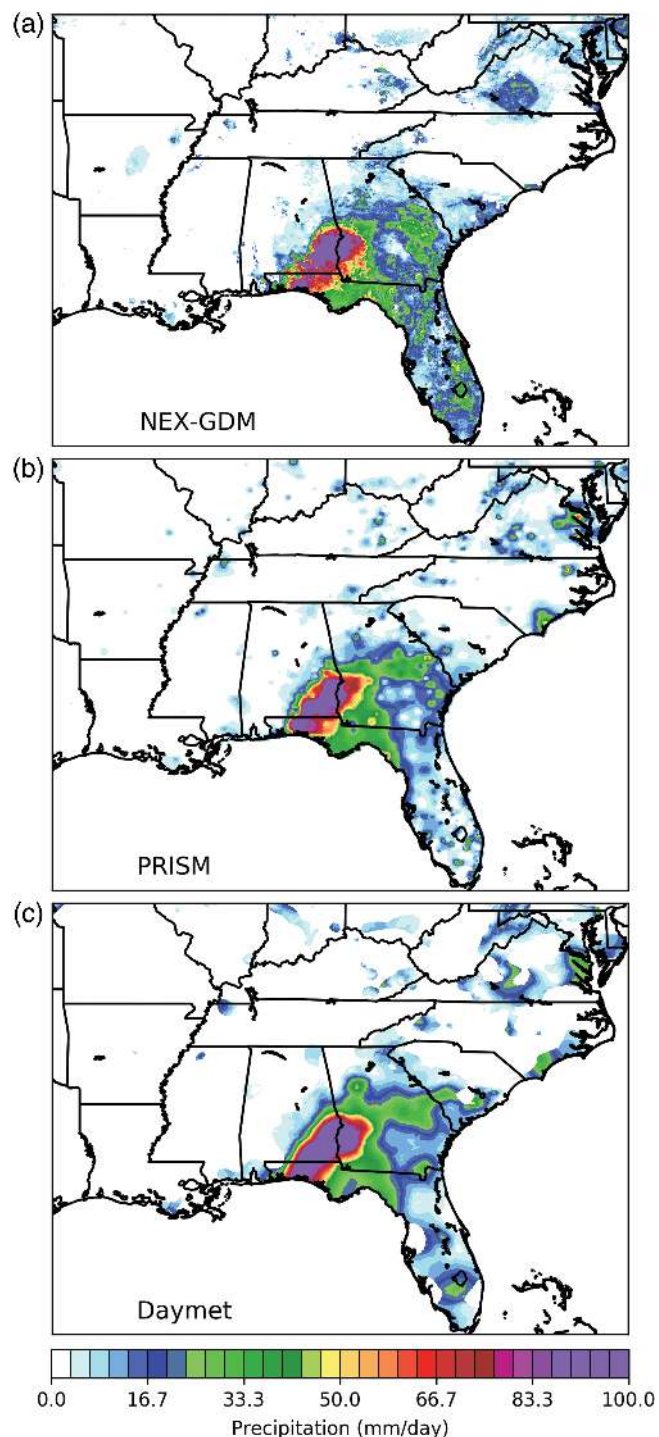


FIGURE 6 Daily precipitation on July 3, 1994, when tropical storm Alberto hit the southeast United States, for (a) NEXGDM, (b) PRISM, and (c) Daymet [Colour figure can be viewed at wileyonlinelibrary.com]

4.2 | Variable importance of spatially continuous data sets

The variable importance analysis shows the change in the relative contribution of spatially continuous data sets (Figures 8 and 9). When the number of spatially continuous data sets increased after 2000 with the addition of the nine MODIS data sets, the importance of each individual spatially continuous data set decreased overall. This indicates that each data set does contribute, and if missing data in a highly important variable exists,

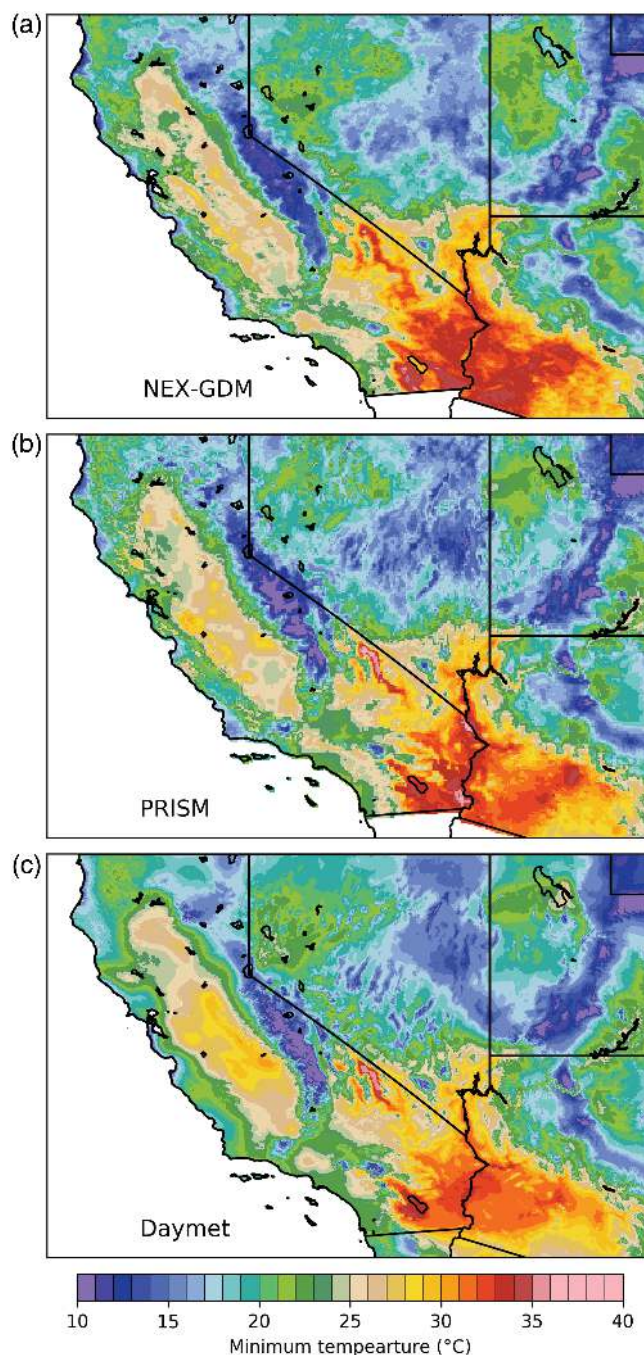


FIGURE 7 Daily minimum temperature on July 22, 2006, when a North American heat wave hit California, for (a) NEX-GDM, (b) PRISM, and (c) Daymet [Colour figure can be viewed at wileyonlinelibrary.com]

other data can be used to compensate for the gap. Furthermore, the substantial contribution from each variable facilitates a smooth transition in terms of accuracy when there is a data gap in one of the spatially continuous data sets. Since the variable importance is relatively evenly balanced and no single variable dominates in terms of overall importance, the addition or loss of one or two data sets out of more than 20 data sets has a small impact on the overall accuracy. Thus, NEX-GDM is less affected by data gaps due to changes in the availability of individual input data sets, so that production and update of the long-term data set is easy compared to other data sets that rely on a smaller or more limited number of input data sets.

TABLE 2 Summary of the statistics of NEX-GDM, PRISM, and Daymet against Fluxnet data from 1997 to 2012

Variable	Statistics	Data set		
		NEX-GDM	PRISM	Daymet
Precipitation	R^2	0.58	0.63	0.50
	RMSE (mm/day)	3.03	2.72	3.81
	Bias (mm/day)	0.25	0.30	0.64
Maximum temperature	R^2	0.94	0.94	0.74
	RMSE (°C)	1.44	1.35	2.51
	Bias (°C)	0.60	0.60	0.57
Minimum temperature	R^2	0.88	0.88	0.77
	RMSE (°C)	1.96	1.92	2.59
	Bias (°C)	-0.74	-0.68	-0.71
Dew point temperature	R^2	0.84	0.96	NA
	RMSE (°C)	2.27	1.39	NA
	Bias (°C)	-0.07	-0.43	NA
Shortwave radiation	R^2	0.76	NA	0.32
	RMSE (MJ/day)	1.63	NA	2.59
	Bias (MJ/day)	-0.31	NA	-0.21

In terms of the variable importance of precipitation input data sets, MERRA data was ranked the most important, while other reanalysis data, elevation data from GTOPO, and GridSat also made important contributions (Figure 8a).

