


Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Environmental Research

journal homepage: www.elsevier.com/locate/envres

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

Allan C. Just^{a,b,c,*} , Kodi B. Arfer^{a,c}, Johnathan Rush^c, Itai Kloog^{c,d}

^a Department of Epidemiology, Brown University School of Public Health, Providence, RI, USA

^b Institute at Brown for Environment and Society, Brown University, Providence, RI, USA

^c Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

^d The Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer Sheva, Israel

ARTICLE INFO

Keywords:

XGBoost

Exposure assessment

Land surface temperature

Climate and health

Temperature and social vulnerability

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 2023. 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 2023 was 1.78 K for the minimum daily temperature, 1.19 K for the mean, and 1.48 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 60% of the mean square error for the minimum temperature and 93% for the maximum. In a national application, we report a stronger relationship between summertime minimum temperature 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.

1. 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 monitoring 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 (Tuholske et al., 2021; Yu et al., 2024). 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 (NASA, 2022), or from hybrid

approaches that downscale these models to a higher resolution (Crosson et al., 2020). An intercomparison of classical modeling and emergent machine-learning approaches to statistical downscaling of global climate models found that complex approaches may offer few benefits without careful refinement (Vandal et al., 2019). Sophisticated interpolation approaches for weather monitors can account for elevation with digital elevation models (DEMs; Thornton et al., 2021), 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

* Corresponding author. Department of Epidemiology, Brown University School of Public Health, Providence, RI, USA.

E-mail address: allan_just@brown.edu (A.C. Just).

<https://doi.org/10.1016/j.envres.2024.120731>

Received 4 October 2024; Received in revised form 20 December 2024; Accepted 28 December 2024

Available online 12 January 2025

0013-9351/© 2024 Published by Elsevier Inc.

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 (Hulley, 2021). 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 (Gutiérrez-Avila et al., 2021; Kloog et al., 2012; Oylar et al., 2015). 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 (Carrión et al., 2021). 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 (Thornton et al., 2021).

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 (Carrión et al., 2021) 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 2023; in a companion paper (Just et al., 2024), 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 (Bernhardt et al., 2018). 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 (Karl and Koscielny, 1982; NOAA, 2013). 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 (Carrión et al., 2021). 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. Finally, to demonstrate the model-dependent interpretation of temperature exposures and to show implications for environmental justice, we show the relation of a summer temperature from XIS (versus the same gridded models) with tract-level social vulnerability (Centers for Disease Control and Prevention, 2018) across the contiguous US.

2. Method

2.1. Study area and time period

XIS-Temperature covers the same area and time period as XIS-PM_{2.5} (Just et al., 2024), namely the contiguous US (excluding large water bodies) for 2003 through 2023. 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).

3. Data

3.1. 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. Our primary source of temperature data was the Meteorological Assimilation Data Ingest System (MADIS; Miller et al., 2005), maintained by the National Oceanic and Atmospheric Administration (NOAA), from which we ingested the National Mesonet and COOP datasets available to registered research organizations. We used an additional data source for comparisons with other models: Weather Underground, a private commercial network of personal weather stations, which we have used previously (Carrión et al., 2021). 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 (State Climate Extremes Committee, 2022).
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. Weighting ensures that in the case of unevenly sparse observations across a day, the mean does not overrepresent times of day with particularly dense observations. (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.

3.2. Predictors

We used the following 13 variables as predictors.

- Longitude and latitude
- The integer day of the year
- An inverse distance weighting (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 (Hulley, 2021), one during the daytime and one at night, represented in kelvins
- Monthly vegetation, quantified as the enhanced vegetation index from Aqua (Didan, 2021)
- Two variables for surface imperviousness (from the National Land Cover Database; Dewitz, 2021): 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 (Just et al., 2024)
- Population density, from the Gridded Population of the World (Center For International Earth Science Information Network-CIESIN-Columbia University, 2018)
- Elevation, from the US Geological Survey's 3D Elevation Program (US Geological Survey, 2017)
- Hilliness, or local relative topography, quantified as the multi-scale topographic dissection index computed from elevation (Oyler et al., 2015)
- 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 (Just et al., 2024).

