https://doi.org/10.5194/egusphere-2024-1886
Preprint. Discussion started: 17 July 2024

A Method for Quantifying Uncertainty in Spatially Interpolated Meteorological Data with Application to Daily Maximum Air Temperature

Conor T. Doherty[1], Weile Wang[1], Hirofumi Hashimoto[1,2], Ian G. Brosnan[1]

[1]NASA Ames Research Center, Moffett Field, CA, 94035, USA
[2]California State University Monterey Bay, Seaside, CA, 93955, USA

Correspondence to: Conor T. Doherty (conor.t.doherty@nasa.gov)

Abstract. Conventional point estimate error statistics are not well-suited to describing spatial and temporal variation in the accuracy of spatially interpolated meteorological variables. This paper describes, applies, and evaluates a method for quantifying prediction uncertainty in spatially interpolated estimates of meteorological variables. The approach presented here, which we will refer to as DNK for "detrend, normal score, krige," uses established methods from geostatistics and we apply them to interpolate data from ground-based weather stations. The method is validated using daily maximum near-surface air temperature (Tmax). Uncertainty is inherent in gridded meteorological data, but this fact is often overlooked when data products provide single-value point estimates without a quantitative description of prediction uncertainty. Uncertainty varies as a function of spatial factors, like distance to the nearest measurement location, and temporal factors, like seasonality in sample heterogeneity. DNK produces not only point estimates but predictive distributions for each location. Predictive distributions quantitatively describe uncertainty suitably for propagation into physical models that take meteorological variables as inputs. We validate the uncertainty quantification by comparing theoretical versus actual coverage of prediction intervals computed at locations where measurement data were held out from the estimation procedure. We find that, for most days, the predictive distributions accurately quantify uncertainty and that theoretical versus actual coverage levels of prediction intervals closely match one another. Even for days with the worst agreement, the predictive distributions meaningfully convey the relative certainty of predictions for different locations in space. After validating the methodology, we demonstrate how the magnitude of prediction uncertainty varies significantly in both space and time. Finally, we examine spatial correlation in predictive distributions by using conditional Gaussian simulation in place of kriging. We conclude that spatial correlation in Tmax errors is relatively small, and that less computationally expensive kriging-based methods will suffice for many applications.

1 Introduction

Interpolated meteorological data products are widely used in the geosciences, but relatively little attention is paid to the errors they contain. For example, when studying terrestrial fluxes of carbon, water, and energy over a large spatial domain (e.g., ≥100 km^2), it is necessary to work with gridded meteorological data. Ground-based weather stations may be sparse or only cover a small fraction of the study area so gridded estimates, rather than station measurements, of meteorological variables are used by models of land surface processes (Zeng et al., 2020; Volk et al., 2024). In many gridded data products, the values are point estimates (i.e., a single number rather than a range or distribution). When given only gridded point estimates, data users do not know, and cannot propagate, the uncertainty in the meteorological inputs to their model. While users may refer to point estimate accuracy statistics for the data product, these statistics only capture errors at locations where measurements are available. For applications that are particularly sensitive to meteorological inputs, such as evapotranspiration modeling, uncertainty in gridded data can contribute significantly to downstream model errors (Doherty et al., 2022). While geostatistical uncertainty quantification is standard practice in other domains like mining (Rossi and Deutsch, 2014), oil and gas exploration (Pyrcz and Deutsch, 2014), and hydrogeology (Kitanidis, 1997), these methods are not used in the most popular near-surface meteorological data products. Understanding uncertainty in gridded meteorological data is necessary to evaluate the robustness of scientific findings, especially when designing and implementing public policy based on those findings (Morgan and Henrion, 1990).

Existing gridded meteorological datasets fall into two main categories of methods: statistical interpolation methods, where there is no dynamic physical model, and data assimilation methods, which combine dynamic physical models with data-driven adjustments. For North America, DayMET (Thornton et al., 1997; Thornton et al., 2021) and PRISM (Daly et al., 2008), which produce estimates of several meteorological variables on fine spatial grids (~1 km^2), are widely used statistical interpolation products. A related product, NEX-GDM (Hashimoto et al., 2019), uses machine learning and a wide range of inputs to produce high resolution gridded meteorological values. Regarding uncertainty, Daly et al. (2008) describes a method for creating prediction intervals, but the resulting maps are not publicly distributed. Thornton et al. (2021) includes an extensive accuracy assessment using cross validation, but the methodology does not produce spatially resolved uncertainty estimates. A wide range of data assimilation products are available including regional products like RTMA (De Pondeca et al., 2011) CONUS404 (Rasmussen et al. 2023), and global ones like MERRA2 (Gelaro et al., 2017) and ERA5 (Hersbach et al., 2020; Bell et al., 2021). Some assimilation products, like ERA5, express uncertainty using an ensemble of model runs, where a greater magnitude of spread in the ensemble is taken to indicate greater uncertainty. However, the computational expense of large-scale climate simulations generally means that the resulting data products have relatively coarse spatial resolution (31 km horizontal resolution for ERA5) and ensembles that are not large enough (tens of ensemble members) to characterize stable empirical distributions. In contrast, the approach described in this work is computationally efficient enough to be run at fine spatial resolution over large areas while also giving a robust description of the predictive distribution.

In this paper we present and analyze a statistical method to produce spatially and temporally resolved uncertainty quantification and apply it to the interpolation of daily maximum near-surface air temperature (Tmax). We will refer to the approach for estimation and uncertainty quantification as DNK for “detrend, normal score, krige.” The basic approach of DNK is well-established in geostatistics, appearing in textbooks such as Olea (1999), Goovaerts (1997), and others. While kriging and related spatial regression methods have previously been used for meteorological data interpolation, they have only been used to produce gridded point estimates. As we will explain, the “detrend” and “normal score” steps in DNK enable us to compute theoretically correct predictive distributions at each prediction location. A central component on this work is to test the validity of predictive distributions generated using DNK and, as such, their utility for uncertainty quantification. Uncertainty is not intrinsic to macro-scale physical phenomenon but rather is a property of the combination of data and a model (Goovaerts, 1997), which means that there is not an objective "correct" predictive distribution for a given unknown value. However, we can assess the validity of a collection of predictive distributions, in aggregate, by testing the rate at which true measurement data fall within prediction intervals relative to those intervals' theoretical coverage. If the validity of predictive distributions can be established, then DNK can accurately quantify uncertainty in gridded meteorological data.

2 Methods

2.1 Input Data

This study uses two sets of input data: daily maximum air temperature at 2m (Tmax) and elevation. Tmax data are provided by Thornton et al. (2022) for stations in the Global Historical Climatology Network (GHCN) (Menne et al., 2012), a database of measurement data from ground-based weather stations across the world. The GHCNd (daily) data are processed as described in Thornton et al. (2021) to correct for temperature sensor biases and inconsistencies in time of observation. Figure 1 shows the spatial distribution of weather stations across the study area. The number of weather stations within the California state boundary ranges between 524 and 542 stations depending on the day of year. Data from stations within the state boundaries are used as ground truth for validation. Data from stations outside the study area contribute to predictions at locations near the boundary but these stations are not, themselves, used as validation locations. Elevation data are sourced from a digital elevation model (DEM) with 90 m resolution for the western United States (Hanser, 2008). The DEM is clipped to the boundaries of the study area and then resampled using mean resampling to a grid with 1 km^2 grid cells.