However, after 2001, the importance of ground radar data or Stage II became significantly higher than other variables, because it can capture precipitation directly from the ground with much higher resolution compared to other data sets. This variable-importance information allows us to conclude that radar data are necessary to further improve the precipitation estimates in NEX-GDM. The high importance of GridSat is due to its frequent measurement of cloud cover compared to the other satellites capturing cloud cover. The elevation data from GTOPO also was important in explaining orographic precipitation.

The reanalysis data and elevation data from GTOPO highly contributed to both maximum and minimum temperature (Figure 8b,c). In 2005–2009 MODIS LST contributed to minimum temperature estimation, however, MODIS LST was less important for maximum temperature. Similar phenomena was observed by Oyler *et al.* (2016), along with the reason that was explained by the land surface process which strongly controlled the LST under clear sky condition (Oyler *et al.*, 2016). Elevation made an important contribution as a substitute for the lapse rate, as PRISM and Daymet explicitly incorporate the lapse rate.

In the calculation of dew point temperature (Figure 9a), wind speed (Figure 9b), and solar radiation (Figure 9c), the reanalysis data had higher importance than the other

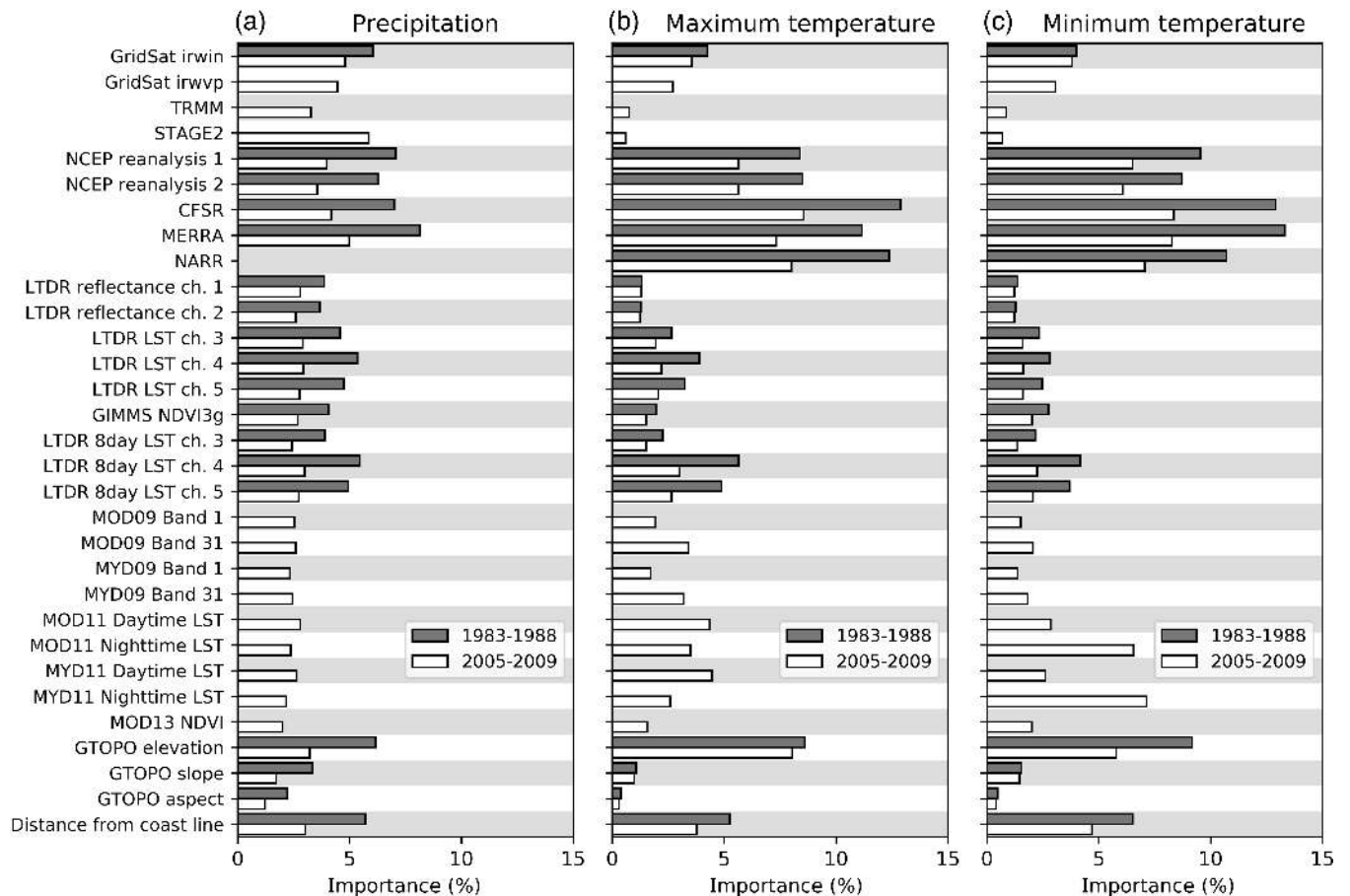


FIGURE 8 Bar chart of annual mean importance of spatial data for (a) precipitation, (b) maximum temperature, and (c) minimum temperature. The black bars are mean from 1983 to 1998. The white bars are mean from 2005 to 2009. The variable that has 0 value indicates missing for the period