We computed distance from water using the North America Rivers and Lakes dataset (<http://web.archive.org/web/20230529213113/https://www.sciencebase.gov/catalog/item/4fb55df0e4b04cb937751e02>). We considered all lakes and reservoirs with areas of at least 1000 km², plus the ocean. 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.

3.3. Models

The core modeling approach used extreme gradient boosting (XGBoost) and IDW with station-level cross validation, as in Just et al. (2024). We conducted tuning as in Just et al. (2024) separately for each of the three DVs, resulting in a separate hyperparameter vector for each (Table 1).

3.4. 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 with 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

Table 1
Selected hyperparameters for the three dependent variables.

Dependent variable	Number of rounds	Max depth	Eta	Gamma	Lambda	Alpha
temp.min	500	9	0.11	0.110	58	0.0100
temp.mean	500	9	0.18	0.077	680	0.0081
temp.max	500	9	0.12	0.066	150	0.0380

the same daily Voronoi-diagram method we used for XIS-PM_{2.5}. We calculated SHAP (Lundberg et al., 2020) for our cross-validated predictions to quantify feature contributions.

4. Results

4.1. Cross-validation

Table 2 shows weighted results for each year of CV. The bias of our predictions per year ranged from -0.012 to $+0.025$ K for min temperature, -0.046 to $+0.003$ K for mean temperature, and -0.081 to -0.024 K for max temperature. Table 3 shows per-region performance for a single year; Figure S1 plots per-region performance for a single DV in every year. Table S1 shows unweighted performance at particularly isolated stations, as a demonstration of how the model performs in sparsely monitored regions. Finally, Tables S2 and S3 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 (Carrión et al., 2021).

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 2023. The result was 0.77 K, comparable in magnitude to our RMSEs from CV.

Fig. 1 shows a mean absolute SHapley Additive exPlanations (SHAP) for each feature (omitting the IDW feature which has much greater absolute SHAP than any other). SHAPs can be interpreted analogously to the terms of a linear-regression model: a SHAP of $+2.5$ for a given predictor and case means that the model attributes a $+2.5$ increase in its prediction for that case to that predictor. We see that although there is substantial variation by region, the largest contributions come from the IDW feature, elevation, longitude, and distance from water.

One may wonder why the SHAPs for LST are so small. Part of the answer appears to be competition with the overwhelmingly effective IDW predictor. When the model for 2010 mean temperature is refit in CV without the IDW predictor, the mean absolute SHAPs increase from 0.028 to 0.055 for daytime LST and 0.030 to 0.038 for nighttime LST. LST is largely missing (in 76% of cases for day and 82% for night): when we examine mean absolute SHAPs among only cases with non-missing LST, we find that these numbers increase further, to 0.150 for day and 0.149 for night.

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

4.2. 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. Fig. 3 maps predictions for the entire study area on the hottest day in 2023. Fig. 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.

4.3. Comparison with other models

Table 4 and 5 show RMSEs (stratified by year and then by seasons of 2023) of daily min temperature from our model and three gridded temperature products: PRISM (4 km in resolution; PRISM Climate Group, 2024), gridMET (4 km; Abatzoglou, 2013) and Daymet (1 km; version 4 revision 1; Thornton et al., 2021). Table S4 shows analogous results to Table 4 for max temperature. The models are tested on observations at Weather Underground stations. For each available year, we take a random sample of 10,000 such stations that lie in the intersection

