Allan C Just[1], Kodi B Arfer[1], Johnathan Rush[1], and Itai Kloog[1,2]

[1]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai
[2]The Department of Geography and Environmental Development, Ben-Gurion University of the Negev

February 9, 2023

Abstract

The challenge of reconstructing air temperature for environmental applications is to accurately estimate past exposures even where monitoring is sparse. We present XGBoost-IDW Synthesis for air temperature (XIS-Temperature), a high-resolution machine-learning model for daily minimum, mean, and maximum air temperature, covering the contiguous US from 2003 through 2021. XIS uses remote sensing (land surface temperature and vegetation) along with a parsimonious set of additional predictors to make predictions at arbitrary points, allowing the estimation of address-level exposures. We built XIS with a computationally tractable workflow for extensibility to future years, and we used weighted evaluation to fairly assess performance in sparsely monitored regions. The weighted root mean square error (RMSE) of predictions in site-level cross-validation for 2021 was 1.89 K for the minimum daily temperature, 1.27 K for the mean, and 1.72 K for the maximum. We obtained higher RMSEs in earlier years with fewer ground monitors. Comparing to three leading gridded temperature models in 2021 at thousands of private weather stations not used in model training, XIS had at most 49% of the mean square error for the minimum temperature and 87% for the maximum. In a national application, we report a stronger relationship between minimum temperature in a heatwave and social vulnerability with XIS than with the other models. Thus, XIS-Temperature has potential for reconstructing important environmental exposures, and its predictions have applications in environmental justice and human health.

Posted on 9 Feb 2023 — CC-BY 4.0 — https://doi.org/10.22541/essoar.167591086.68441300/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

XIS-Temperature: A daily spatiotemporal machine-learning model for air temperature in the contiguous United States

Allan C. Just[1*], Kodi B. Arfer[1], Johnathan Rush[1], Itai Kloog[1,2]

[1]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[2]The Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer Sheva, Israel

Corresponding Author: allan.just@mssm.edu
Address: Allan Just, One Gustave L. Levy Place, Box 1057, New York, NY 10029 USA

Abstract

The challenge of reconstructing air temperature for environmental applications is to accurately estimate past exposures even where monitoring is sparse. We present XGBoost-IDW Synthesis for air temperature (XIS-Temperature), a high-resolution machine-learning model for daily minimum, mean, and maximum air temperature, covering the contiguous US from 2003 through 2021. XIS uses remote sensing (land surface temperature and vegetation) along with a parsimonious set of additional predictors to make predictions at arbitrary points, allowing the estimation of address-level exposures. We built XIS with a computationally tractable workflow for extensibility to future years, and we used weighted evaluation to fairly assess performance in sparsely monitored regions. The weighted root mean square error (RMSE) of predictions in site-level cross-validation for 2021 was 1.89 K for the minimum daily temperature, 1.27 K for the mean, and 1.72 K for the maximum. We obtained higher RMSEs in earlier years with fewer ground monitors. Comparing to three leading gridded temperature models in 2021 at thousands of private weather stations not used in model training, XIS had at most 49% of the mean square error for the minimum temperature and 87% for the maximum. In a national application, we report a stronger relationship between minimum temperature in a heatwave and social

vulnerability with XIS than with the other models. Thus, XIS-Temperature has potential for reconstructing important environmental exposures, and its predictions have applications in environmental justice and human health.

Keywords

- XGBoost
- exposure assessment
- land surface temperature
- climate and health
- temperature and social vulnerability

Synopsis

Improved estimates of air temperature across the United States will improve future analyses on the health impacts of temperature and exposure disparities.

Introduction

Reconstructions of outdoor air temperature are an important exposure-assessment tool in characterizing the effect of extreme weather on human health. Epidemiological studies and health-impact assessments rely on accurate exposure modeling, and many people do not live close to weather stations. Large populations within a metropolitan area may be assigned the temperature from the nearest weather station (e.g., an outlying airport), yet temperatures can vary substantially across the area, even block-to-block, due to factors such as varying land cover and urban heat islands. While there are a number of available temperature models, developed for various purposes, that are used in health studies, they vary in accuracy and resolution.

Gridded temperature estimates are often built from numerical weather models and assimilation systems,[1] or from hybrid approaches that downscale these models to a higher

resolution.[2] Sophisticated interpolation approaches for weather monitors can account for elevation with digital elevation models (DEMs),[3] but they may not capture temperature variation driven by hyper-local land-use differences, such as those that occur within urban heat archipelagos, which may also be underrepresented within long-term climate-monitoring networks. Satellite remote sensing offers important predictors for land-use regression of air temperature, ranging from land-cover classifications to vegetation indices. The Moderate Resolution Imaging Spectroradiometer (MODIS) sensor on NASA's Terra and Aqua satellites offer daily thermal infrared-derived land surface temperature (LST). These LST products cover the top few millimeters of the earth's surface at a 1-km resolution. Recent reprocessing of MODIS data and advancements in the LST retrieval algorithms have reduced geolocation error and improved sensor calibration.[4] Although the relation between LST and air temperature is complex, we and many others have integrated LST into geostatistical models trained with air-temperature monitors.[5–7] In a recent model comparison that reconstructed air temperature in the Northeastern US at 1 km of resolution, we found that a machine-learning approach based on gradient boosting outperformed several other approaches, including generalized additive mixed models with spatial smoothing.[8] Machine learning is increasingly used to integrate remotely sensed predictors for higher-resolution predictions, but it is computationally demanding. Machine learning also needs reproducible data-ingestion pipelines to be extensible and to remain as up-to-date as the popular interpolation models.[3]

Gridded models are subject to a tradeoff between spatial resolution and computational demands as the resulting datasets expand. But even 1-km grid cells can fail to capture temperature gradients that are important for human health. In this study, we extend our prior machine-learning framework[8] and switch to a point-based model that incorporates both

rasterized and continuous fields. With a point-based approach, we can make daily predictions anywhere in the contiguous United States, such as at exact locations for geocoded addresses. We call this model XGBoost-IDW Synthesis (XIS) and build a reusable and extensible data pipeline to generate our daily XIS-Temperature predictions for 2003 through 2021; in a companion paper,[9] we use the same approach for modeling fine particulate air pollution ($PM_{2.5}$). Popular gridded models report only daily minimum and maximum temperature because they rely on interpolation of observed extrema. With large quality-controlled time-resolved observation series, one can construct accurate daily mean temperatures, without the assumption of diurnal symmetry (and consequent bias) that is inherent in averaging daily minima and maxima together.[10] We fit separate models for the daily minimum, mean, and maximum temperature, because all three variables are relevant in applications, including epidemiology.