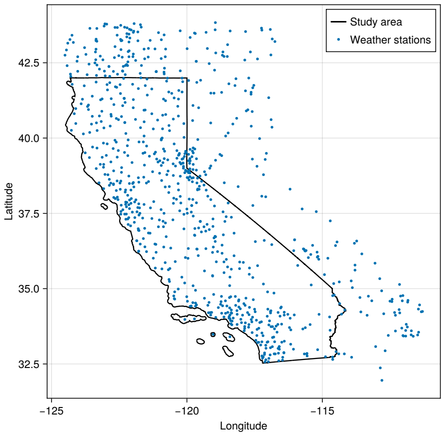

Figure 1: Study area and locations of weather stations. Black lines mark the bounds of the study area (the state of California). Blue dots mark the locations of GHCN weather stations that were active in 2022. The number of weather stations used on each day varies depending on the availability of data. Weather stations outside the study area were not used as prediction locations but were used in estimation for grid cells near the study area boundaries.

2.2 Estimation Methods for Spatial Interpolation

The primary estimation method we consider is Ordinary Kriging (OK) (see e.g., Goovaerts, 1997; Olea, 1999; Anderes, 2012), which gives analytical solutions for both a point estimate and the variance of a predictive distribution. For a random field that is (1) covariance stationary and (2) Gaussian, the OK prediction mean and variance completely characterize the predictive distribution. In general, spatially distributed Tmax data satisfy neither assumption, so data transformations must be applied (and later reverted) to remove spatial trends and induce normality. These transformations are described in Sect. 2.3 Data Transformations. Under assumptions of stationarity and normality, OK is equivalent to Gaussian process regression with a constant unknown mean. We apply OK locally at each prediction location using measurement data within a 100 km radius of the location in the estimation. Figure 2G-H shows an example of the gridded estimates produced by OK from the point measurement data shown in Fig. 2E-F.

In addition to OK, we also demonstrate spatial uncertainty quantification using conditional Gaussian simulation (CGS). Samples generated by CGS are equally probable "realizations" of the underlying random field that produced the measurement data. Under assumptions (1) and (2), local predictive distributions (i.e., the marginal predictive distribution at a given grid cell) generated by CGS are equivalent to the distribution defined by the kriging variance (Goovaerts, 2001). However, CGS produces realizations that are spatially coherent with respect to the model of spatial covariance (see Sect. 2.4 Variography), whereas gridded estimates produced by OK do not. As such, CGS can be used to express "spatial uncertainty," or spatial correlation in errors, as described by the joint predictive distribution over multiple grid cells. CGS is performed using a method based on a decomposition of the conditional distribution covariance matrix, commonly referred to in geostatistics literature as the "LU method" of CGS (Alabert, 1987). Both OK and CGS are performed using the GeoStats.jl library (Hoffimann, 2018).

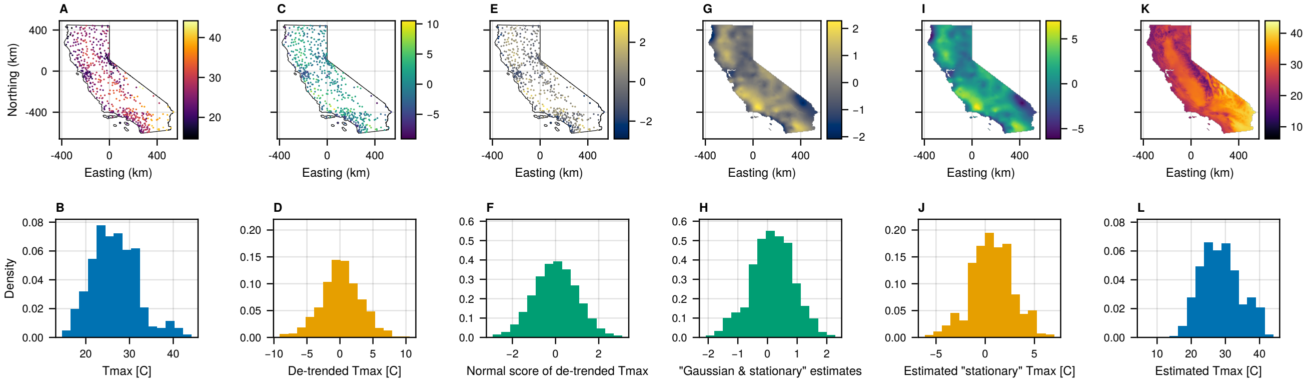

Figure 2: Illustration of data processing steps. Each column corresponds to a step in the processing and estimation pipeline. The top row shows point (panels A, C, E) and gridded (panels G, I, K) data values. The bottom row shows histograms of the data at each processing step. Panels A-B show the measured Tmax values. Panels C-D show the Tmax values after spatial trends have been estimated and subtracted. Panels E-F show the detrended Tmax data after the normal score transform has been applied. Panels G-H show gridded estimates produced

by OK in the detrended and Gaussian space. Panels I-J show gridded estimates after reverting the normal score transform using quantile information from the distributions in panels C-D. Panels K-L show the final gridded estimates after re-adding spatial trends.

2.3 Pre- and Post-Estimation Data Transformations

There are two transformations that are applied to the Tmax data prior to estimation, and then reverted after, in the DNK procedure. The first attempts to model and then remove local trends ("detrend") to bring the data closer to stationarity. The second transforms the data to be approximately Gaussian by applying a "normal score transformation." These transformations enable us to utilize properties of the multivariate Gaussian to derive predictive distributions (see e.g., Olea 1999). After estimation (interpolation) is performed in the de-trended and Gaussian space, the transformations are reverted to bring the data back to their original non-Gaussian and nonstationary distribution. We describe some practical challenges related to these transformations in the Discussion.

The first transformation models and then subtracts local spatial trends. The purpose of trend modeling is to identify and remove variation at coarse spatial scales that would otherwise make the data nonstationary. We use the term "trend" to refer both to large scale variation in longitude-latitude space and variation due to change in elevation (lapse rate). While some prior approaches assume a fixed lapse rate (Hart et al., 2008), we follow DayMET (Thornton et al., 1997; Thornton et al. 2021) and PRISM (Daly et al., 2008) in allowing lapse rate to vary in space. Local trends are estimated by regressing Tmax values on spatial coordinates and elevation for all weather stations within a 100 km search radius of a given station (including the station itself). The residual at the station is saved as the detrended Tmax. The detrended station-level Tmax values are shown in Fig. 2B-C. A similar procedure is applied to "add back" the trend after estimation: the trend parameters are estimated at locations centered on each grid cell, again using data from weather stations within 100 km. The cell-wise trend value is calculated using the regression parameter estimates and the corresponding coordinates of the cell and elevation from a digital elevation model. The gridded estimates with the trend added back is shown in Fig. 2K-L.