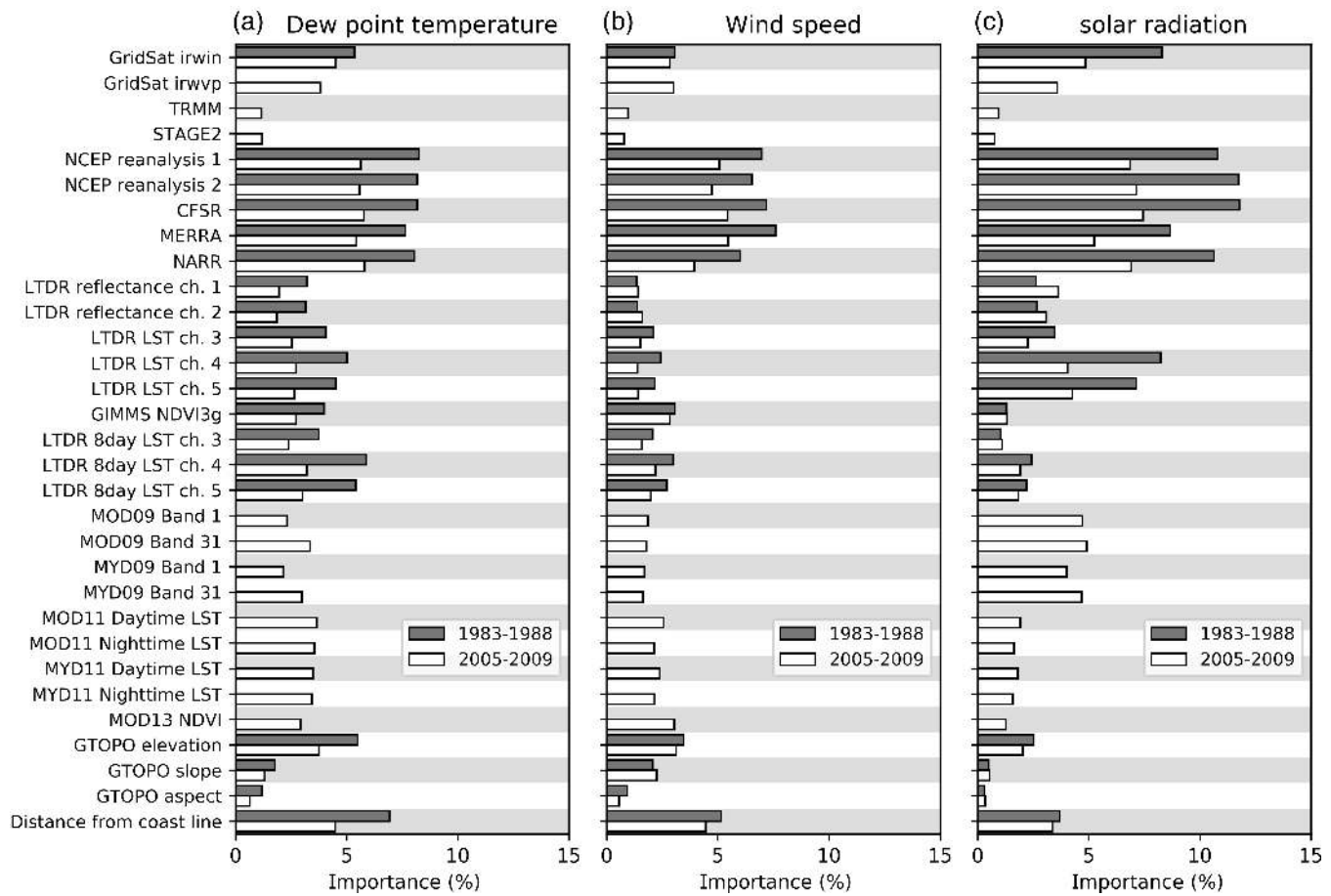


FIGURE 9 Same as Figure 8 except for (a) dew point temperature, (b) wind speed, and (c) solar radiation

variables, because the satellites and radars cannot directly measure those variables. Therefore, the importance of the reanalysis data contributed more than the other data sets (Figures 8 and 9). The spatial distribution of the variable importance does provide additional insights into where each input variable becomes important (Figures 10–12).

Stage II data was the most important input for precipitation in 2009, but it contributed mainly in the eastern United States, except for the Appalachian Mountains region (Figure 10a). It is well known that radar does not work as well in mountainous terrain as in flat areas because of overshooting, ground clutter, and less gauge measurement (Fulton *et al.*, 1998). As a result, the importance value of Stage II in the Rocky Mountains was lower than in the eastern United States. In contrast, the elevation of GTOPO has high importance values in the Rocky Mountains where it explains orographic precipitation (Figure 10h). Thus, relying only on Stage II will not suffice for the interpolation schemes.

The reanalysis data contributed primarily to wet regions due to the difficulty in predicting convective rain in dry regions. Meanwhile, satellite data, such as Gridsat and TRMM data, contributed more homogeneously and broadly over conterminous US than the reanalysis data.

Both maximum and minimum temperature show that elevation data (i.e., GTOPO) contributed well in the western

United States and Appalachian Mountains (Figures 11 and 12). However, the contribution of elevation to minimum temperature was much smaller than that of maximum temperature due to the MODIS LST contribution. MODIS night-time LST was important in the southwest United States for the minimum temperature calculation (Figures 12d,f). Reanalysis data was the most important input for estimation of both maximum and minimum temperature across the eastern United States.

4.3 | Contribution of satellite data to capture the urban heat island

Satellite data, especially LST, is essential for capturing spatial patterns at the continental scale (Figures 11 and 12). Other than LST, we demonstrate that NDVI can contribute to estimation of urban heat island effects in New York (Figure 13) by comparing the temperature estimates with satellite data and without satellite data. New York City experienced the urban heat island on July 2011 (Meir *et al.*, 2013). As an independent source of data to validate the spatial distribution of urban heat island effects, we used the EARTH NETWORK data, which is a private network of weather observations mainly located in urban areas (data courtesy of Earth Networks, Inc.). The quality of the EARTH NETWORK data was not homogeneous compared

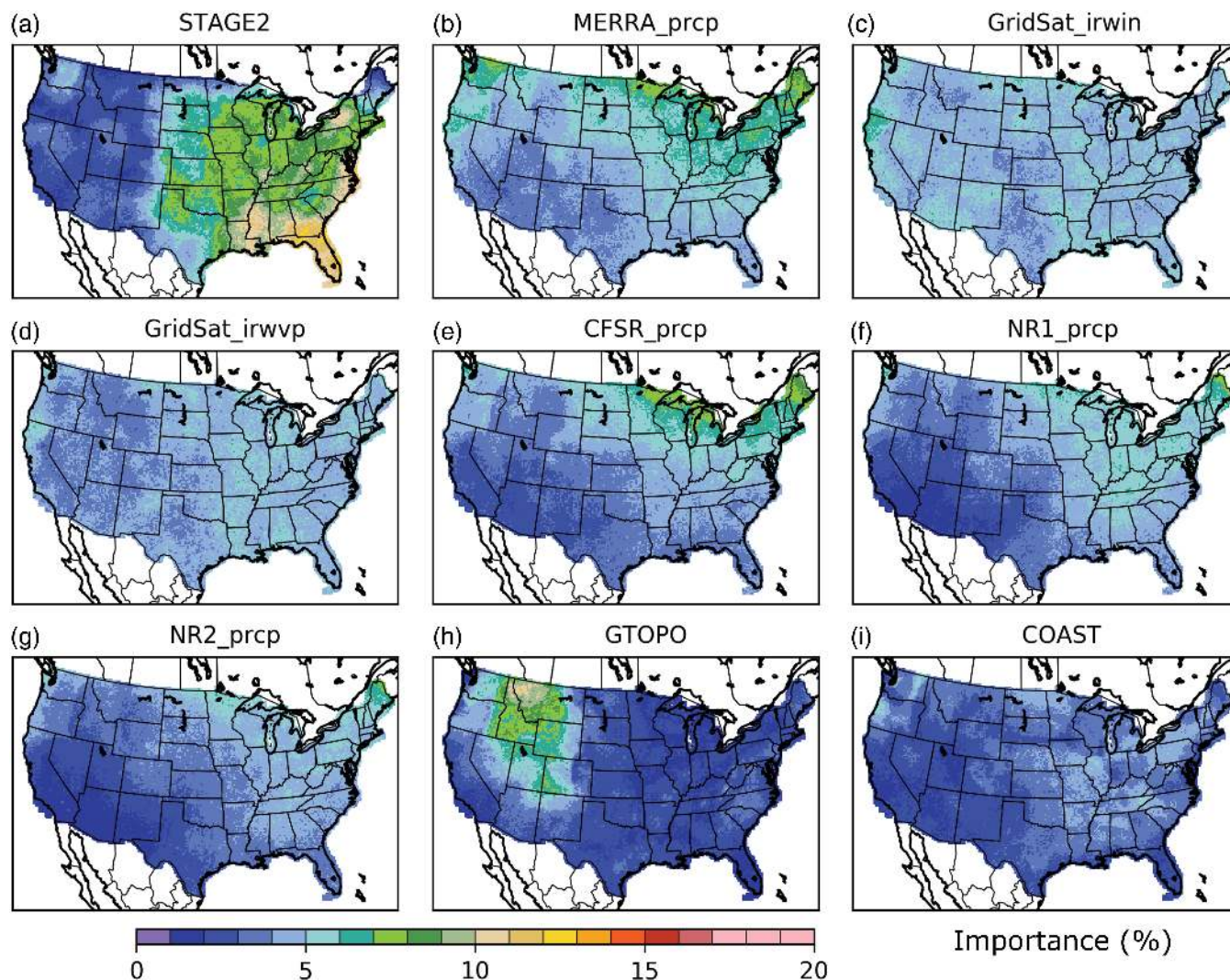


FIGURE 10 Spatial patterns of annual mean importance of input data for precipitation in 2009. The nine input data sets of highest average importance are shown in descending order from (a) to (i) [Colour figure can be viewed at wileyonlinelibrary.com]

to the governmental weather station data, and when we plotted the maximum temperature, the spatial distribution of temperature data was such that the urban heat island effect was hardly apparent across New York City. To avoid the issue of heterogeneous data quality, we calculated diurnal temperature range (DTR), which is the difference between maximum temperature and minimum temperature, to cancel the bias of maximum temperature and minimum temperature for each station. After making this correction, it is possible to recognize the urban heat island effects from the DTR of the EARTH NETWORK data (Figure 13c). The diurnal temperature range shows smaller values in the urban area (Figure 13d). NEX-GDM, using satellite data, clearly shows a similar spatial pattern with the EARTH NETWORK stations (Figure 13a). Meanwhile, NEX-GDM without satellite data shows smaller diurnal temperature ranges only on the coastlines (Figure 13b). Staten Island was surrounded by low values, and the centre of Brooklyn also shows higher values than the coastlines. The importance of the spatially continuous data sets in this analysis

shows that the GIMMS NDVI and MODIS NDVI added information about urban extent, and allowed NEX-GDM to capture the urban heat island effects in New York City (Figure 14).