We present detailed performance metrics for XIS-Temperature using a site-level cross-validation across the contiguous US with stratification by year, season, and NOAA climate region.[11,12] Because weather stations are found more often in densely populated areas, we use weights to appropriately quantify performance across the study region, including suburban and rural areas.[8] It is often difficult to tell which particular weather stations have been used in training large models, raising the threat of data leakage in model comparison. We consider three gridded models popular for applied research in the US, and compare them to XIS on thousands of private weather stations that were not used for training any of the models. The comparison models and their resolutions are PRISM (4 km),[13], gridMET (4 km),[14] and Daymet (1 km).[3] Finally, to demonstrate the model-dependent interpretation of temperature exposures and to show implications for environmental justice, we show the relation of a peak summer temperature from

XIS (versus the same gridded models) with tract-level social vulnerability[15] across the contiguous US.

Method

Study area and time period

XIS-Temperature covers the same area and time period as XIS-$PM_{2.5}$,[9] namely the contiguous US (excluding large water bodies) for 2003 through 2021. Like XIS-$PM_{2.5}$, XIS-Temperature represents space as floating-point longitude-and-latitude pairs and represents days as midnight-to-midnight intervals of Central Standard Time (UTC−6).

Data

Temperature

A key input for geostatistical models of environmental conditions is the set of observations used for training. We separately modeled three metrics of daily temperature as dependent variables (DVs): minimum (hereinafter "min") temperature, mean temperature, and maximum (hereinafter "max") temperature. We used two sources of temperature data: 1. the Meteorological Assimilation Data Ingest System (MADIS),[16] maintained by the National Oceanic and Atmospheric Administration (NOAA), from which we ingested the National Mesonet and COOP datasets available to registered research organizations, and 2. Weather Underground, a private commercial network of personal weather stations, which we have used previously.[8] For MADIS, we started with individual observations timestamped to the second, whereas for Weather Underground, we used precomputed daily means and extrema. We filtered and quality-checked the data per year and source as follows:

1. Drop station-times with a missing temperature, time, longitude, or latitude.

2. (MADIS only) Keep only station-times passing at least MADIS quality-control stages 1 and 2 checks for validity and consistency (`temperatureDD` equal to `S`, `V`, `K`, or `k`).
3. To handle instances where nearby stations might be duplicates, group stations into clusters in which no two stations are more than 50 m apart. In each cluster, keep only the station with the most common station identifier. Identify these clusters as stations henceforth, using the lexicographically first location as the location for the cluster.
4. Drop stations outside the study area.
5. Remove rows with observations that are beyond NOAA's record historical extrema for the region.[17]
6. Among observations that are equal (or very close) to 0 °F or 0 °C, try to distinguish which are real measurements and which represent missing values. We do this by dropping any such "zero observations" with no other observation at the same station within 5 days that is both nonzero and within 3 K of the zero observation.
7. Drop to one observation per station-time, preferring observations that appear earlier in the input.
8. (MADIS only) Ensure that each station-day covers at least 18 distinct hours in UTC−6, then aggregate into days. Compute the min as simply the minimum observation on each date, and likewise for the max. Compute the mean with all observations on the date, weighted according to the number of seconds in the date to which each observation is closest. (Note that in general, the daily Weather Underground values have been computed differently, including a different time zone.)
9. Remove daily observations that are part of a run of equal values, spanning more than 3 consecutive nonmissing station-days, for any of min, mean, or max temperature.

10. For spatial consistency, compare observations that are within 100 km of two other observations. If these neighbors have an elevation difference from the original observation no greater than 500 m, and both differ from the original observation by more than 20 K, drop the original observation. Run this check separately for each DV, but drop the entire row (i.e., all DVs) if an observation fails on any of them.
11. Drop stations with less than 30 days of observations.

Thanks to the inclusion of Weather Underground, the size of the entire temperature dataset for each year could be computationally burdensome; for example, in 2020, there were 35,825,729 observations of each DV from 117,276 stations. We suspected that with a subset of the data, we could obtain similar performance as with all of it. We opted to prioritize Weather Underground stations that cover the most area not covered by MADIS. To determine appropriate subsets, we conducted a learning-curve analysis on the mean-temperature DV in 30 random days of 2018. We held out a random fifth of stations and computed per-station weights by summing observation weights (computed with Voronoi diagrams of the study region using the stations available each day, as described previously[9]). The process started with all the MADIS stations and none of the Weather Underground stations, then added the next 2,500 remaining Weather Underground stations of highest weight (representing the largest areas without monitors in the study region) at each subsequent step. At each step, we used IDW to predict observations at the held-out stations using the training stations, then computed the weighted root mean square error (RMSE). We found that the earliest step with a weighted RMSE within 0.01 K of the best was step 5, with 1.81 K. The minimum weight of the Weather Underground stations used in this step was 2,636 km^2. Thus, for our real models, we used only Weather Underground stations with a weight of at least 2,636 km^2 times $n/30$, where n is the number of days in the year being analyzed

(365 or 366) and 30 is the number of days in the learning-curve analysis. Ultimately, the per-year rate at which observations in our analytic dataset came from Weather Underground ranged from 10% to 40%.