The second transformation, a "normal score transformation," transforms the data to be approximately Gaussian. This is done by mapping quantiles of the empirical distribution of detrended Tmax values to the corresponding quantiles of a standard normal distribution. The transformed station-level data are shown in Fig. 2C-D. The quantile information from the original empirical distribution is saved so that the normal score transformation can be reverted after estimation. The gridded estimates, after reverting the normal score transform, are shown in Fig. 2I-J. This transformation is implemented in TableTransforms.jl.

2.4 Variography

Both OK and CGS rely on a model of spatial covariance to produce gridded Tmax estimates and prediction uncertainty. Traditionally in geostatistics, spatial variation is represented using a semi-variogram, which is a function that describes the decrease in correlation between two locations as the distance between them increases (see e.g., Olea, 1999; Goovaerts, 1997).

In this study, variography is performed using functions from the GeoStats.jl library (Hoffimann, 2018). We model the theoretical semi-variogram using the pentaspherical function:

$$\gamma(h) = (s-n)\left[\left(\frac{15}{8}\left(\frac{h}{r}\right) - \frac{5}{4}\left(\frac{h}{r}\right)^3 + \frac{3}{8}\left(\frac{h}{r}\right)^5\right)\cdot 1_{(0,r)}(h) + 1_{[r,\infty)}(h)\right] + n\cdot 1_{(0,\infty)}(h) \tag{1}$$

where the variable h is the lag distance and s, n, and r are parameters estimated to fit the empirical semi-variogram representing the sill (value of γ as $h \to \infty$), nugget (value of γ as $h \to 0$), and range (roughly the value of h where γ “levels off”), respectively. $1_{(l,u)}(h)$ is an indicator function that is 1 if $l < h < u$ and 0 otherwise. At shorter lag distances, the pentaspherical model produces semi-variances between that of exponential and spherical models, which have been used previously to interpolate near-surface air temperature (Menafoglio et al., 2013; Hudson and Wackernagel, 1994). While the Tmax data also seem to reasonably support the use of the exponential model, we found that fitting the pentaspherical model was more numerically stable given the spatial sparsity of the weather station data. The theoretical model is fit to the empirical semi-variogram by minimizing the sum of squared errors with equal weight given to each lag bin. Model fitting is performed using the detrended and normal scored data (i.e., the data in Figure 2E-F).

2.5 Validation of Predictive Distributions

We implement a validation scheme that evaluates the accuracy of the predictive distributions in quantifying prediction uncertainty. Uncertainty is not intrinsic to the physical phenomenon and, as such, there is not an objective “correct” predictive distribution. However, a collection of “valid” predictive distributions should produce statistics that reflect appropriate levels of confidence in aggregate.

To assess the validity of local predictive distributions, we use a strategy based on prediction intervals described by Deutsch (1997). For each day of year, we perform leave-one-out (LOO) cross-validation with each weather station. The measurement at the left-out station is not used in trend modeling, variography, or estimation. Predictions are made for the point at the center of the grid cell containing the weather station that will be used for validation. Estimates of the multiples of 0.5 percentiles are produced for each predictive distribution, from which centered prediction intervals are calculated with p% coverage for $p \in \{1, 2, \ldots, 99\}$. For each p% prediction interval, let $p_{low} = (1-p)/2$ and $p_{upp} = (1+p)/2$ be the lower and upper bound of the theoretical prediction interval. Then for $Tmax_i$, the measured Tmax at weather station i, define an indicator function:

$$\xi(Tmax_i\,; p) = \begin{cases} 1, \text{if } Tmax_i \in (p_{low}, p_{upp}) \\ 0, \text{otherwise} \end{cases} \tag{2}$$

Then for each p% prediction interval, we compute an average over all n stations as $\overline{\xi(p)} = \frac{1}{n}\sum_{i=1}^{n}\xi(Tmax_i\,; p)$. For exactly valid predictive distributions, $\overline{\xi(p)} = p$ for any p. To summarize the errors, we calculate the mean bias error as $\frac{1}{99}\sum_{p=1}^{99}\overline{\xi(p)} - p$ and the mean absolute error (MAE) as $\frac{1}{99}\sum_{p=1}^{99}\left|\overline{\xi(p)} - p\right|$. The MAE weights errors from being too confident and too conservative equally. The bias indicates whether the predictive distributions are too confident (negative) or too conservative (positive) on average.

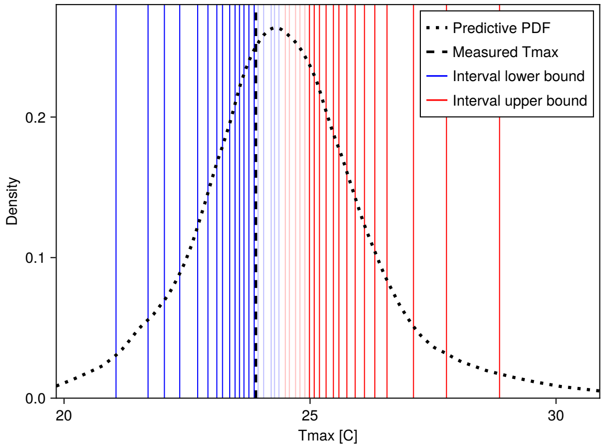

Figure 3: Prediction intervals versus measured Tmax. The lower bounds (blue) and upper bounds (red) of prediction intervals with coverage $p \in \{5, 10, \ldots, 95\}$ for a single prediction location are shown. (Validation statistics are computed using $p \in \{1, 2, \ldots, 99\}$, but we show fewer intervals here for legibility.) Intervals that contain the measured Tmax (black dashed line) are drawn as solid lines and the intervals that do not are drawn as partially transparent. The dotted black line shows a kernel density estimate of the predictive distribution.

3 Results

We first present the validation of local uncertainty quantification using the predictive distributions from OK. Figure 4 shows the average accuracy of the predictive distributions using the LOO validation scheme described in in Sect. 2.5. Figure 4A shows the mean absolute error (MAE) in terms of the predicted versus actual proportion of Tmax values that fall within a given prediction interval. For example, an MAE of 0.01 indicates that for a p% prediction interval, the true value fell within that interval $(p \pm 1)$% of the time. The largest MAE of 0.046 occurs on DOY 96. The median MAE is 0.013 and 84% of days had an MAE less than 0.02. Fig. 4B shows the average bias for each day. For example, a bias of -0.01 indicates that for a p % prediction interval, the true value fell within that prediction interval $(p - 1)$% of the time. Positive bias indicates that the prediction intervals are too conservative on average, and negative bias indicates that the prediction intervals are too confident on average. The largest positive bias of +0.046 occurred on DOY 96 and the largest negative bias of -0.040 occurred on DOY

343. The median bias was -0.002. There are temporal patterns with the bias metrics errors of similar sign and magnitude persisting for periods ranging from a few days to multiple weeks.