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	907,131	4876	10.53	2.56	10.90	2.14	11.91	2.78
2004	1,266,937	5828	10.24	2.26	10.48	1.85	11.35	2.37
2005	1,778,765	8400	10.35	2.19	10.73	1.71	11.71	2.21
2006	2,335,197	9660	10.03	2.16	10.38	1.56	11.32	2.04
2007	2,689,582	10,839	10.76	2.09	11.11	1.54	12.10	2.06
2008	2,928,358	11,903	10.78	1.88	11.04	1.38	11.94	1.79
2009	3,190,170	12,416	10.69	1.90	10.94	1.39	11.80	1.79
2010	3,439,466	13,506	10.65	1.86	11.07	1.34	12.11	1.76
2011	3,830,866	14,825	10.90	1.90	11.30	1.36	12.28	1.76
2012	5,330,791	21,039	10.11	1.82	10.45	1.24	11.40	1.61
2013	5,929,147	22,068	11.07	1.80	11.29	1.28	12.09	1.69
2014	6,235,048	22,784	11.09	1.79	11.22	1.25	12.03	1.64
2015	6,245,577	23,163	10.58	1.76	10.71	1.24	11.52	1.62
2016	6,317,539	22,291	10.27	1.82	10.51	1.30	11.43	1.66
2017	6,492,587	22,851	10.36	1.82	10.68	1.29	11.66	1.64
2018	6,101,552	22,136	11.13	1.78	11.29	1.21	12.12	1.54
2019	6,454,618	25,514	11.16	1.73	11.40	1.20	12.26	1.53
2020	7,269,937	26,785	10.47	1.80	10.67	1.17	11.53	1.47
2021	7,835,606	27,255	10.66	1.77	10.81	1.17	11.66	1.47
2022	8,037,915	27,371	11.62	1.81	11.71	1.20	12.42	1.50
2023	8,280,718	29,270	10.43	1.78	10.71	1.19	11.61	1.48

Table 3
Weighted SDs and RMSEs (K) for 2023 broken down by region.

Region	Observations	Sites	Min temp.		Mean temp.		Max temp.	
			SD	RMSE	SD	RMSE	SD	RMSE
Ohio Valley	687,096	2529	8.52	1.22	8.73	0.75	9.63	1.05
Upper Midwest	506,838	1812	10.66	1.35	11.06	0.79	12.16	1.14
Northeast	1,119,757	3833	9.17	1.30	9.25	0.82	10.21	1.15
Northwest	771,008	2697	8.40	2.07	9.44	1.40	11.02	1.75
South	831,782	2962	9.42	1.38	9.24	0.89	9.68	1.19
Southeast	1,043,562	3766	8.05	1.19	7.36	0.75	7.58	1.09
Southwest	899,707	3197	10.09	2.43	10.79	1.73	11.65	2.02
West	2,016,434	6938	9.60	2.42	10.10	1.65	11.00	1.94
Northern Rockies and Plains	404,534	1536	10.79	1.87	11.44	1.23	12.59	1.50

of all four modeling regions and provide at least 347 days of observations (about 95% of a common year), so we only analyze years with at least this many stations available. We compute 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, and on the final year of comparisons, since PRISM’s date scheme is misaligned by one day and the next year of predictions is not yet available. For min temperature, with averaging across years, our model has 47% of the MSE of PRISM, 47% of gridMET, and 56% of Daymet. Without weighting, these figures become 39% of PRISM, 38% of gridMET, and 49% of Daymet. Yearly weighted biases range from -0.83 to -0.61 K for PRISM, -0.81 to -0.54 K for gridMET, -0.68 to -0.39 K for Daymet, and -0.06 to $+0.08$ K for XIS. For max

temperature, our results are less impressive than for min temperature, because the three competitor models are much improved compared to min temperature: XIS obtains 87% of the MSE of PRISM, 87% of gridMET, and 97% of Daymet. The yearly weighted biases range from -0.95

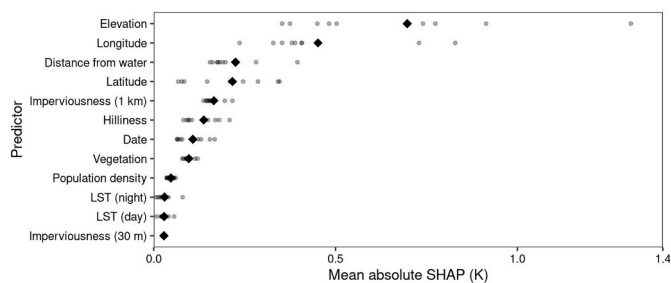


Fig. 1. Mean absolute SHAP of each predictor in 2010 for mean temperature (the IDW feature, which has much greater absolute SHAP than everything else, is omitted). Small dots show per-region means. Diamonds show overall means.