Predictors

We used the following 13 variables as predictors.

- Longitude and latitude
- The integer day of the year
- An IDW feature, which is an interpolation of the relevant temperature metric (min, mean, or max) at sites within 100 km, weighted by the distance (thus, the IDW exponent is 1)
- Two overpasses per day of Aqua LST,[18] one during the daytime and one at night, represented in kelvins
- Monthly vegetation, quantified as the enhanced vegetation index from Aqua[19]
- Two variables for surface imperviousness (from the National Land Cover Database[20]): one for the imperviousness at a single 30-m grid cell and one for the Gaussian-filtered imperviousness in a 1-km square around the query point[9]
- Population density, from the Gridded Population of the World[21]
- Elevation, from the US Geological Survey's 3D Elevation Program[22]
- Hilliness, or local relative topography, quantified as the multi-scale topographic dissection index computed from elevation[6]
- Distance from water, in kilometers

Given the goal of sharing an efficient geospatial data-processing workflow, we reused variable construction with XIS-$PM_{2.5}$ for the majority of predictors.[9]

We computed distance from water using a global coastline shapefile from OpenStreetMap and Great Lakes shapefiles by the US Geological Survey, including Lake St. Clair. Other bodies of water were not considered. Distances were capped at 500 km so that our model did not use this variable as an index of far-inland locations in place of the longitude and latitude features.

Models

The core modeling approach used extreme gradient boosting (XGBoost) and IDW as in XIS-$PM_{2.5}$ with station-level cross validation.[9] We conducted tuning as in XIS-$PM_{2.5}$[9] separately for each of the three DVs, resulting in a separate hyperparameter vector for each (Table 1).

Table 1: Selected hyperparameters for the three dependent variables.

dv	nrounds	max_depth	eta	gamma	lambda	alpha
temp.min	500	9	0.078	0.38	360	0.570
temp.mean	500	9	0.090	0.82	810	0.011
temp.max	500	9	0.061	0.22	250	0.080

Evaluation

We used station-wise cross-validation (CV) as for XIS-$PM_{2.5}$, but with 5 rather than 10 folds for speed, in the face of much larger datasets. The concerns that motivated the use of absolute-error metrics for XIS-$PM_{2.5}$ did not apply to the temperature data, so we gave XGBoost a square-error objection function, evaluated the models RMSE, and measured baseline variability with the standard deviation (SD). In order to account for the highly variable density of observations across the study region, we weighted observations by their spatial coverage with the same daily Voronoi-diagram method we used for XIS-$PM_{2.5}$. We calculated SHAP[23] for our cross-validated predictions to quantify feature contributions.

Results

Cross-validation

Table 2 shows weighted results for each year of CV. The bias of our predictions per year ranged from -0.004 to +0.048 K for min temperature, -0.044 to +0.019 K for mean temperature, and -0.087 to -0.022 K for max temperature. Table 3 shows per-region performance for a single year; Figure 5 in the Supplementary Information (SI) plots per-region performance for a single DV in every year. Table 6 in the SI shows unweighted performance at particularly isolated stations, as a demonstration of how the model performs in sparsely monitored regions. Finally, Tables 7 and 8 in the SI show unweighted CV results among station-days that are particularly hot or cold, which is of particular relevance for epidemiologic applications examining the health impacts of extreme weather and similar to an analysis for our previous temperature model.[8]

Table 2: Weighted SDs and RMSEs (K) from yearly CV.

Year	Observations	Sites	Min temp.		Mean temp.		Max temp.	
			SD	RMSE	SD	RMSE	SD	RMSE
2003	1,007,880	5,241	10.55	2.61	10.91	2.16	11.92	2.82
2004	1,435,459	6,426	10.26	2.26	10.49	1.85	11.35	2.41
2005	2,030,065	9,229	10.38	2.21	10.74	1.73	11.73	2.29
2006	2,694,789	10,869	10.05	2.23	10.38	1.69	11.33	2.18
2007	3,212,596	12,530	10.79	2.25	11.11	1.75	12.11	2.25
2008	3,620,854	14,090	10.82	1.99	11.06	1.52	11.96	1.96
2009	4,019,316	15,111	10.73	1.96	10.95	1.48	11.82	1.90
2010	4,428,537	16,604	10.69	1.92	11.09	1.43	12.14	1.88
2011	4,989,079	18,451	10.93	1.97	11.31	1.49	12.29	1.89
2012	6,546,637	24,811	10.13	1.87	10.46	1.34	11.43	1.72
2013	7,297,032	26,322	11.10	1.86	11.30	1.36	12.12	1.78
2014	7,884,056	27,970	11.10	1.84	11.22	1.35	12.06	1.75
2015	8,354,364	29,892	10.61	1.81	10.75	1.36	11.56	1.75
2016	9,003,820	30,839	10.37	1.85	10.64	1.42	11.55	1.79
2017	9,740,974	33,230	10.42	1.91	10.78	1.44	11.78	1.90
2018	9,999,329	34,869	11.17	1.90	11.40	1.42	12.26	1.94
2019	10,682,830	39,077	11.17	1.84	11.39	1.29	12.35	1.86
2020	11,681,725	39,947	10.46	1.90	10.65	1.28	11.61	1.76
2021	11,646,537	39,074	10.66	1.89	10.79	1.27	11.72	1.72

Table 3: Weighted SDs and RMSEs (K) for 2021 broken down by region.