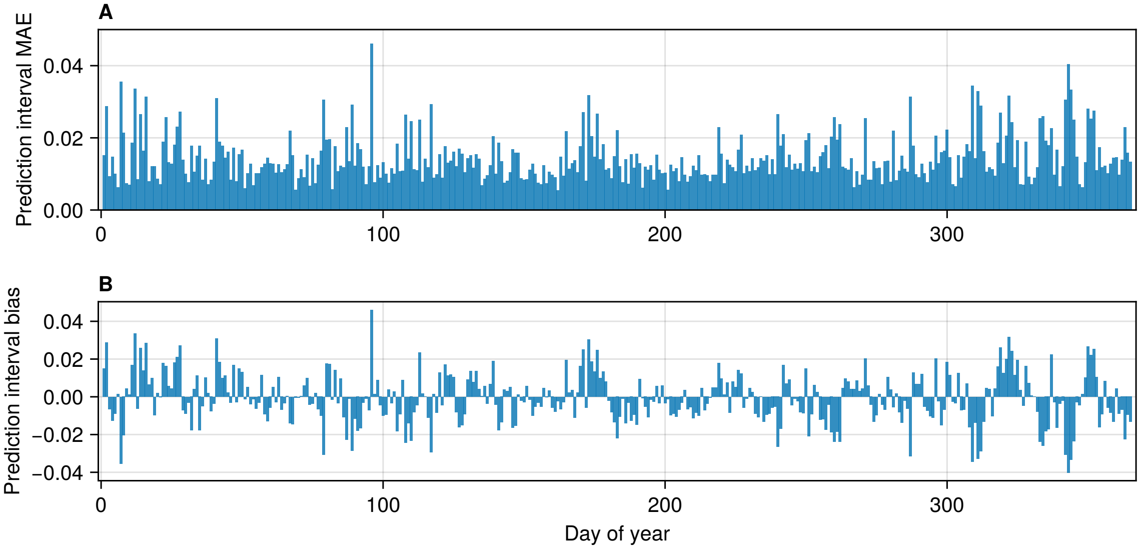

Figure 4: Predictive distribution quantitative validation statistics. Validity is assessed by comparing the theoretical coverage versus actual rate that measurement data fall within prediction intervals. The calculation uses an indicator function ξ described in Eq. (1). Panel A shows the mean absolute error (MAE) for each day of year, calculated as $\frac{1}{99}\sum_{p=1}^{99}\left|\overline{\xi(p)} - p\right|$. Panel B shows the mean bias error for each day of year, $\frac{1}{99}\sum_{p=1}^{99}\overline{\xi(p)} - p$. A positive bias indicates that predictive distributions were too conservative on average and negative bias indicates they were too confident on average.

Figure 5 shows validation results for three individual days. In addition to showing the day with the median MAE, we highlight the two days with the most significant errors in the sample to give a sense of the "worst case" accuracy in uncertainty quantification. Values in the x-direction correspond to theoretical prediction intervals centered on the median. Values in the y-direction are the actual proportion of true Tmax values that fall within the given theoretical interval. For example, a point at (0.25, 0.3) would mean that 30% of true values fell within the corresponding 25% prediction intervals. Figure 5A shows results for DOY 77, which had the median MAE of 0.013. For DOY 77, the actual proportions closely track the theoretical prediction intervals with the largest error occurring at the 29% intervals, which contained 26.4% of the true values. Figures 5B and 5C show the same information for DOY 96 and 343, which are the days with largest positive and negative biases of +0.046 and -0.040, respectively. The largest error for DOY 96 occurs at the 59% prediction intervals, which contained 67.3% of true Tmax values. The largest error for DOY 343 occurs at the 49% prediction intervals, which contained 41.6% of true Tmax values.

https://doi.org/10.5194/egusphere-2024-1886
Preprint. Discussion started: 17 July 2024

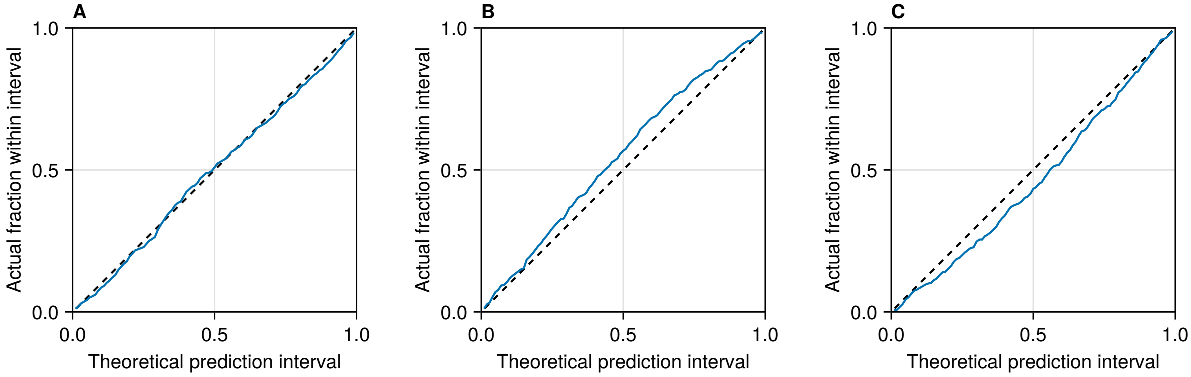

Figure 5: Theoretical versus actual coverage of prediction intervals. X-axes correspond to the coverage of the theoretical prediction interval and Y-axes correspond to the proportion of values that actually fell within the intervals. Panel A shows results for DOY 77, which had the median MAE of all the days in the sample. Panels B and C shows DOY 96 and 343, which had the largest positive and negative biases, respectively.

When applying our method across a full spatial domain, we observe that local predictive distributions vary in space and time. Figure 6 shows predicted Tmax and uncertainty statistics (spread of predictive distributions) for four different days, where each row corresponds to a day. The days are approximately equally spaced and cover different seasons, including the first days of January (Fig. 6A-6C), April (Fig. 6D-6F), July (Fig. 6G-6I), and October (Fig. 6J-6L). The first column (Fig. 6A, 6D, 6G, and 6J) shows the median of the predictive distribution for each 1 km^2 grid cell. The second column (Fig. 6B, 6E, 6H, and 6K) shows the magnitude of the 50% prediction interval of each predictive distribution (the 75th percentile minus the 25th percentile). The third column (Fig. 6C, 6F, 6I, and 6L) shows the magnitude of the 90% prediction interval of each predictive distribution (the 95th percentile minus the 5th percentile). Each column uses a single common color gradient (the gradients do not vary between rows). The Tmax prediction uncertainty varies in both space and time. Spatial patterns in prediction uncertainty are strongly influenced, though not exclusively determined, by the spatial density of weather stations near the prediction location. Prediction uncertainty also increases when the variance in Tmax at nearby stations is greater (the nearby stations have "lower agreement" with one another). Comparing the uncertainty maps in time (between rows), the magnitude, spatial patterns, and magnitude of variation in prediction uncertainty all vary. Spatial patterns also differ based on the type of the prediction interval. The 90% prediction intervals (rightmost column) appear to be controlled primarily by the local spatial density of measurements, with the locations of weather stations standing out as local minima. The patterns in the 50% prediction intervals (center column) are more complex and dependent on local agreement between stations. The fact that many of the locations with the largest uncertainties are near the Pacific coast suggest that the trend model is failing to capture local Tmax trends, where cooling from the ocean confounds the usual negative correlation between temperature and elevation. Using a more sophisticated approach to trend modeling may improve accuracy and reduce uncertainty at these locations (see the Discussion for more on this topic).