5 | DISCUSSION

5.1 | Limitations in analyses using NEX-GDM

For long-term analyses using NEX-GDM, extensive care to address inhomogeneity is required. Gridded data sets rely heavily on a sparse network of observational input data sets leading to spurious results if time series corrections are not applied (McGuire *et al.*, 2012; Oyler *et al.*, 2015b). The causes of artificial long-term trends can include changes of observation time, sensor degradation, shift in measurement locations, and innovations or changes to the measurement instrument itself. To avoid those artificial long-term trends, some previous data sets applied inhomogeneity techniques (Vose *et al.*, 2014; Oyler *et al.*,

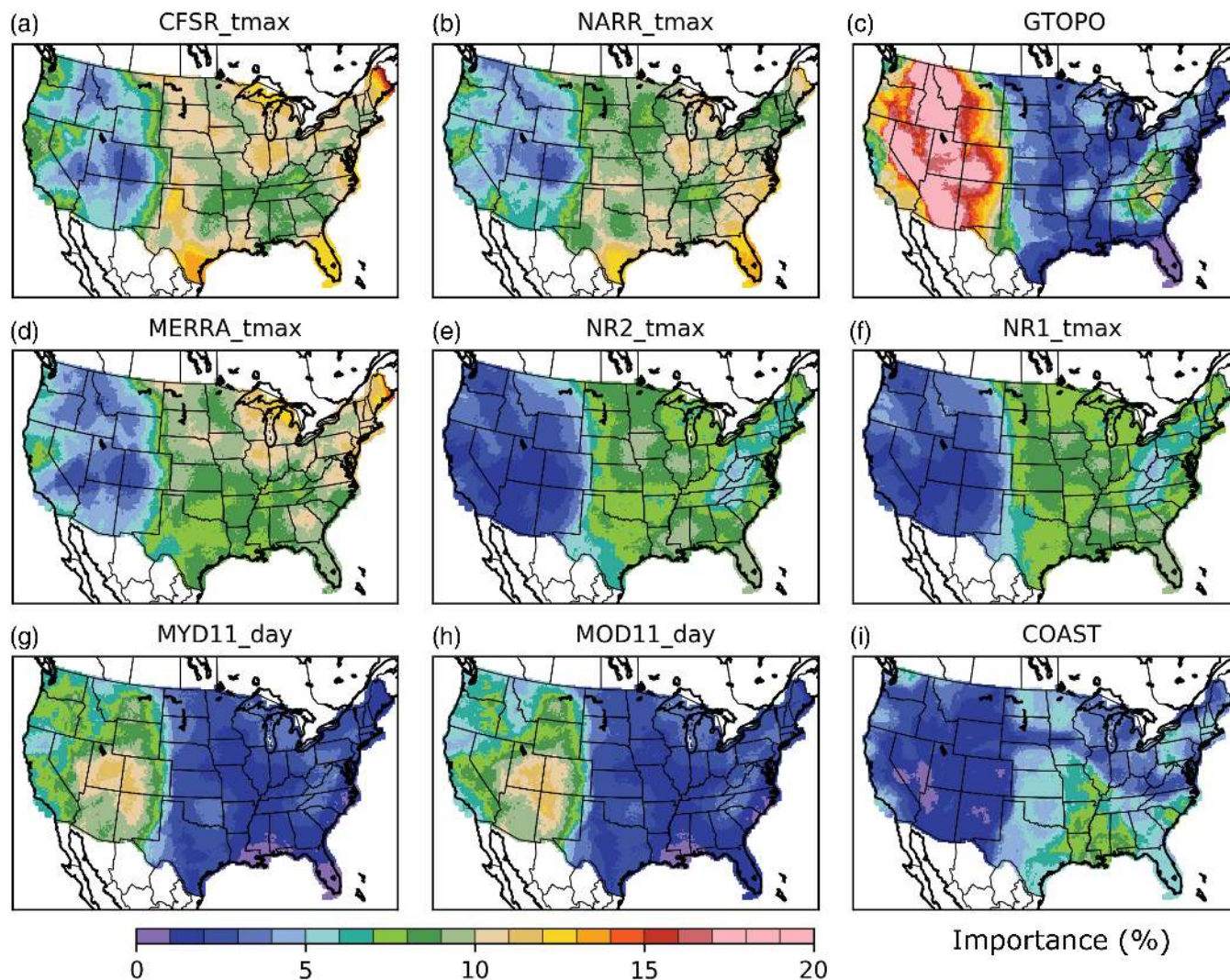


FIGURE 11 Same as Figure 10 except for maximum temperature [Colour figure can be viewed at wileyonlinelibrary.com]

2015a), while others added anomalies from climate models to long-term means created from interpolated data (Abatzoglou, 2013; Livneh *et al.*, 2013). NEX-GDM applied only simple bias correction to GHCN-D. Also, discontinuities in spatially continuous data set can create artificial long-term trends. Although the smoothing process of gridding can mitigate the artificial long-term trends to some extent, the users must be cautious in analysing the long-term trend by comparing it with adjusted weather station data. Further data set comparison analyses are in Appendix S2.

The AINA algorithm is still vulnerable to scarcity of station data, especially for solar radiation data. As with other gridded data, the error in spatial variability of the NEX-GDM data set is larger in earlier decades than in recent decades. It also can create artificial long-term trends. Therefore, long-term analysis of NEX-GDM of variables of sparse observation such as solar radiation and wind speed requires caution.

Another drawback of NEX-GDM is additional uncertainty in estimated values at high elevations and around the

top of the mountains due to the use of the random forest approach. Theoretically, the random forest algorithm cannot extrapolate its estimates beyond the maximum and minimum values from the input data. Thus, estimates for elevations that are higher than the elevation of the highest weather station in the training data set are likely to have higher uncertainty and lower reliability. For the same reason, isolated stations at high elevations can have a large effect on the surrounding mountainous regions. There are few high-elevation stations in eastern conterminous United States, and the isolated observation data at these stations (such as Mount Washington, NH) can have too much influence on the surrounding area, which can generate anomalous values relative to lower elevation regions.

For the same reasons with difficulty in analysing high elevation region, NEX-GDM may not be suitable for analysis of climate extremes. When only a single weather station in a region captures the extreme phenomena, the extreme value can be smoothed or even omitted because the random selection of samples for algorithm training could leave out the weather station that observed any particular extreme