Region	Observations	Sites	Min temp.		Mean temp.		Max temp.	
			SD	RMSE	SD	RMSE	SD	RMSE
Ohio Valley	1,395,207	4,550	9.77	1.33	9.84	0.98	10.74	1.49
Upper Midwest	981,677	3,131	11.82	1.52	12.02	0.95	12.96	1.45
Northeast	1,377,490	4,495	10.07	1.43	10.13	0.91	11.05	1.37
Northwest	939,829	3,135	8.10	2.24	9.31	1.48	11.19	1.88
South	1,556,537	5,243	9.52	1.45	9.08	1.09	9.54	1.63
Southeast	1,507,076	5,081	8.41	1.32	7.70	0.95	8.04	1.50
Southwest	1,143,976	3,967	9.93	2.47	10.34	1.64	11.10	2.03
West	1,917,297	6,678	9.54	2.61	10.07	1.73	11.03	2.11
Northern Rockies and Plains	827,448	2,794	11.05	1.97	11.89	1.29	13.25	1.72

To demonstrate the difference between mean temperature modeled directly and mean temperature represented as the average of min and max, we computed the weighted root mean square difference between our mean predictions and the average of our min and max predictions for 2021. The result was 0.77 K, comparable in magnitude to our RMSEs from CV.

Figure 1 shows the mean absolute SHAP of each feature for mean temperature in 2010. Although there is substantial variation by region, the largest contributions come from the IDW feature, elevation, longitude, and distance to water.

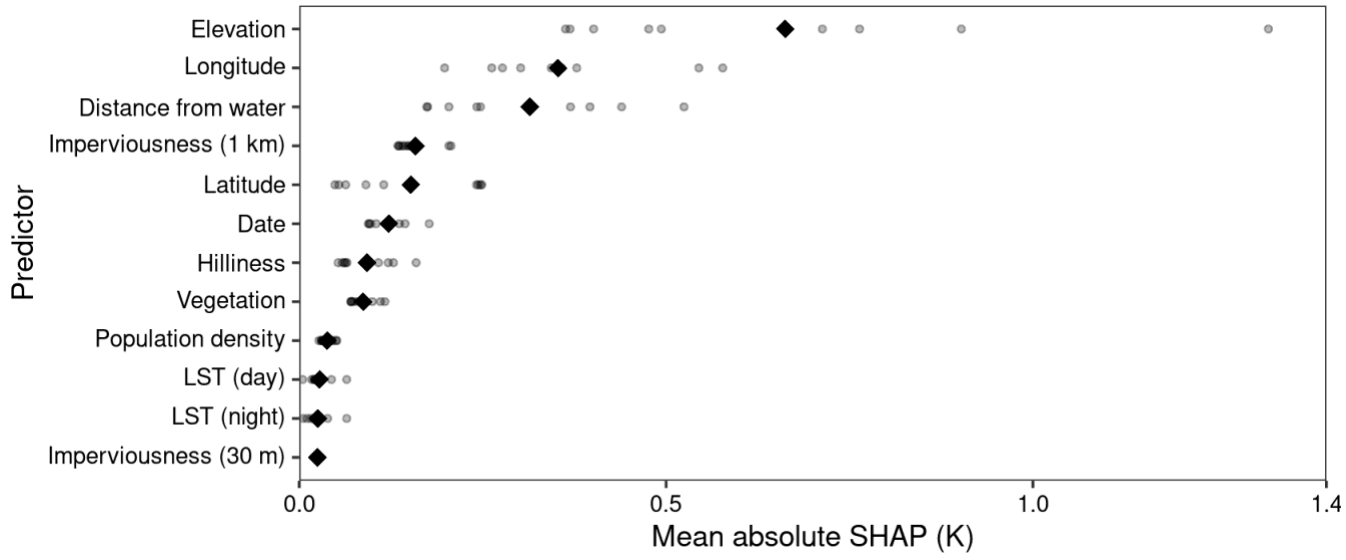

Figure 1: Mean absolute SHAP for mean temperature of each predictor in 2010 (the IDW feature, which has much greater absolute SHAP than everything else, is omitted). Small dots show per-region means. Diamonds show overall means.

Figure 2 shows one year of daily predictions and error (i.e., the difference from observations) for a single representative station.