https://doi.org/10.5194/egusphere-2024-1886
Preprint. Discussion started: 17 July 2024

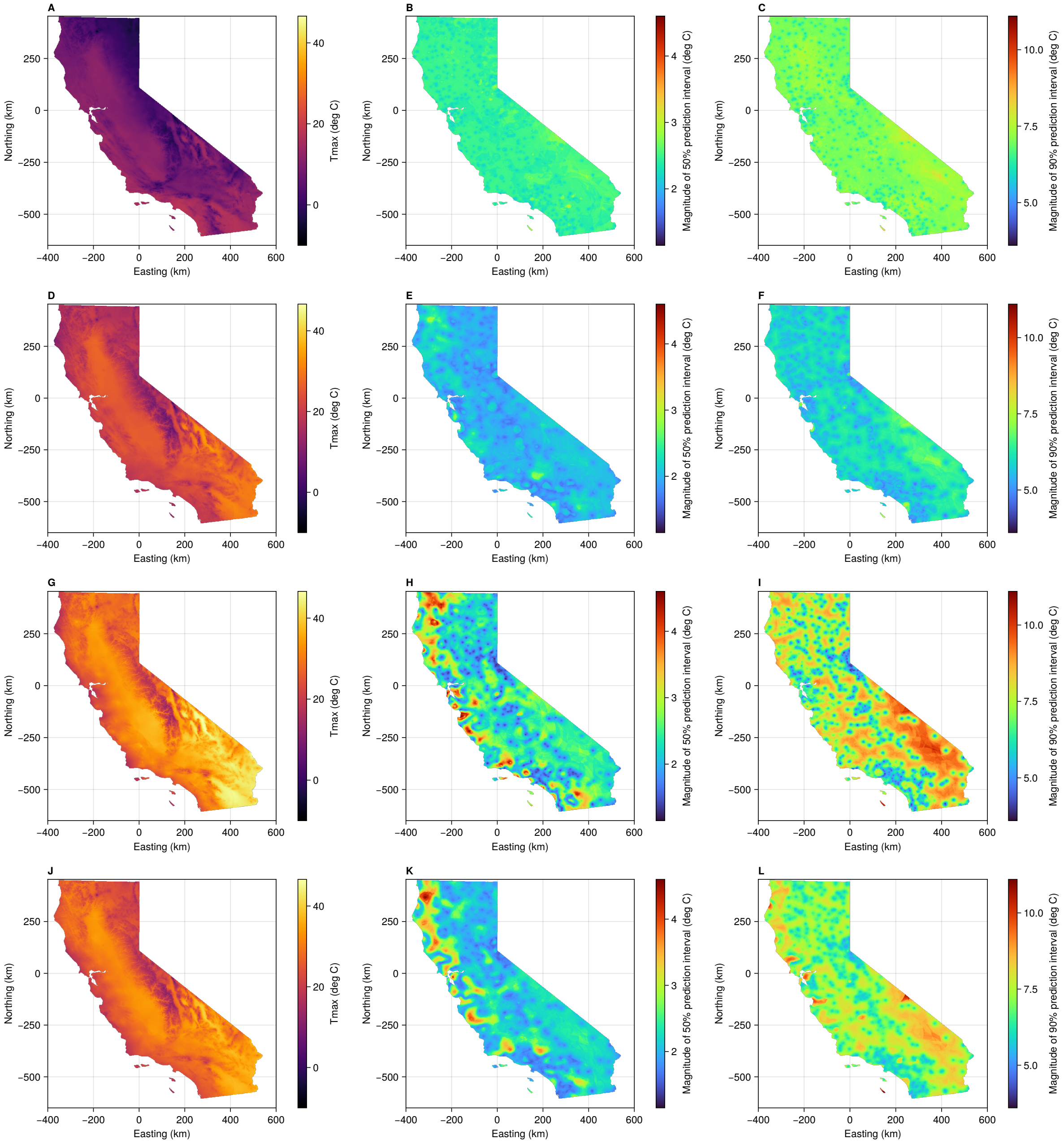

Figure 6: Tmax and prediction uncertainty maps for four different days. Each row shows results for a different day. The days are DOY 1 (panels A-C), DOY 91 (panels D-F), DOY 182 (panels G-I), and DOY 274 (panels J-L) from 2022. The first column (panels A, D, G, J) shows the median of the Tmax predictive distribution at each grid cell. The second column (panels B, E, H, K) shows the magnitude of the 50% prediction interval for each grid cell. The third column (panels C, F, I, L) shows the magnitude of the 90% prediction interval for each grid cell.

In addition to local (cell-wise marginal) uncertainty, we can represent spatial uncertainty using CGS to sample from the joint predictive distribution over multiple grid cells. Figure 7 shows the predictive distributions for grid cells containing two weather stations, IDs USC00043747 ("station 1") and USW00053119 ("station 2"), near Hanford, California. The centers of the grid cells containing the stations are 1.4 km apart. The red lines show kernel density estimates of the joint and marginal predictive distributions generated using CGS. The blue lines show the same predictive distributions generated using OK, where the "joint distribution" is generated by sampling independently from the marginal distributions. While the OK samples show no correlation between locations (by construction), the samples produced by CGS show a correlation of approximately 0.5 between predictions at the two locations. This reveals the fact that errors in predictions and nearby locations are likely to be correlated with one another. The marginal distributions generated using OK and CGS are virtually identical at each of the two locations (Fig. 7A and 7C). This is because the predictive distribution generated by OK is correct in a marginal sense, even though kriging maps have too little variance (are "too smooth") relative to the distribution of measurement data and the theoretical true underlying random field. Reproducing the model of spatial covariance and accounting for spatial correlation in errors (uncertainty) requires sampling from the joint predictive distribution using CGS.