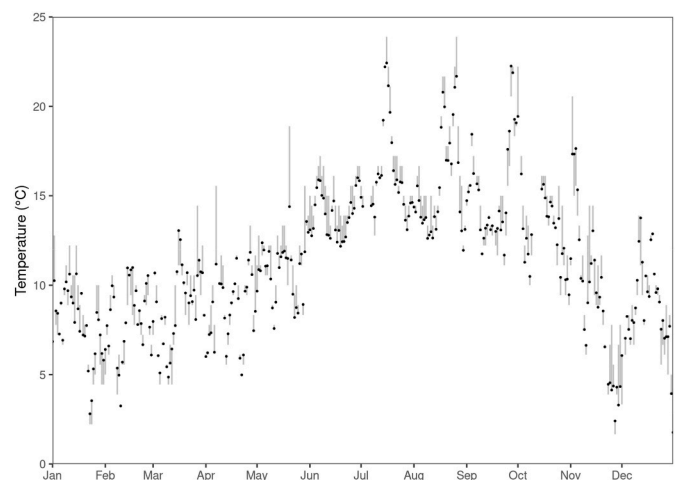


Fig. 2. A plot of predicted min temperature from CV in 2010 (points), and the distance from the observed value (line segments), for a station in the Chatsworth neighborhood of Los Angeles. This station was selected to have the yearly per-station unweighted RMSE closest to the median among all stations that had an observation for at least 347 days of 2010. Its RMSE is 1.27 K.

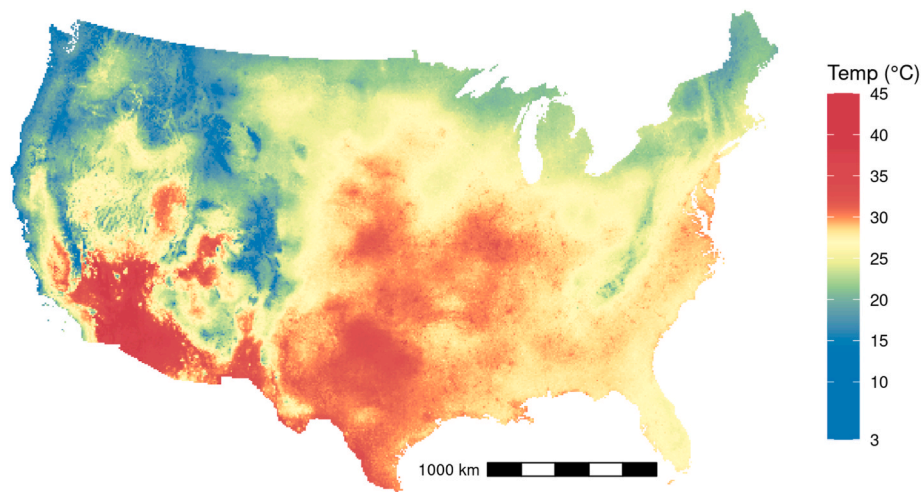


Fig. 3. Predicted mean temperature for Jul 27, 2023 across the study area, shown in the US National Atlas projection. We chose this date for having the highest mean temperature in 2023 across all stations. The underlying prediction grid has cells about 9097 m apart.