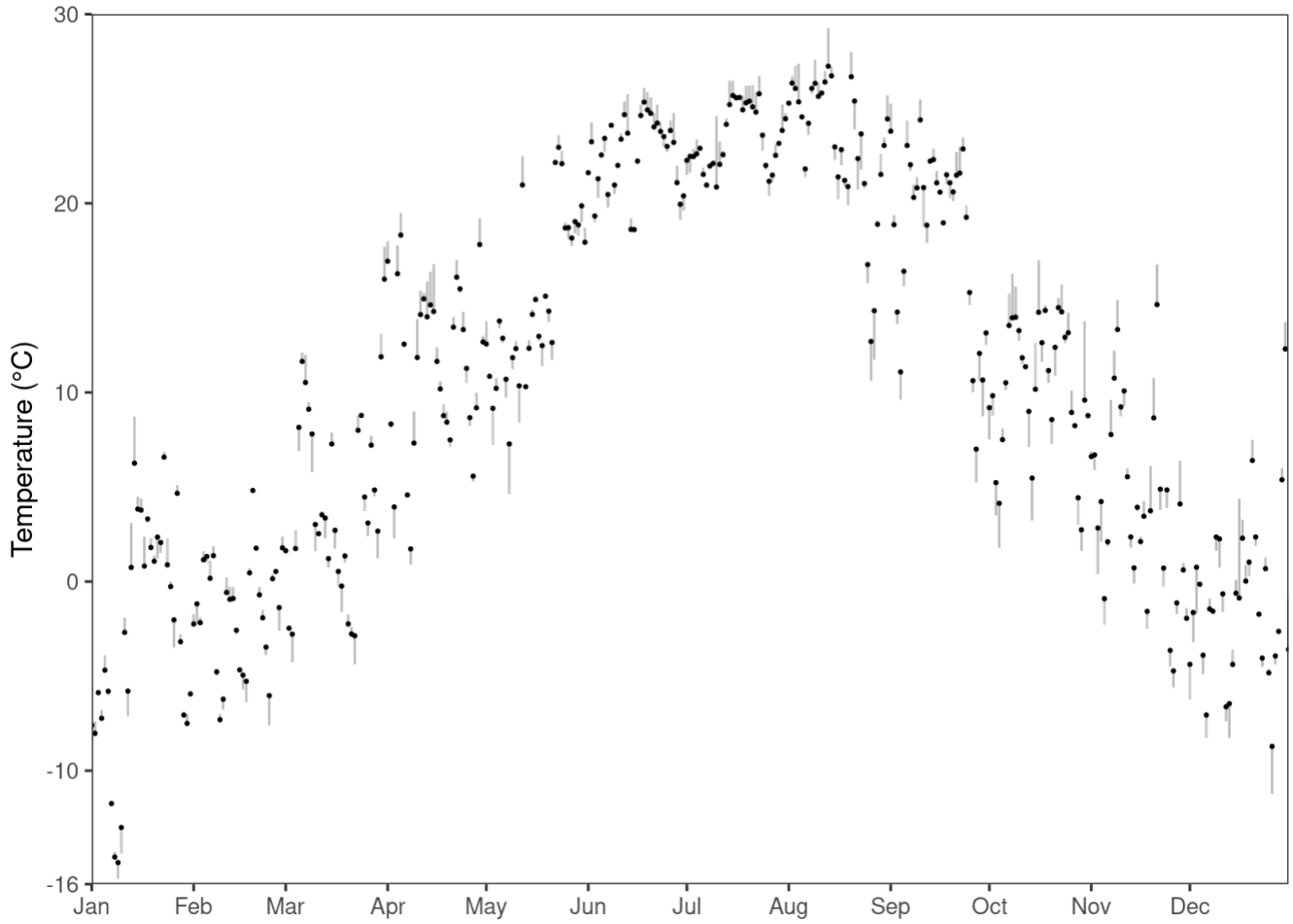

Figure 2: A plot of predicted min temperature from CV in 2010 (points), and the distance from the observed value (line segments), for a Weather Underground station near Oklahoma City. This station was selected to have the yearly per-station unweighted RMSE closest to the median among all stations that had an observation for every day in 2010. Its RMSE is 1.12 K.

New predictions

For the following plots and analyses, we fit XIS to all the training data we had for each year and made predictions for new point-days. Figure 3 maps predictions for the entire study area on the hottest day in 2021. Figure 4 shows predictions for the same day in the New York City area, with discernible fine-scale variation in temperature, such as cooler air in Central Park than in adjacent built-up areas within the island of Manhattan.

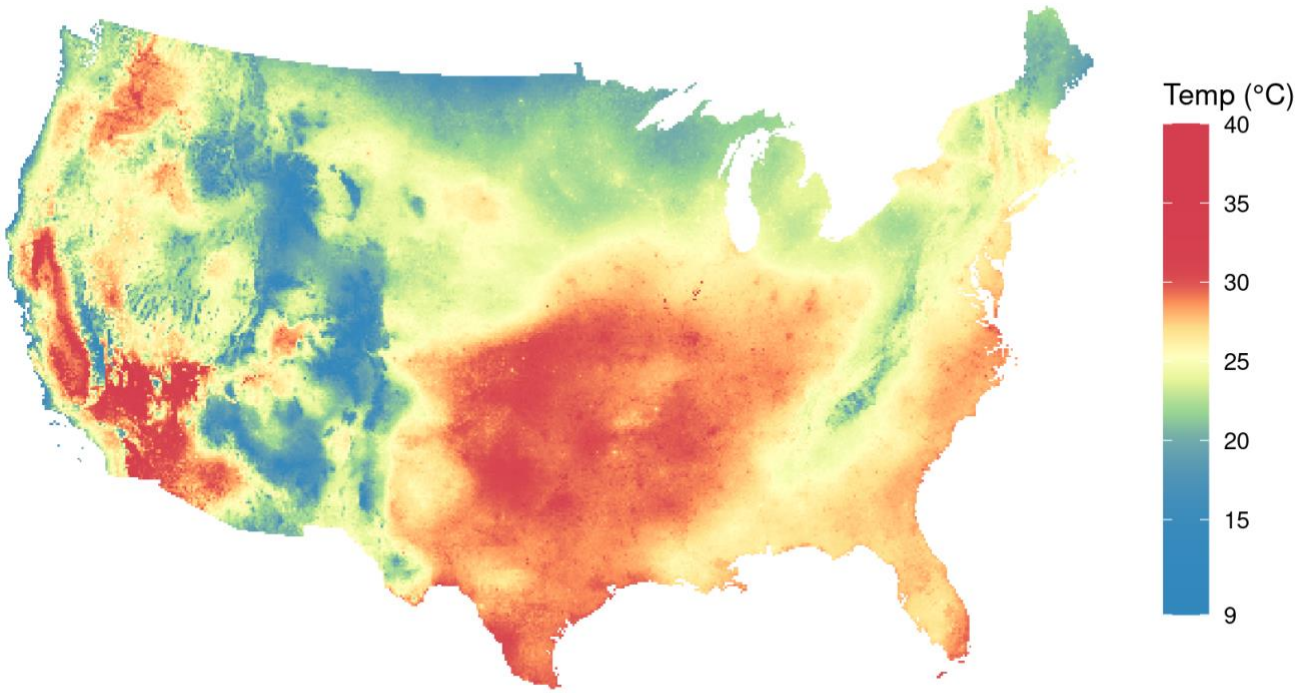

Figure 3: Predicted mean temperature for 11 Aug 2021 across the study area, shown in the US National Atlas projection. We chose this date for having the highest mean temperature in 2021 across all stations.