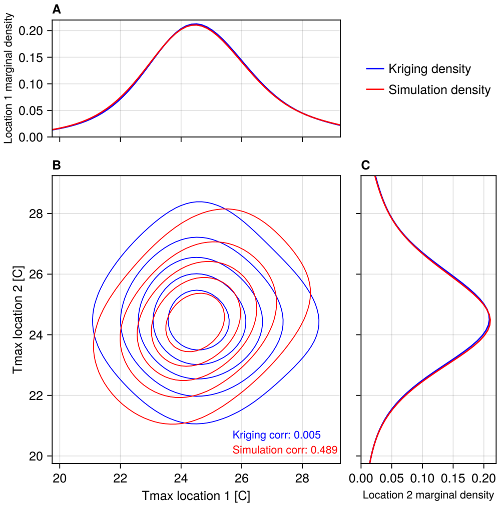

Figure 7: Joint and marginal predictive distributions using OK versus CGS. Predictive distributions were generated for two nearby grid cells that contain weather stations. Panels A and C show the marginal (local) predictive distributions at the two locations and panel B shows the joint (spatial) predictive distribution. The predictive distributions from OK are in blue and the distributions generated by CGS are in red. The two estimation methods produce the same marginal distributions, but the joint distributions are different. The distribution from CGS reflects the spatial correlation in predictions between the two locations.

The impact of spatial correlation in predictive distributions is observable but minimal for Tmax at the two stations from the previous example. Figure 8 shows histograms of predictive distributions for Tmax estimates at Station 1 when conditioned on different information about Tmax at Station 2. Measurements from both stations are held out from the estimation procedure. The dotted black line is the Tmax measured at Station 1. The blue histogram is the unconditional marginal distribution for the grid cell containing Station 1. The orange and green histograms show empirical conditional distributions at Station 1 given that the error at Station 2 is less that 2 degrees ("conditional 2") and less than 1 degree ("conditional 1"), respectively. These empirical distributions are generated by filtering the simulation ensemble. Conditioning the predictive distribution reduces the standard deviation from 2.0 C in the unconditional distribution to 1.7 C and 1.6 C for conditional 2 and conditional 1, respectively. If the mean of the predictive distribution is taken as a point estimate, the error for the unconditional distribution is +0.24 C. The error is reduced to -0.04 for conditional 2 but then increases (in magnitude) to -0.15 for conditional 1. The spread of the predictive distribution is reduced when conditioning on information about a nearby location, but the reduction is relatively small. The small reduction is because the semi-variogram model has a large nugget, meaning that there will still be nonnegligible prediction variance (uncertainty) after conditioning on nearby data. The large nugget estimate may be partly an artifact of the semi-variogram estimation process, given that there are relatively few pairs of weather stations within 1-10 km of one another. However, there does appear to be significant "real" variance in measured Tmax values even at short spatial scales. For example, the two stations used in this example are 1.4 km apart and at nearly identical elevation but have an average difference in Tmax of nearly 1.5 C in 2022. This difference may be explained by site-specific effects, such as different land cover below the weather stations, rather than physical variation that would persist under idealized homogenous conditions.

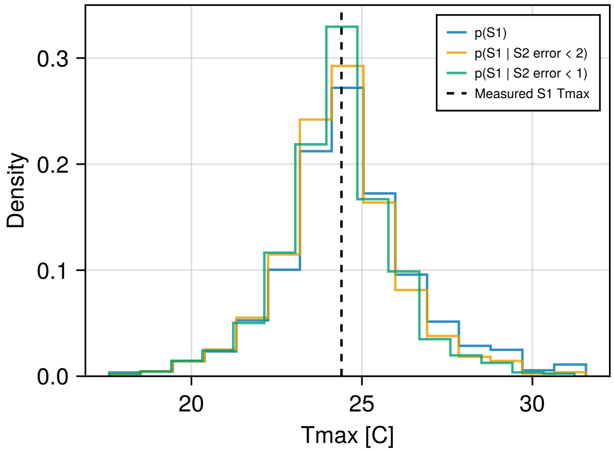

Figure 8: Marginal predictive distributions without and with additional spatial information. Three empirical predictive distributions at Station 1 (S1) are shown including the unconditional (blue) and two distributions conditioned on predictive accuracy at the nearby Station 2 (S2). Conditioning is performed by filtering the ensemble and leaving only realizations where the prediction at S2 was within 2 C (orange) or within 1 C (green) of the measured value. Conditioning on additional spatial information reduces the spread of the marginal predictive distribution, but by a relatively small amount.

4. Discussion and Conclusion

This work presents and validates an approach to quantifying spatially and temporally resolved prediction uncertainty in interpolated meteorological data products. In the quantitative validation study, the DNK method produced highly accurate uncertainty quantification. Even in the least accurate cases, the predictive distributions are qualitatively informative and are still sufficiently accurate to be useful for many applications. Accuracy assessment of point estimates is useful, but fundamentally cannot describe prediction uncertainty at times and locations where measurement data are unavailable. In this application, we have measurements at hundreds of weather stations and predict Tmax values at hundreds of thousands of grid cells, some of which are nearly 100 km from the nearest measurement. This means that the number of locations where we can assess prediction uncertainty using actual measurement data is vanishingly small compared to the number of locations where we do not have measurements. We show that the magnitude of prediction uncertainty varies significantly in space and in time, and that using average error statistics will overestimate prediction uncertainty in some cases and dramatically underestimate it in others.

The DNK methodology described in this paper could be applied to a wide range of applications due to its generality and relatively low computational cost. The main application area motivating this work was modeling of land surface fluxes of water and carbon, given that rates of evapotranspiration (Volk et al., 2024) and primary production (Zeng et al., 2020) are particularly sensitive to near-surface meteorological conditions. However, gridded meteorological data are used to make predictions and draw conclusions about many other phenomena including crop yield (Lobell et al., 2015), vegetation phenology (White et al., 1997), economic productivity (Burke et al., 2015), human conflict (Hsiang et al., 2013), and others. Using gridded predictive distributions rather than point estimates can help ensure the robustness of scientific conclusions given uncertainty in model inputs. The decision to use cell-wise marginal (OK) or full joint (CGS) predictive distributions depends on various factors including the size of grid cells, the distances between locations being compared, and sensitivity of the analysis to spatial correlation in prediction errors (e.g., for causal inference). Users of gridded meteorological data products can propagate uncertainty through their analysis by running models multiple times using either random samples or a preset collection of quantiles from the distribution. This approach does not require any additional modeling choices or assumptions because the relevant information about uncertainty in the model inputs is expressed by the predictive distributions. Also, computationally expensive models may use the gridded variable uncertainty to prioritize sensitivity analyses and reduce the total number of model runs required.

Producing a predictive distribution using DNK, rather than a single point estimate, requires only marginally more computation. The main computation required in OK is solving for the kriging weights λ_s, which requires solving a $(S + 1) \times (S + 1)$ system of linear questions for S stations. The prediction mean μ_{OK} and variance σ_{OK}^2 are both functions of the S station data and the weights λ_s. Drawing samples from the predictive distribution only requires drawing samples from $\mathrm{Normal}(\mu_{OK}, \sigma_{OK}^2)$ and then, for each sample, applying the reverse normal score transformation and adding back the trend. Computing the joint conditional distribution and sampling using CGS is more computationally expensive, as it requires solving the system of kriging equations as well as factorizing a $C \times C$ matrix, where C is the number of prediction grid cells and $C \gg S$. There exist more computationally scalable methods for conditional Gaussian simulation (Gómez-Hernández and Journel, 1993; Gutjahr et al., 1997; Gómez-Hernández and Srivastava, 2021) that we do not discuss here. However, as we show in this paper, the spatial correlation in Tmax errors is weak at multi-kilometer scales and using CGS for this purpose has less practical benefit relative to the additional computation cost.