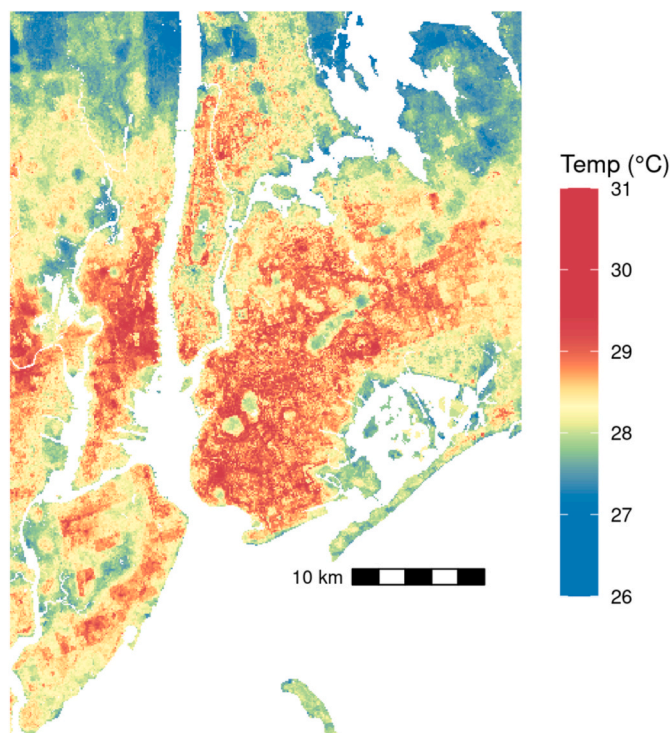


Fig. 4. Predicted mean temperature for Jul 27, 2023 in the New York City area. Areas of water have been masked out of the plot. In the center is Brooklyn; Staten Island lies to the southwest, Manhattan to the northwest, and Queens to the north. The underlying prediction grid has cells about 110 m apart.

Table 4
Comparison of minimum temperature weighted SD and RMSEs (K) for PRISM, gridMET, Daymet, and XIS.

Year	Observations	SD	PRISM	gridMET	Daymet	XIS
2015	3,613,296	10.67	2.56	2.55	2.36	1.68
2016	3,612,812	10.39	2.62	2.61	2.36	1.81
2017	3,602,753	10.56	2.64	2.68	2.40	1.82
2018	3,593,725	11.25	2.67	2.68	2.45	1.89
2019	3,600,723	11.22	2.60	2.59	2.36	1.72
2020	3,609,000	10.54	2.62	2.61	2.39	1.79
2021	3,508,692	10.77	2.57	2.59	2.32	1.79

to -0.58 K for PRISM, -0.73 to -0.41 K for gridMET, -0.68 to -0.31 K for Daymet, and -0.72 to -0.28 K for XIS.

To examine how XIS’s higher spatial resolution contributed to its improved performance, we also tried making XIS predictions for the 2023 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 2.293 K, compared to 2.291 K for using the true locations and 2.645 K for Daymet.

4.4. Model application to social vulnerability

We examined how minimum temperature in 2010-06-01 through 2010-08-31 related to the social vulnerability index in 2010 (Centers for Disease Control and Prevention, 2018). 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 mean of 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 0.75 K ([0.70, 0.79]), where the latter is a 95% CI, meaning that a change from minimum to maximum vulnerability was associated with a 0.75-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: 0.24 K ([0.19, 0.28]) for PRISM, 0.28 K ([0.24, 0.32]) for gridMET, and 0.20 K ([0.17, 0.23]) for Daymet.

5. Discussion

We present a daily spatiotemporal air temperature model for the contiguous US that covers 21 years. Our model, XIS-Temperature, builds on a large time-resolved dataset of ground observations, NOAA’s MADIS. As expected, our model shows substantial accuracy, which increases in more recent years, since the number of observations available increases tenfold from 2003 to 2023.