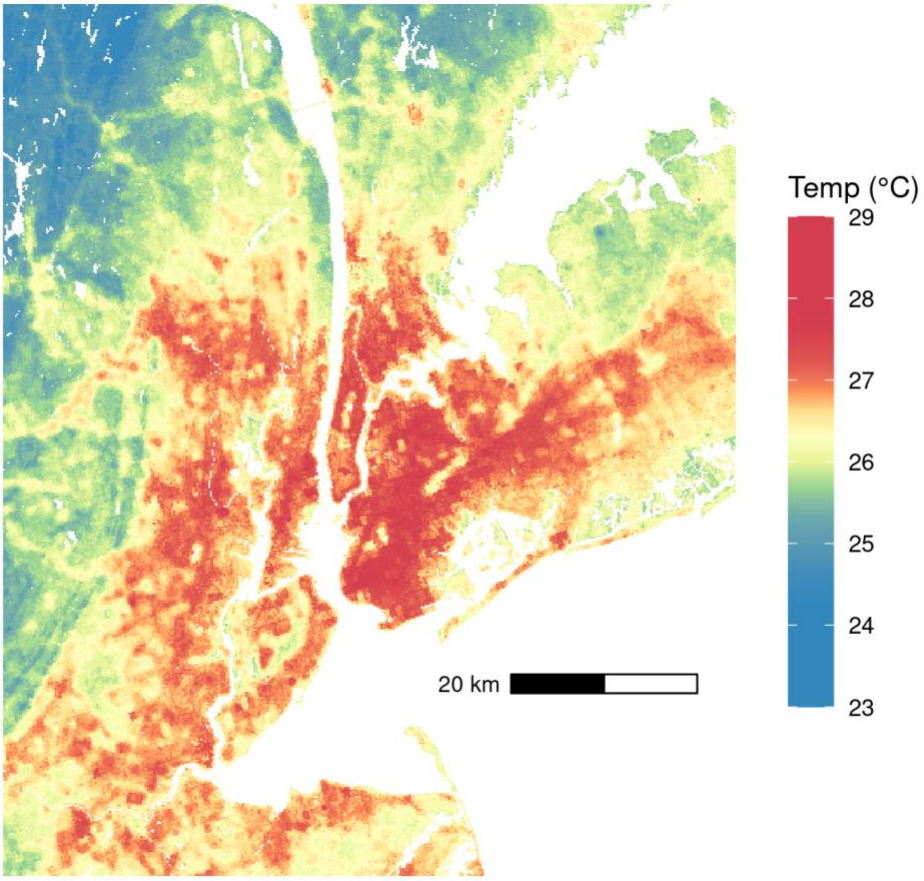

Figure 4: Predicted mean temperature for 11 Aug 2021 in the New York City area.

Comparison with other models

Tables 4 and 5 show RMSEs (stratified by year and then for seasons of 2021) of daily min temperature from our model and three gridded temperature products: PRISM, gridMET, and Daymet. Table 9 in the SI shows analogous results to Table 4 for max temperature. The models are tested on observations at Weather Underground stations that were not used for training XIS. For each year, we take a random sample of 10,000 such stations that lie in the intersection of all four modeling regions, so we only analyze years with at least this many stations available. We recompute weights for these observations with the same algorithm we used for the main CV. We

omit December 31st on leap years, since Daymet provides no predictions on these days. With averaging across years, our model has 28% of the MSE of PRISM, 34% of gridMET, and 46% of Daymet. Without weighting, these figures become 32% of PRISM, 35% of gridMET, and 52% of Daymet. Yearly weighted biases range from -0.94 to -0.74 K for PRISM, -0.91 to -0.72 K for gridMET, -0.88 to -0.57 K for Daymet, and -0.14 to +0.03 K for XIS. For max temperature, our results are relatively less impressive, because gridMET and Daymet are much improved over min temperature: XIS obtains 18% of the MSE of PRISM, 69% of gridMET, and 81% of Daymet. The yearly weighted biases range from -1.02 to -0.61 K for PRISM, -0.79 to -0.43 K for gridMET, -0.70 to -0.30 K for Daymet, and -0.46 to -0.13 K for XIS.

Table 4: Comparison of weighted RMSEs (K) of min temperature with PRISM, gridMET, and Daymet.

Year	Observations	SD	PRISM	gridMET	Daymet	XIS
2014	2,874,710	11.12	3.29	2.69	2.50	1.53
2015	2,749,242	10.60	3.15	2.66	2.47	1.61
2016	2,718,814	10.47	3.11	2.69	2.46	1.61
2017	2,807,208	10.46	3.30	3.44	2.61	1.83
2018	2,729,138	11.29	3.23	2.70	2.49	1.75
2019	2,702,679	11.25	3.23	2.96	2.50	1.78
2020	3,114,743	10.51	3.26	3.66	2.58	1.82
2021	3,034,309	10.68	3.15	2.96	2.51	1.75

Table 5: Weighted RMSEs (K) of min temperature for the various models in 2021, broken down by season. We use December from 2020 instead of 2021 so as to analyze a contiguous winter. Thus the winter row includes the random samples of sites from two different years and has more distinct sites than the other seasons.

Season	Observations	Sites	SD	PRISM	gridMET	Daymet	XIS
Winter	777,358	16,919	8.59	3.66	3.36	2.78	1.90
Spring	702,355	9,700	7.91	3.19	2.96	2.56	1.80
Summer	791,595	9,398	5.29	2.20	2.42	2.07	1.57
Fall	751,520	9,062	8.14	3.14	2.99	2.53	1.75

To examine how XIS's higher spatial resolution contributed to its improved performance, we also tried making XIS predictions for the 2021 test observations using the centroids of

Daymet's 1-km grid cells instead of the true locations. The result was an unweighted RMSE for min temperature of 1.70 K, compared to 1.68 K for using the true locations and 2.24 K for Daymet.