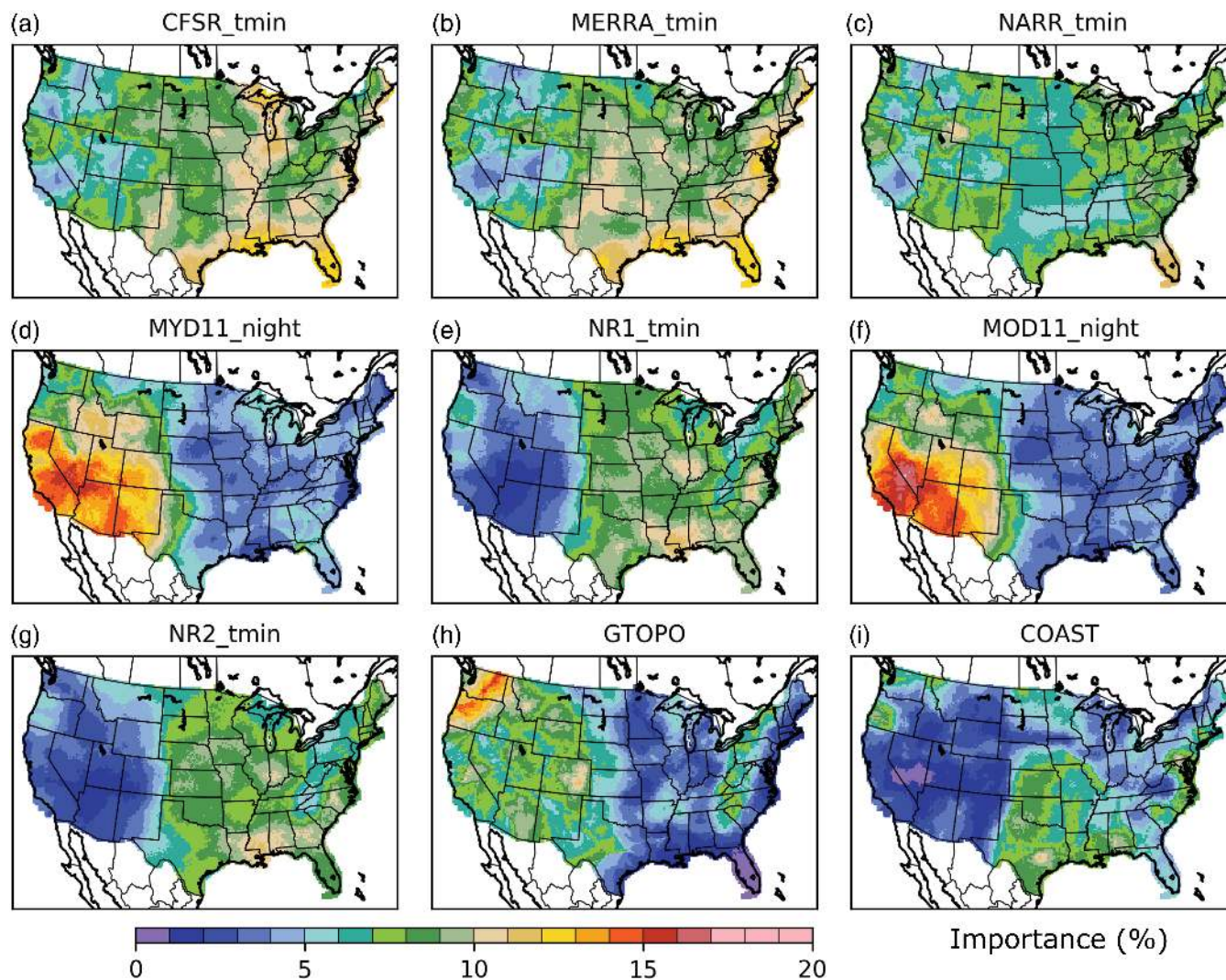


FIGURE 12 Same as Figure 10 except for minimum temperature [Colour figure can be viewed at wileyonlinelibrary.com]

event or events. As a general rule, NEX-GDM is not recommended for the studies that seek to analyse regions that are smaller in total than mean distance between weather stations.

5.2 | Scalability of AINA

Availability of ground observation data sets is critical to the ability of the AINA approach to produce accurate spatial data sets, because the ground observation is the response variable used by the random forest algorithm. As long as a sufficient number of ground observations are available, AINA can produce a spatial map of any type of climate variable. Here, we produced precipitation, minimum, maximum and dew point temperatures, wind speed, and radiation data sets for NEX-GDM. In future work, we plan to produce snow cover data sets from ISD. Application of AINA to other regions is also straightforward, especially for Europe and Far East Asia, where the density of climate observation stations is comparable to the United

States. Theoretically, AINA can be applied to regions where the network of ground observation stations is sparse. However, if the density of the observation stations is too sparse, the decision trees of the random forest cannot adequately explain the spatial variability within the climate data, and it is not recommended that the AINA be applied for such a region.

When applying AINA before 1979, the paucity of spatial data directly impacts the confidence of the gridded estimates. Neither radar nor satellite data are available before 1979, and thus NEX-GDM must rely on only reanalysis data sets. Unfortunately, the available reanalysis data, such as NCEP1, have a coarse resolution (>100 km), which is not sufficient for capturing spatial heterogeneity within the climate variables included in the NEX-GDM data set. To avoid issues caused by the coarse resolution, we plan to run regional climate models to create high-resolution spatial data sets to be used as climate inputs to ecosystem and hydrological models in future research.

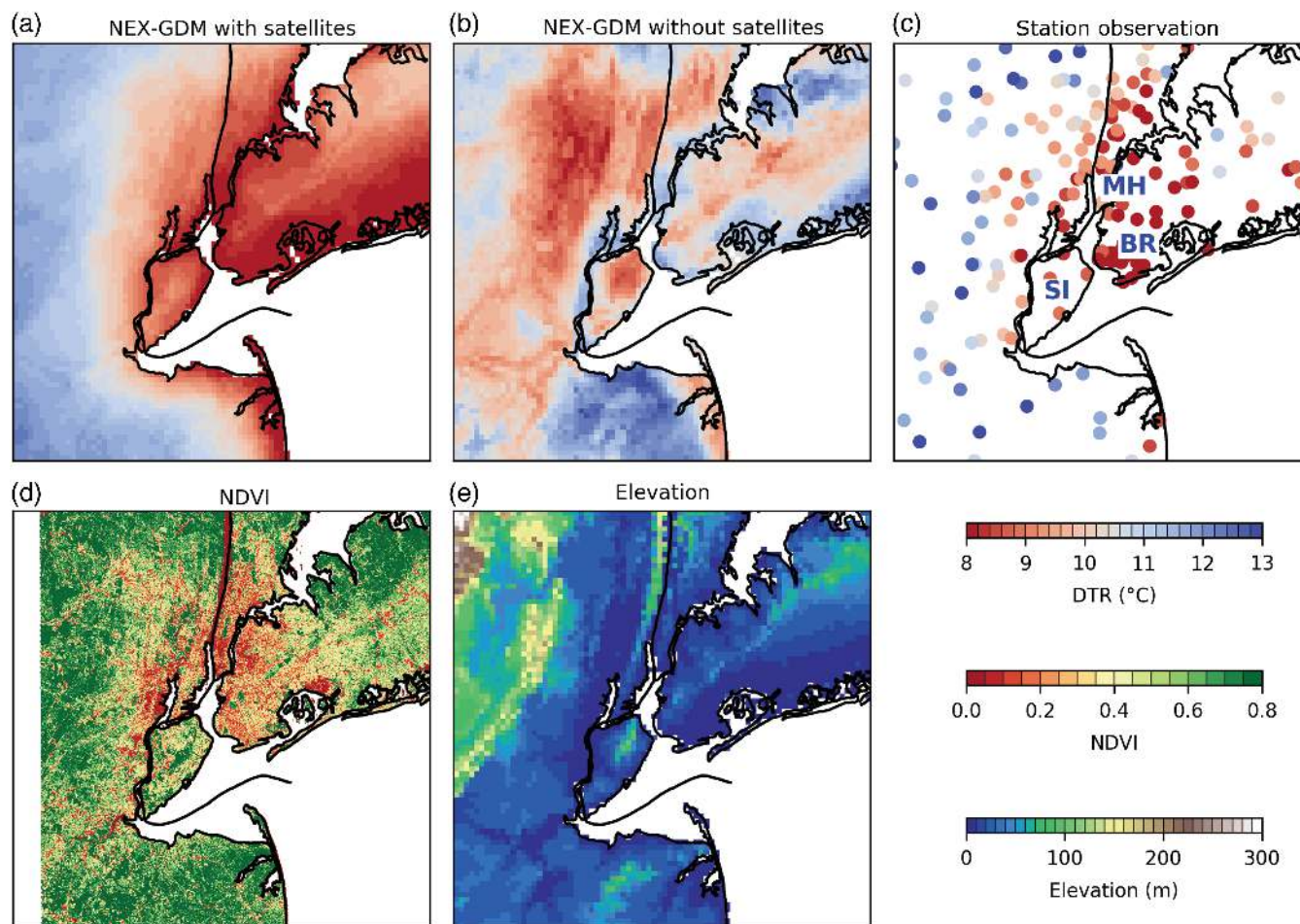


FIGURE 13 Monthly average of diurnal temperature range in the New York City on July 2011 of (a) NEX-GDM with satellites, (b) NEX-GDM without satellites, and (c) station observation of the EARTH NETWORK. Panel (d) is the NDVI of Landsat 5 on July 7, 2011. Panel (e) shows topography from GTOPO 30. The blue texts in panel (c) show the location of Brooklyn and Staten Island [Colour figure can be viewed at wileyonlinelibrary.com]

One additional caveat is that the solar radiation data used assumes that there is no local horizontal obstruction, such as surrounding mountains or building at the scale of each individual pixel. Therefore, when applying AINA to spatial scales that are finer than the input data of geographic data sets, collection of data on local obstructions may be needed.