Users of this methodology should be aware of some practical limitations. First, we reiterate that there is no objectively correct Tmax predictive distribution for a given location. Uncertainty is a property of the measurement data and modeling decisions, and making different modeling decisions will produce different predictive distributions. Different modeling decisions could lead to larger or smaller errors in point estimates on average, but still produce predictive distributions that are valid (i.e., where true values fall in prediction intervals at the prescribed rate). In addition, even large samples from predictive distributions will necessarily suffer from deficiencies inherent in the data-generating process. Covariance stationarity and multi-Gaussianity are strong assumptions that are relied upon for the validity of the predictive distributions, and the transformations made to satisfy these assumptions are imperfect. The normal score transformation requires estimating an empirical distribution from the weather station data. Given that we generally wish to draw samples larger than the size of the measurement data, reverting the normal score transformation necessarily requires interpolation and extrapolation of that distribution (see Goovaerts (1997) for a detailed discussion). In practice, this can produce artifacts like clusters of similar values, particularly near under-sampled edges of the distribution. Regarding stationarity, trend modeling can also strongly influence predictive distributions. Like uncertainty itself, a "spatial trend" is not an objectively observable phenomenon. Reliably estimating spatial trends can be difficult when measurement data are sparse or when other physical phenomena, like cooling or warming due to coastal proximity, confound the basic estimation procedure. Limitations in trend modeling contributed to our use of OK (locally constant unknown mean) rather than Simple Kriging (SK) (locally constant known mean), despite the latter being theoretically justifiable for detrended data. Results using OK versus SK were very similar, with OK performing marginally better likely due to coarse scale variation that was still present after modeling and subtracting spatial trends.

There are many potential avenues for future work building on the methods and results described in this paper. One important area for further study is the analyzing the effects of trend estimation on characteristics and robustness of predictive distributions. For a covariance stationary and multi-Gaussian random field, predictive distributions will be valid over a sufficiently large sample. This indicates that invalid predictive distributions are driven primarily by the transformations we apply (and revert) to make the data stationary and Gaussian. Relatedly, it would be useful to find ways of incorporating additional physical information not explained by a large-scale spatial trend. The strength of data products like NEX-GDM and PRISM come from the use of additional physical information (e.g., coastal proximity, slope, aspect) in predictions. It is not immediately clear how this information could be incorporated into the underlying mathematical model from which our predictive distributions are derived, but doing so could produce more precise estimates accompanied by valid predictive distributions. Lastly, the DNK method should be tested for interpolation of meteorological variables other than Tmax. The methodology seems likely to transfer to certain variables like daily minimum air temperature and humidity, but it is not a given that all variables can be approached the same way.

Going forward, it will be valuable to not only produce the "best available" gridded meteorological data products, but also to produce spatially and temporally resolved uncertainty quantification. Considering uncertainty in model inputs is important for drawing robust scientific conclusions. It is also important for guiding the design and implementation of science-informed policies. Given uncertain information, conclusions about the "best" policy option may differ when using a deterministic versus probabilistic benefit-cost analysis (Morgan and Henrion, 1990). Similarly, there may be asymmetric consequences for over- or underestimation of a given model input. In this scenario, using a predictive distribution rather than a point estimate allows policymakers to quantitatively assess tradeoffs between maximizing expected outcomes and minimizing risk. More broadly, accounting for uncertainty in scientific models is necessary not only for designing informed policies, but for building and maintaining trust in science-informed policymaking.

Code Availability

All code to perform analysis and generate figures is available on Zenodo at https://zenodo.org/doi/10.5281/zenodo.12171025 and on GitHub at https://github.com/conordoherty/met-uncertainty-paper.

Data Availability

Tmax data from Thornton et al. (2022) are available for download from the ORNL DAAC at https://doi.org/10.3334/ORNLDAAC/2132. All other required data are included in the code repository and can be downloaded from Zenodo or GitHub.

Author Contributions

CD and IB developed the concept and acquired funding. CD developed the methodology, developed the software, performed the experiments and formal analysis, and wrote the initial draft of the manuscript. WW contributed to methodology development and formal analysis. HH contributed to methodology development, validation, and software testing. All authors contributed to reviewing and editing the manuscript.

Competing Interests

The authors declare that they have no conflict of interest.

Acknowledgements

CD is supported by an appointment to the NASA Postdoctoral Program at NASA Ames Research Center, administered by Oak Ridge Associated Universities under contract with NASA. CD thanks Forrest Melton for support on the initial conceptualization and proposal for this work, A.J. Purdy and Lee Johnson for helpful feedback on the manuscript, and Kyle Kabasares for performing additional software testing for reproducibility. CD also thanks Júlio Hoffimann for his generous assistance in using the GeoStats.jl library, which he created.

References

Alabert, F., 1987. The practice of fast conditional simulations through the LU decomposition of the covariance matrix. Mathematical Geology 19, 369–386. https://doi.org/10.1007/BF00897191

Anderes, E.B., 2012. Kriging, in: El-Shaarawi, A.H., Piegorsch, W.W. (Eds.), Encyclopedia of Environmetrics. Wiley. https://doi.org/10.1002/9780470057339.vak003.pub2

Bell, B., Hersbach, H., Simmons, A., Berrisford, P., Dahlgren, P., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Radu, R., Schepers, D., Soci, C., Villaume, S., Bidlot, J., Haimberger, L., Woollen, J., Buontempo, C., Thépaut, J., 2021. The ERA5 global reanalysis: Preliminary extension to 1950. Quart J Royal Meteoro Soc 147, 4186–4227. https://doi.org/10.1002/qj.4174

Burke, M., Hsiang, S.M., Miguel, E., 2015. Global non-linear effect of temperature on economic production. Nature 527, 235–239. https://doi.org/10.1038/nature15725

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

De Pondeca, M.S.F.V., Manikin, G.S., DiMego, G., Benjamin, S.G., Parrish, D.F., Purser, R.J., Wu, W.-S., Horel, J.D., Myrick, D.T., Lin, Y., Aune, R.M., Keyser, D., Colman, B., Mann, G., Vavra, J., 2011. The Real-Time Mesoscale Analysis at NOAA's National Centers for Environmental Prediction: Current Status and Development. Weather and Forecasting 26, 593–612. https://doi.org/10.1175/WAF-D-10-05037.1

Deutsch, C., 1997. Direct assessment of local accuracy and precision. Geostatistics wollongong 96, 115–125.