Model application to social vulnerability

We examined how minimum temperature on 17 Jul 2010, the day of 2010 with the highest mean of min temperatures across all stations, related to the social vulnerability index in 2010.[15] We fit a mixed-effects linear regression model where the unit of analysis was the 71,712 US Census tracts in our study area and the dependent variable was the minimum temperature at the center of population of each tract. The model had a fixed effect for vulnerability, per-county random slopes of vulnerability, and per-county random intercepts (with the slopes and intercepts modeled as correlated). The fixed effect of vulnerability was estimated as 0.69 K ([0.65, 0.74]), where the latter is a 95% CI, meaning that a change from minimum to maximum vulnerability was associated with a 0.69-K higher minimum temperature on this day.

We fit similar mixed models with temperature estimates from the gridded temperature products to which we compared XIS earlier, and obtained substantially smaller estimates for this effect: 0.20 K ([0.15, 0.25]) for PRISM, 0.26 K ([0.21, 0.30]) for gridMET, and 0.16 K ([0.13, 0.20]) for Daymet.

Discussion

We present a daily spatiotemporal air temperature model for the contiguous US that covers 19 years. Our model, XIS-Temperature, builds on a large time-resolved dataset of ground observations, NOAA's MADIS, and is augmented with observations from private weather stations in more sparsely monitored areas. As expected, our model shows substantial accuracy,

which increases in more recent years, since the number of observations available increases tenfold from 2003 to 2021.

We compared XIS predictions for min and max temperatures with three leading gridded models at 10,000 private weather stations not used in our model training, reweighted spatially to increase representativeness for the full study region. We have substantially lower RMSE than all three competitors in every year of the comparison. When we further stratified our model comparison by season in 2021, XIS had the least RMSE for each season, as well as the least variability in RMSE across seasons. A sensitivity analyses generating XIS predictions at the same centroids used by Daymet's 1-km grid (as opposed to exact locations of weather stations) showed that our improved accuracy is not explained by differences in resolution. Overall, testing on a large network of private weather stations demonstrates that using XIS-Temperature obtains lower exposure measurement error overall, as well as lower seasonal variation in the error.

We fit separate models for min, mean, and max temperature because all three DVs have useful applications in estimating impacts of temperature. Our primary data source, MADIS, provides time-resolved air-temperature data; thus, we did not need to rely on the inexact date-shifting used by other models.[3,6] We calculated a daily time-weighted mean temperature for MADIS data, and trained a separate model for mean temperature, to avoid the assumption of diurnal symmetry; that is, the assumption that the daily mean is reasonably approximated by the mean of the daily extrema.[10] Given the inherent difficulty in estimating extrema, as well as the higher SD we observed for max temperature compared to mean and min, it is not surprising that our mean-temperature models have lower RMSE than our extrema models.

As a demonstration of the application of the XIS model to social vulnerability, we constructed a national multi-level regression for the relation of tract-level minimum temperature

with social vulnerability, nested within counties, on the hottest day in 2010. Comparing the most vulnerable to the least vulnerable tracts, we saw a substantially larger difference in temperature when using XIS than when using any of the competing models. Differences in overall accuracy are the most likely explanation for these model-dependent findings, although we also highlight the advantage of our point-based model to resolve stark disparities in temperature between nearby neighborhoods. Our application shows the model-dependent interpretation of the complex relation between temperature and vulnerability for this one day; a more thorough evaluation of temperature disparities across time is ongoing.

The limitations of our model include temporal coverage bounded by our inclusion of data from NASA's Aqua satellite. XIS-Temperature only goes back as far as 2003, whereas Daymet goes back to 1980. Furthermore, because we fit our model annually and incorporated new stations as they came online, (improving our accuracy for later years), our model may not be well suited for studying long-term climate change. Our 2021 model performance is worst in the West and Southwest regions, which may be related to more complex topoclimatic relations. Future inclusion of predictors related to snow cover may help in those regions, particularly in winter and spring, which were the hardest seasons to predict for XIS as well as for the competing models. Our SHAP analysis suggests that the LST variables contribute little to predictions, although we had expected them to contribute in complex terrain, particularly for min temperature.[24] Future XIS development could adopt the approach of constructing measures of monthly relative LST variation over local windows[6] to identify 1-km pixels that are hotter or colder than nearby pixels, rather than directly including the daily (and often missing) LST values.

The parsimony and automation of XIS-Temperature enable further development, refinement, and the inclusion of new predictors. Thus we expect further improvement as we

extend XIS into the future. Not only have we demonstrated better predictive accuracy and smaller bias than three leading gridded models, assessed at a large network of private weather stations, but we have shown a strong model-dependent relation of extreme heat and social vulnerability, highlighting the importance of using improved exposure models such as XIS-Temperature in health-impacts analyses.

Acknowledgments

Research reported in this publication was supported by the Environmental influences on Child Health Outcomes (ECHO) program, Office of The Director, National Institutes of Health, under Award Numbers U2C OD023375, U24 OD023382, U24 OD023319, UH3 OD023337, and an ECHO Opportunities and Infrastructure Fund award to ACJ, as well as National Institutes of Health grants R01 ES031295, R01 DK127139, P30 ES023515, and UL1 TR004419.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