5.3 | Near real-time data set

In addition to the long-term historical data sets included in NEX-GDM and produced using the full suite of available data sets, we will also produce near a real-time data set for applications that require continuous updates and recent information for the analysis, for instance, natural disaster preparedness and nature resource management. Near real-time processing using NEX-GDM requires several data sets for input, though some data sets used for the full data set are not produced in real-time. This inherently presents a compromise between the timeliness of near-real time updates and data accuracy. The variable-importance measure provided by the AINA approach will guide the selection of the minimum required input data sets to produce data products

with a satisfactory level of accuracy for public usage. For example, Figure 10 reveals that Stage II data is important in estimating the spatial distribution of precipitation in the eastern conterminous United States, and so we will produce the near real-time data set following the availability of radar data. Although the near real-time data set could be less accurate than the data set using all the available data with a longer delay, the near real-time data sets can help users who need to analyse spatial patterns of surface climate variables in real time, for example to analyse a current natural disaster, schedule irrigation, or forecast floods.

6 | CONCLUSION

We created the NEX-GDM data set by applying the AINA framework to the conterminous United States to produce high-resolution climate fields from 1979 to 2017. Our interpolation methodology is novel compared to the commonly used existing data sets because the machine-learning algorithm incorporated a wide range of different types of spatially continuous data sets. Statistics from comparisons with

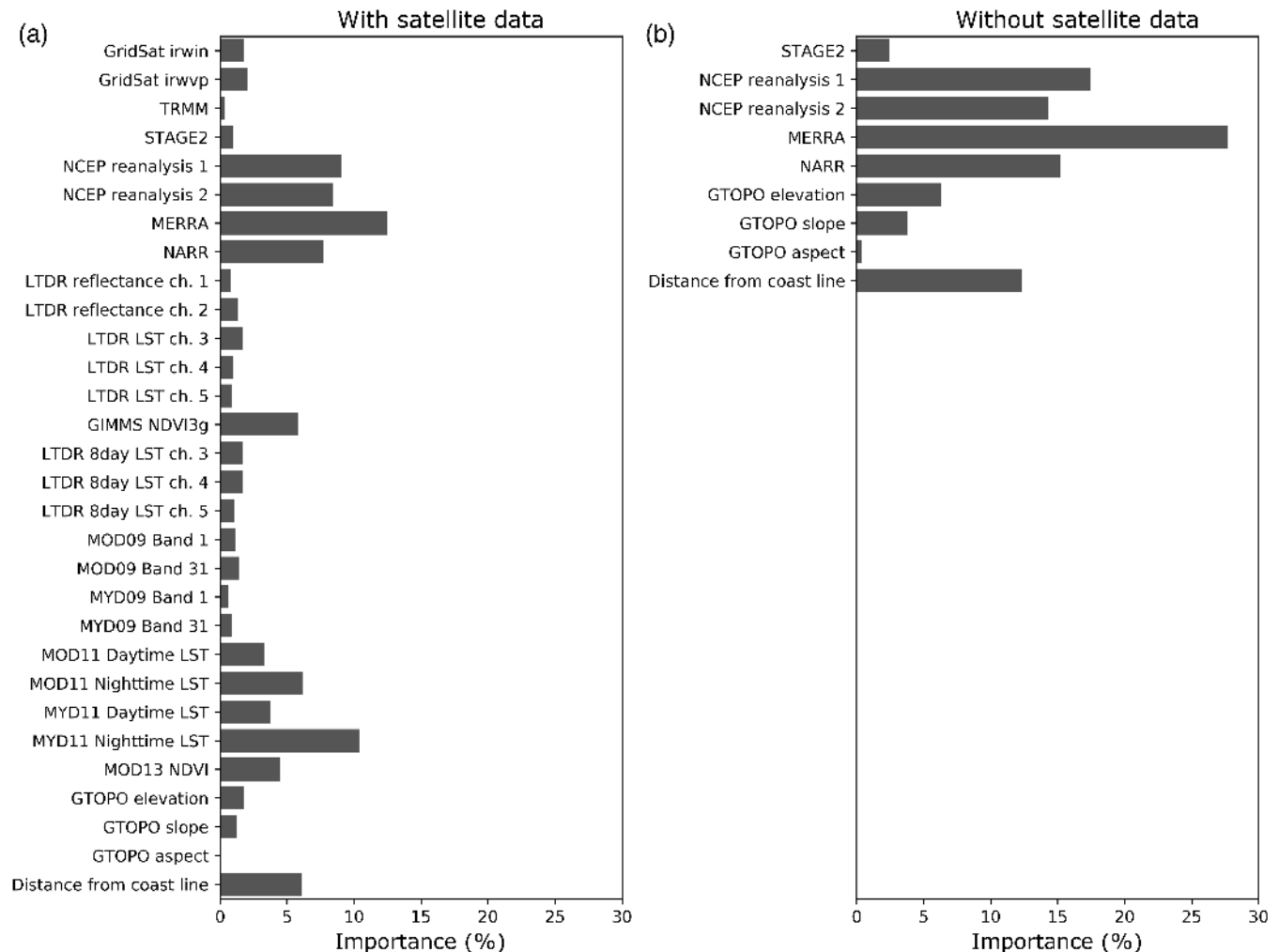


FIGURE 14 Monthly mean variable importance of input variables for calculating minimum temperature in New York City on July 2011 (a) with satellites and (b) without satellites

existing data sets produced with completely different algorithms shows that the accuracy of NEX-GDM is comparable with those data, and useful for ecosystem modelling. Also, we quantified, through the variable importance analysis, the importance of the contribution of satellite data, which are not used by other gridded climate data products, to the NEX-GDM. In the future, we will produce other climate variables taking advantage of the scalability of the AINA framework. The NEX-GDM data set is publicly available on the NEX facility.

ACKNOWLEDGEMENTS

This research was supported by funding from NASA's Earth science program. Computational resources from the NASA Earth Exchange helped facilitate the analysis and storage of the model products.

ORCID

Hirofumi Hashimoto  <https://orcid.org/0000-0001-7706-1854>

REFERENCES

- Abatzoglou, J.T. (2013) Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1), 121–131. <https://doi.org/10.1002/joc.3413>.
- Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw, K.T., Pilegaard, K., Schmid, H.P., Valentini, R., Verma, S., Vesala, T., Wilson, K., Wofsy, S., Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw, K.T., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S., Vesala, T., Wilson, K. and Wofsy, S. (2001) FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, 82(11), 2415–2434. [https://doi.org/10.1175/1520-0477\(2001\)082<2415:FANTTS>2.3.CO;2](https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2).
- Behnke, R., Vavrus, S., Allstadt, A., Albright, T., Thogmartin, W.E. and Radeloff, V.C. (2016) Evaluation of downscaled, gridded climate data for the conterminous United States. *Ecological Applications*, 26(5), 1338–1351. <https://doi.org/10.1002/15-1061>.
- Belgiu, M. and Drăguț, L. (2016) Random forest in remote sensing: a review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>.
- Bradski, G. (2000) The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 120, 122–125.