Doherty, C.T., Johnson, L.F., Volk, J., Mauter, M.S., Bambach, N., McElrone, A.J., Alfieri, J.G., Hipps, L.E., Prueger, J.H., Castro, S.J., Alsina, M.M., Kustas, W.P., Melton, F.S., 2022. Effects of meteorological and land surface modeling uncertainty on errors in winegrape ET calculated with SIMS. Irrig Sci 40, 515–530. https://doi.org/10.1007/s00271-022-00808-9

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., Merkova, D., Nielsen, J.E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S.D., Sienkiewicz, M., Zhao, B., 2017. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). J. Climate 30, 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Gómez-Hernández, J.J., Journel, A.G., 1993. Joint Sequential Simulation of MultiGaussian Fields, in: Soares, A. (Ed.), Geostatistics Tróia '92, Quantitative Geology and Geostatistics. Springer Netherlands, Dordrecht, pp. 85–94. https://doi.org/10.1007/978-94-011-1739-5_8

Gómez-Hernández, J.J., Srivastava, R.M., 2021. One Step at a Time: The Origins of Sequential Simulation and Beyond. Math Geosci 53, 193–209. https://doi.org/10.1007/s11004-021-09926-0

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103, 3–26. https://doi.org/10.1016/S0016-7061(01)00067-2

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press. https://doi.org/10.1093/oso/9780195115383.001.0001

Gutjahr, A., Bullard, B., Hatch, S., 1997. General joint conditional simulations using a fast fourier transform method. Mathematical Geology 29, 361–389. https://doi.org/10.1007/BF02769641

Hanser, S.E., 2008. Elevation in the Western United States (90 meter DEM).

Hart, Q.J., Brugnach, M., Temesgen, B., Rueda, C., Ustin, S.L., Frame, K., 2009. Daily reference evapotranspiration for California using satellite imagery and weather station measurement interpolation. Civil Engineering and Environmental Systems 26, 19–33. https://doi.org/10.1080/10286600802003500

Hashimoto, H., Wang, W., Melton, F.S., Moreno, A.L., Ganguly, S., Michaelis, A.R., Nemani, R.R., 2019. High-resolution mapping of daily climate variables by aggregating multiple spatial data sets with the random forest algorithm over the conterminous United States. Intl Journal of Climatology 39, 2964–2983. https://doi.org/10.1002/joc.5995

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., De Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J., 2020. The ERA5 global reanalysis. Quart J Royal Meteoro Soc 146, 1999–2049. https://doi.org/10.1002/qj.3803

Hoffimann, J., 2018. GeoStats.jl – High-performance geostatistics in Julia. JOSS 3, 692. https://doi.org/10.21105/joss.00692

Hsiang, S.M., Burke, M., Miguel, E., 2013. Quantifying the Influence of Climate on Human Conflict. Science 341, 1235367. https://doi.org/10.1126/science.1235367

Hudson, G., Wackernagel, H., 1994. Mapping temperature using kriging with external drift: Theory and an example from scotland. Intl Journal of Climatology 14, 77–91. https://doi.org/10.1002/joc.3370140107

Kitanidis, P.K., 1997. Introduction to Geostatistics: Applications in Hydrogeology. Cambridge University Press.

Lobell, D.B., Thau, D., Seifert, C., Engle, E., Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sensing of Environment 164, 324–333. https://doi.org/10.1016/j.rse.2015.04.021

Menafoglio, A., Secchi, P., Dalla Rosa, M., 2013. A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. Electron. J. Statist. 7. https://doi.org/10.1214/13-EJS843

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., Houston, T.G., 2012. An Overview of the Global Historical Climatology Network-Daily Database. Journal of Atmospheric and Oceanic Technology 29, 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1

Morgan, M.G., Henrion, M., 1990. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press.

Olea, R.A., 1999. Geostatistics for Engineers and Earth Scientists. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4615-5001-3

Pyrcz, M., Deutsch, C.V., 2014. Pyrcz, M.J., and Deutsch, C.V., Geostatistical Reservoir Modeling, 2nd Edition, Oxford University Press, New York, p. 448.

Rasmussen, R.M., Chen, F., Liu, C.H., Ikeda, K., Prein, A., Kim, J., Schneider, T., Dai, A., Gochis, D., Dugger, A., Zhang, Y., Jaye, A., Dudhia, J., He, C., Harrold, M., Xue, L., Chen, S., Newman, A., Dougherty, E., Abolafia-Rosenzweig, R., Lybarger, N.D., Viger, R., Lesmes, D., Skalak, K., Brakebill, J., Cline, D., Dunne, K., Rasmussen, K., Miguez-Macho, G., 2023. CONUS404: The NCAR–USGS 4-km Long-Term Regional Hydroclimate Reanalysis over the CONUS. Bulletin of the American Meteorological Society 104, E1382–E1408. https://doi.org/10.1175/BAMS-D-21-0326.1

Rossi, M.E., Deutsch, C.V., 2014. Mineral Resource Estimation. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-1-4020-5717-5

Thornton, M.M., Shrestha, R., Wei, Y., Thornton, P.E., Kao, S.-C., Wilson, B.E., 2022. Daymet: Station-Level Inputs and Cross-Validation for North America, Version 4 R1. https://doi.org/10.3334/ORNLDAAC/2132

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

Thornton, P.E., Shrestha, R., Thornton, M., Kao, S.-C., Wei, Y., Wilson, B.E., 2021. Gridded daily weather data for North America with comprehensive uncertainty quantification. Sci Data 8, 190. https://doi.org/10.1038/s41597-021-00973-0

Volk, J.M., Huntington, J.L., Melton, F.S., Allen, R., Anderson, M., Fisher, J.B., Kilic, A., Ruhoff, A., Senay, G.B., Minor, B., Morton, C., Ott, T., Johnson, L., Comini De Andrade, B., Carrara, W., Doherty, C.T., Dunkerly, C., Friedrichs, M., Guzman, A., Hain, C., Halverson, G., Kang, Y., Knipper, K., Laipelt, L., Ortega-Salazar, S., Pearson, C., Parrish, G.E.L., Purdy, A., ReVelle, P., Wang, T., Yang, Y., 2024. Assessing the accuracy of OpenET satellite-based evapotranspiration data to support water resource and land management applications. Nat Water 2, 193–205. https://doi.org/10.1038/s44221-023-00181-7

White, M.A., Thornton, P.E., Running, S.W., 1997. A continental phenology model for monitoring vegetation responses to interannual climatic variability. Global Biogeochemical Cycles 11, 217–234. https://doi.org/10.1029/97GB00330

Zeng, J., Matsunaga, T., Tan, Z.-H., Saigusa, N., Shirai, T., Tang, Y., Peng, S., Fukuda, Y., 2020. Global terrestrial carbon fluxes of 1999–2019 estimated by upscaling eddy covariance data with a random forest. Sci Data 7, 313. https://doi.org/10.1038/s41597-020-00653-5