(1) NASA. *NLDAS-2 Forcing Dataset Information*. https://ldas.gsfc.nasa.gov/nldas/v2/forcing (accessed 2022-06-23).
(2) Crosson, W. L.; Al-Hamdan, M. Z.; Insaf, T. Z. Downscaling NLDAS-2 Daily Maximum Air Temperatures Using MODIS Land Surface Temperatures. *PLOS ONE* **2020**, *15* (1), e0227480. https://doi.org/10.1371/journal.pone.0227480.
(3) Thornton, P. E.; Shrestha, R.; Thornton, M.; Kao, S.-C.; Wei, Y.; Wilson, B. E. Gridded Daily Weather Data for North America with Comprehensive Uncertainty Quantification. *Sci Data* **2021**, *8* (1), 190. https://doi.org/10.1038/s41597-021-00973-0.
(4) Hulley, Glynn. MODIS/Aqua Land Surface Temperature/3-Band Emissivity Daily L3 Global 1km SIN Grid Day V061, 2021. https://doi.org/10.5067/MODIS/MYD21A1D.061.
(5) Kloog, I.; Chudnovsky, A.; Koutrakis, P.; Schwartz, J. Temporal and Spatial Assessments of Minimum Air Temperature Using Satellite Surface Temperature Measurements in Massachusetts, USA. *Science of The Total Environment* **2012**, *432*, 85–92. https://doi.org/10.1016/j.scitotenv.2012.05.095.
(6) Oyler, J. W.; Ballantyne, A.; Jencso, K.; Sweet, M.; Running, S. W. Creating a Topoclimatic Daily Air Temperature Dataset for the Conterminous United States Using Homogenized Station Data and Remotely Sensed Land Skin Temperature. *Int. J. Climatol* **2015**, *35* (9), 2258–2279. https://doi.org/10.1002/joc.4127.

(7) Gutiérrez-Avila, I.; Arfer, K. B.; Wong, S.; Rush, J.; Kloog, I.; Just, A. C. A Spatiotemporal Reconstruction of Daily Ambient Temperature Using Satellite Data in the Megalopolis of Central Mexico from 2003 to 2019. *Int J Climatol* **2021**, *41* (8), 4095–4111. https://doi.org/10.1002/joc.7060.
(8) Carrión, D.; Arfer, K. B.; Rush, J.; Dorman, M.; Rowland, S. T.; Kioumourtzoglou, M.-A.; Kloog, I.; Just, A. C. A 1-Km Hourly Air-Temperature Model for 13 Northeastern U.S. States Using Remotely Sensed and Ground-Based Measurements. *Environ. Res.* **2021**, *200*, 111477. https://doi.org/10.1016/j.envres.2021.111477.
(9) Just, A.; Arfer, K.; Rush, J.; Lyapustin, A.; Kloog, I. XIS-Pm2.5: A Daily Spatiotemporal Machine-Learning Model for Pm2.5 in the Contiguous United States. *Earth and Space Science Open Archive* **2022**, 26. https://doi.org/10.1002/essoar.10512861.1.
(10) Bernhardt, J.; Carleton, A. M.; LaMagna, C. A Comparison of Daily Temperature-Averaging Methods: Spatial Variability and Recent Change for the CONUS. *A Comparison of Daily Temperature-Averaging Methods* **2018**, *31* (3), 979–996. https://doi.org/10.1175/JCLI-D-17-0089.1.
(11) NOAA. *U.S. Climate Regions | Monitoring References | National Centers for Environmental Information (NCEI)*. https://www.ncdc.noaa.gov/monitoring-references/maps/us-climate-regions.php (accessed 2020-10-07).
(12) Karl, T. R.; Koscielny, A. J. Drought in the United States: 1895–1981. *Journal of Climatology* **1982**, *2* (4), 313–329. https://doi.org/10.1002/joc.3370020402.
(13) PRISM Climate Group, Oregon State U. https://www.prism.oregonstate.edu/ (accessed 2022-11-03).
(14) Abatzoglou, J. T. Development of Gridded Surface Meteorological Data for Ecological Applications and Modelling. *International Journal of Climatology* **2013**, *33* (1), 121–131. https://doi.org/10.1002/joc.3413.
(15) Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry/ Geospatial Research, Analysis, and Services Program. *CDC/ATSDR Social Vulnerability Index (SVI) 2018 Database US*. https://www.atsdr.cdc.gov/placeandhealth/svi/index.html (accessed 2022-06-17).
(16) *NCEP Meteorological Assimilation Data Ingest System (MADIS)*. https://madis.ncep.noaa.gov/ (accessed 2022-09-14).
(17) State Climate Extremes Committee (SCEC) | National Centers for Environmental Information (NCEI). *Records*. http://web.archive.org/web/20220812134705id_/https://www.ncei.noaa.gov/access/monitoring/scec/records.csv (accessed 2022-09-15).
(18) Hulley, Glynn. MODIS/Aqua Land Surface Temperature/3-Band Emissivity Daily L3 Global 1km SIN Grid Day V061, 2021. https://doi.org/10.5067/MODIS/MYD21A1D.061.
(19) Didan, Kamel. MODIS/Aqua Vegetation Indices Monthly L3 Global 1km SIN Grid V061, 2021. https://doi.org/10.5067/MODIS/MYD13A3.061.
(20) Dewitz, J. National Land Cover Database (NLCD) 2019 Products, 2021. https://doi.org/10.5066/P9KZCM54.
(21) Center For International Earth Science Information Network-CIESIN-Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 11, 2018. https://doi.org/10.7927/H4JW8BX5.
(22) US Geological Survey. *1 Arc-second Digital Elevation Models (DEMs) - USGS National Map 3DEP Downloadable Data Collection*.

https://www.sciencebase.gov/catalog/item/4f70aa71e4b058caae3f8de1 (accessed 2022-06-15).
(23) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* **2020**, *2* (1), 56–67. https://doi.org/10.1038/s42256-019-0138-9.
(24) Oyler, J. W.; Dobrowski, S. Z.; Holden, Z. A.; Running, S. W. Remotely Sensed Land Skin Temperature as a Spatial Predictor of Air Temperature Across the Conterminous United States. *Journal of Applied Meteorology and Climatology* **2016**, *55* (7), 1441–1457. https://doi.org/10.1175/JAMC-D-15-0276.1.