- Breiman, L. (2001) Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Daly, C., Neilson, R.P. and Phillips, D.L. (1994) A statistical–topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology*, 33(2), 140–158. [https://doi.org/10.1175/1520-0450\(1994\)033<0140:ASTMFM>2.0.CO;2](https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2).
- Daly, C., Gibson, W.P., Taylor, G.H., Doggett, M.K. and Smith, J.I. (2007) Observer bias in daily precipitation measurements at United States cooperative network stations. *Bulletin of the American Meteorological Society*, 88(6), 899–912. <https://doi.org/10.1175/BAMS-88-6-899>.
- Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J. and Pasteris, P.P. (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, 28(15), 2031–2064. <https://doi.org/10.1002/joc.1688>.
- Daly, C., Smith, J.I. and Olson, K.V. (2015) Mapping atmospheric moisture climatologies across the conterminous United States. *PLoS ONE*, 10(10), e0141140. <https://doi.org/10.1371/journal.pone.0141140>.
- Davey, C.A. and Pielke, R.A. (2005) Microclimate exposures of surface-based weather stations: implications for the assessment of long-term temperature trends. *Bulletin of the American Meteorological Society*, 86(4), 497–504. <https://doi.org/10.1175/BAMS-86-4-497>.
- Fulton, R.A., Breidenbach, J.P., Seo, D.-J., Miller, D.A. and O’Bannon, T. (1998) The WSR-88D rainfall algorithm. *Weather and Forecasting*, 13(2), 377–395. [https://doi.org/10.1175/1520-0434\(1998\)013<0377:TWRA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1998)013<0377:TWRA>2.0.CO;2).
- Gelaro, R., McCarty, W., Suárez, M.J., Todling, R., Molod, A., Takacs, L., Randles, C.A., Darmenov, A., Bosilovich, M.G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A.M., Gu, W., Kim, G.K., Koster, R., Lucchesi, R., Mervola, D., Nielsen, J.E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S.D., Sienkiewicz, M., Zhao, B., Gelaro, R., McCarty, W., Suárez, M.J., Todling, R., Molod, A., Takacs, L., Randles, C.A., Darmenov, A., Bosilovich, M.G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., Silva, A.M. da., Gu, W., Kim, G.K., Koster, R., Lucchesi, R., Mervola, D., Nielsen, J.E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S.D., Sienkiewicz, M. and Zhao, B. (2017) The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, 30(14), 5419–5454. <https://doi.org/10.1175/JCLI-D-16-0758.1>.
- Gesch, D.B. and Larson, K.S. (1996) Techniques for development of global 1-kilometer digital elevation models. *Pecora Thirteen, Human Interactions with the Environment-Perspectives from Space*. Sioux Falls, SD.
- Higuchi, A., Hirose, H., Toyoshima, K., Ushio, T., Mega, T., Shige, S., Yamamoto, M., and Yatagai, A. (2016) Detection of precipitation-related variables revealed by geostationary meteorological satellites, and these applications. In: Proceedings of 2016 Annual Conference, Japan Society of Hydrology and Water Resources, vol. 18.
- Hsu, K., Gao, X., Sorooshian, S. and Gupta, H.V. (1997) Precipitation estimation from remotely sensed information using artificial neural networks. *Journal of Applied Meteorology*, 36(9), 1176–1190. [https://doi.org/10.1175/1520-0450\(1997\)036<1176:PEFRSI>2.0.CO;2](https://doi.org/10.1175/1520-0450(1997)036<1176:PEFRSI>2.0.CO;2).
- Huete, A., Justice, C. and van Leeuwen, W. (1999) *MODIS Vegetation Index (MOD 13) ATBD version 3*.
- Huffman, G.J., Bolvin, D.T., Nelkin, E.J., Wolff, D.B., Adler, R.F., Gu, G., Hong, Y., Bowman, K.P. and Stocker, E.F. (2007) The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, 8(1), 38–55. <https://doi.org/10.1175/JHMS60.1>.
- Huntzinger, D.N., Post, W.M., Wei, Y., Michalak, A.M., West, T.O., Jacobson, A.R., Baker, I.T., Chen, J.M., Davis, K.J., Hayes, D.J., Hoffman, F.M., Jain, A.K., Liu, S., McGuire, A.D., Neilson, R.P., Potter, C., Poulter, B., Price, D., Raczka, B.M., Tian, H.Q., Thornton, P., Tomelleri, E., Viovy, N., Xiao, J., Yuan, W., Zeng, N., Zhao, M. and Cook, R. (2012) North American Carbon Program (NACP) regional interim synthesis: terrestrial biospheric model intercomparison. *Ecological Modelling*, 232, 144–157.
- Hutchinson, M.F. (1995) Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Systems*, 9(4), 385–403. <https://doi.org/10.1080/02693799508902045>.
- Jing, W., Yang, Y., Yue, X. and Zhao, X. (2016) A comparison of different regression algorithms for downscaling monthly satellite-based precipitation over North China. *Remote Sensing*, 8(10), 835. <https://doi.org/10.3390/rs8100835>.
- Jolly, W.M., Graham, J.S., Michaelis, A., Nemani, R.R. and Running, S.W. (2005) A flexible, integrated system for generating meteorological surfaces derived from point sources across multiple geographic scales. *Environmental Modelling and Software*, 20(7), 873–882. <https://doi.org/10.1016/j.envsoft.2004.05.003>.
- Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Jenne, R. and Joseph, D. (1996) The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3), 437–471.
- Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.K., Hnilo, J.J., Fiorino, M. and Potter, G.L. (2002) NCEP–DOE AMIP-II reanalysis (R-2). *Bulletin of the American Meteorological Society*, 83(11), 1631–1643. <https://doi.org/10.1175/BAMS-83-11-1631>.
- Knapp, K.R., Ansari, S., Bain, C.L., Bourassa, M.A., Dickinson, M.J., Funk, C., Helms, C.N., Hennon, C.C., Holmes, C.D., Huffman, G.J., Kossin, J.P., Lee, H.T., Loew, A. and Magnusdottir, G. (2011) Globally gridded satellite observations for climate studies. *Bulletin of the American Meteorological Society*, 92(7), 893–907. <https://doi.org/10.1175/2011BAMS3039.1>.
- Saha, S., Moorthi, S., Pan, H.L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.T., Chuang, H.Y., Juang, H.-M.H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Van Den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R.W., Rutledge, G. and Goldberg, M. (2010) The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, 91(8), 1015–1057. <https://doi.org/10.1175/2010BAMS3001.1>.
- Livneh, B., Rosenberg, E.A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K.M., Maurer, E.P. and Lettenmaier, D.P. (2013) A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: update and extensions. *Journal of Climate*, 26(23), 9384–9392. <https://doi.org/10.1175/JCLI-D-12-00508.1>.
- McGuire, C.R., Nufio, C.R., Bowers, M.D. and Guralnick, R.P. (2012) Elevation-dependent temperature trends in the Rocky Mountain front range: changes over a 56- and 20-year record. *PLoS ONE*, 7(9), e44370. <https://doi.org/10.1371/journal.pone.0044370>.
- Meir, T., Orton, P.M., Pullen, J., Holt, T., Thompson, W.T. and Arend, M.F. (2013) Forecasting the New York City urban heat Island and sea breeze during extreme heat events. *Weather and Forecasting*, 28(6), 1460–1477. <https://doi.org/10.1175/WAF-D-13-00012.1>.
- Menne, M.J. and Williams, C.N. (2009) Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22(7), 1700–1717. <https://doi.org/10.1175/2008JCLI2263.1>.
- Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P.C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E.H., Ek, M.B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D. and Shi, W. (2006) North American regional reanalysis. *Bulletin of the American Meteorological Society*, 87(3), 343–360. <https://doi.org/10.1175/BAMS-87-3-343>.
- Menne, M.J., Williams, C.N. and Vose, R.S. (2009) The U.S. historical climatology network monthly temperature data, version 2. *Bulletin of the American Meteorological Society*, 90(7), 993–1007. <https://doi.org/10.1175/2008BAMS2613.1>.
- Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E. and Houston, T.G. (2012) An overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, 29(7), 897–910. <https://doi.org/10.1175/JTECH-D-11-00103.1>.
- Mourtzinis, S., Edreirab, J.I.R., Conleya, S.P. and Grassinib, P. (2017) From grid to field: assessing quality of gridded weather data for agricultural applications. *European Journal of Agronomy*, 82, 163–172. <https://doi.org/10.1016/J.EJA.2016.10.013>.
- National Renewable Energy Laboratory. (1992) *User’s Manual: National Solar Radiation Data Base (1961–1990)*.

- National Renewable Energy Laboratory. (2007) *National Solar Radiation Database 1991–2005 Update: User's Manual*.
- NCDC. (1981) *Daily meteorological data for U.S. cooperative stations from NCDC TD3200*. Boulder, CO: Research Data Archive, National Center for Atmospheric Research, Computational and Information Systems Laboratory.
- NCDC. (2002) *National Climatic Data Center data documentation for data set 6421 (DSI-6421): enhanced hourly wind station data for the contiguous United States*.
- Nemani, R., Hashimoto, H., Votava, P., Melton, F., Wang, W., Michaelis, A., Mutch, L., Milesi, C., Hiatt, S. and White, M. (2009) Monitoring and forecasting ecosystem dynamics using the Terrestrial Observation and Prediction System (TOPS). *Remote Sensing of Environment*, 113(7), 1497–1509. <https://doi.org/10.1016/j.rse.2008.06.017>.
- Newman, A.J., Clark, M.P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., Mizukami, N., Brekke, L. and Arnold, J.R. (2015) Gridded ensemble precipitation and temperature estimates for the contiguous United States. *Journal of Hydrometeorology*, 16(6), 2481–2500. <https://doi.org/10.1175/JHM-D-15-0026.1>.
- NOAA. (1998) *Automated Surface Observing System (ASOS) User's Guide*.
- Oshiro, T.M., Perez, P.S. and Baranauskas, J.A. (2012) How many trees in a random forest? In: Perner, P. (Ed.) *Machine Learning and Data Mining in Pattern Recognition*. Berlin-Heidelberg: Springer, pp. 154–168. https://doi.org/10.1007/978-3-642-31537-4_13.
- Oyler, J.W., Ballantyne, A., Jencso, K., Sweet, M. and Running, S.W. (2015a) Creating a topoclimatic daily air temperature dataset for the conterminous United States using homogenized station data and remotely sensed land skin temperature. *International Journal of Climatology*, 35(9), 2258–2279. <https://doi.org/10.1002/joc.4127>.
- Oyler, J.W., Dobrowski, S.Z., Ballantyne, A.P., Klene, A.E. and Running, S.W. (2015b) Artificial amplification of warming trends across the mountains of the western United States. *Geophysical Research Letters*, 42(1), 153–161. <https://doi.org/10.1002/2014GL02803>.
- Oyler, J.W., Dobrowski, S.Z., Holden, Z.A. and Running, S.W. (2016) Remotely sensed land skin temperature as a spatial predictor of air temperature across the conterminous United States. *Journal of Applied Meteorology and Climatology*, 55(7), 1441–1457. <https://doi.org/10.1175/JAMC-D-15-0276.1>.
- Pal, M. (2005) Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1), 217–222. <https://doi.org/10.1080/01431160412331269698>.
- Parmentier, B., McGill, B.J., Wilson, A.M., Regetz, J., Jetz, W., Guralnick, R., Tuanmu, M.-N. and Schilthauer, M. (2015) Using multi-timescale methods and satellite-derived land surface temperature for the interpolation of daily maximum air temperature in Oregon. *International Journal of Climatology*, 35(13), 3862–3878. <https://doi.org/10.1002/joc.4251>.
- Pierce, D.W., Cayan, D.R. and Thrasher, B.L. (2014) Statistical downscaling using localized constructed analogs (LOCA). *Journal of Hydrometeorology*, 15(6), 2558–2585. <https://doi.org/10.1175/JHM-D-14-0082.1>.
- Reges, H.W., Doesken, N., Turner, J., Newman, N., Bergantino, A., Schwalbe, Z., Reges, H.W., Doesken, N., Turner, J., Newman, N., Bergantino, A. and Schwalbe, Z. (2016) CoCoRaHS: The evolution and accomplishments of a volunteer rain gauge network. *Bulletin of the American Meteorological Society*, 97(10), 1831–1846. <https://doi.org/10.1175/BAMS-D-14-00213.1>.
- Shi, Y. and Song, L. (2015) Spatial downscaling of monthly TRMM precipitation based on EVI and other geospatial variables over the Tibetan Plateau from 2001 to 2012. *Mountain Research and Development*, 35(2), 180–194. <https://doi.org/10.1659/MRD-JOURNAL-D-14-00119.1>.
- Smith, R.J. (2009) Use and misuse of the reduced major axis for line-fitting. *American Journal of Physical Anthropology*, 140(3), 476–486. <https://doi.org/10.1002/ajpa.21090>.
- Smith, A., Lott, N. and Vose, R. (2011) The integrated surface database: recent developments and partnerships. *Bulletin of the American Meteorological Society*, 92(6), 704–708. <https://doi.org/10.1175/2011BAMS3015.1>.
- Thiessen, A.H. (1911) Precipitation averages for large areas. *Monthly Weather Review*, 39(7), 1082–1089. [https://doi.org/10.1175/1520-0493\(1911\)39<1082b:PAFLA>2.0.CO;2](https://doi.org/10.1175/1520-0493(1911)39<1082b:PAFLA>2.0.CO;2).
- Thornton, P.E., Running, S.W. and White, M.A. (1997) Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology*, 190(3–4), 214–251. [https://doi.org/10.1016/S0022-1694\(96\)03128-9](https://doi.org/10.1016/S0022-1694(96)03128-9).
- Vermote, E.F. and Vermeulen, A. (1999) *MODIS Algorithm Technical Background Document: Atmospheric Correction Algorithm: Spectral Reflectances (MOD09)*, p. 107.
- Vose, R.S., Applequist, S., Squires, M., Durre, I., Menne, C.J., Williams, C.N., Fenimore, C., Gleason, K. and Arndt, D. (2014) Improved historical temperature and precipitation time series for U.S. climate divisions. *Journal of Applied Meteorology and Climatology*, 53(5), 1232–1251. <https://doi.org/10.1175/JAMC-D-13-0248.1>.
- Wan, Z. (1999) *MODIS Land-Surface Temperature Algorithm Theoretical Basis Document (LST ATBD)*.
- Wang, J. and Kotamarthi, V.R. (2014) Downscaling with a nested regional climate model in near-surface fields over the contiguous United States. *Journal of Geophysical Research: Atmospheres*, 119(14), 8778–8797. <https://doi.org/10.1002/2014JD021696>.
- Wessel, P. and Smith, W.H.F. (1996) A global, self-consistent, hierarchical, high-resolution shoreline database. *Journal of Geophysical Research: Solid Earth*, 101(B4), 8741–8743. <https://doi.org/10.1029/96JB00104>.
- Willmott, C.J., Rowe, C.M. and Philpot, W.D. (1985) Small-scale climate maps: a sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring. *American Cartographer*, 12(1), 5–16. <https://doi.org/10.1559/152304085783914686>.
- Wu, Z., Ahlström, A., Smith, B., Ardö, J., Eklundh, L., Fensholt, R. and Lehsten, V. (2017) Climate data induced uncertainty in model-based estimations of terrestrial primary productivity. *Environmental Research Letters*, 12(6), 064013. <https://doi.org/10.1088/1748-9326/aa6fd8>.
- Yang, F., Zhu, A.-X., Ichii, K., White, M.A., Hashimoto, H. and Nemani, R.R. (2008) Assessing the representativeness of the AmeriFlux network using MODIS and GOES data. *Journal of Geophysical Research*, 113(G4), G04036. <https://doi.org/10.1029/2007JG000627>.
- Zhang, G. and Lu, Y. (2011) Bias-corrected random forests in regression. *Journal of Applied Statistics*, 39(1), 151–160. <https://doi.org/10.1080/02664763.2011.578621>.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Hashimoto H, Wang W, Melton FS, *et al.* High-resolution mapping of daily climate variables by aggregating multiple spatial data sets with the random forest algorithm over the conterminous United States. *Int J Climatol*. 2019;39: 2964–2983. <https://doi.org/10.1002/joc.5995>