We compared XIS predictions for min and max temperatures with three leading gridded models at 10,000 private weather stations, 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 2023, XIS had the least RMSE for each season, as well as the least variability in RMSE across seasons. A sensitivity

Table 5

Weighted SD and RMSEs (K) of minimum temperature for the various models in 2023, broken down by season. We use December from 2022 instead of 2023 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	891,385	17,970	9.85	3.23	3.89	3.28	2.61
Spring	914,624	10,000	8.78	3.10	3.20	2.78	2.37
Summer	913,267	10,000	5.75	2.84	2.89	2.53	2.34
Fall	904,499	10,000	8.72	3.07	3.14	2.91	2.54

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 (Oyler et al., 2015; Thornton et al., 2021). 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 (Bernhardt et al., 2018). 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 in the summer of 2010 with social vulnerability, nested within counties, similar to our previous analysis in the Northeast US (Carrion et al., 2021). 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; a more thorough evaluation of temperature disparities, as we have previously shown for the Northeast US (Carrion et al., 2024), 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. While we only train on data that have passed quality control from MADIS and we detail a number of further filtering steps, future refinements of XIS could explore an adaptive buddy check (where the threshold for data exclusion depends on the local space-time variability) versus the use of hardcoded thresholds (Dee et al. 2001); Our 2023 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, which was the hardest season to predict for XIS as well as for the competing models. Given the importance of the IDW of temperature measurements and elevation in our SHAP analysis, there may be further feature engineering that offers improvement. A future refinement of our IDW could incorporate environmental lapse rates (e.g., considering both horizontal distance and elevation gradients in IDW construction) (Daly et al., 2008). 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 (Oyler et al., 2016). Future XIS development could adopt the approach of constructing measures of monthly relative LST variation over local windows (Oyler et al., 2015) to identify 1-km pixels that are hotter or colder than nearby pixels, rather than directly including the daily (and often missing) LST values.

Applicability of the XIS-Temperature model in environmental epidemiology includes several important opportunities for future research. Health studies of climate-related variables would benefit from the simultaneous consideration of temperature and mass-based humidity metrics (Baldwin et al., 2023). The reusable structure of the XIS framework lends itself to other spatiotemporal variables and there is a XIS-Humidity under development. Important exposure modeling opportunities also include the propagation of spatial uncertainty (related to people's time-activity patterns) and model prediction uncertainty. XIS-Temperature could also be used for important analyses of other outcomes, such as energy demand at the complex nexus of disparities in heat, housing, and energy justice (Carrion et al., 2024).

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.

CRediT authorship contribution statement

Allan C. Just: Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Conceptualization. **Kodi B. Arfer:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Johnathan Rush:** Writing – review & editing, Software. **Itai Kloog:** Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Allan Just reports financial support was provided by National Institutes of Health. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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, P20 AG089308, 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.envres.2024.120731>.

Data availability

Intermediate data files to reproduce our analyses are openly available on Zenodo at <https://zenodo.org/doi/10.5281/zenodo.7331250>. These data were derived from a combination of public-domain and restricted-access sources described in our Method.

References

- Abatzoglou, J.T., 2013. Development of gridded surface meteorological data for ecological applications and modelling. *Int. J. Climatol.* 33, 121–131. <https://doi.org/10.1002/joc.3413>.
- Baldwin, J.W., Benmarhnia, T., Ebi, K.L., Jay, O., Lutsko, N.J., Vanos, J.K., 2023. Humidity's role in heat-related health outcomes: a heated debate. *Environ. Health Perspect.* 131, 055001. <https://doi.org/10.1289/EHP11807>.
- Bernhardt, J., Carleton, A.M., LaMagna, C., 2018. A comparison of daily temperature-averaging methods: spatial variability and recent change for the CONUS. *A Comparison of Daily Temperature-Averaging Methods* 31, 979–996. <https://doi.org/10.1175/JCLI-D-17-0089.1>.
- Carrión, D., Arfer, K.B., Rush, J., Dorman, M., Rowland, S.T., Kioumourtzoglou, M.-A., Kloog, I., Just, A.C., 2021. A 1-km hourly air-temperature model for 13 northeastern U.S. states using remotely sensed and ground-based measurements. *Environ. Res.* 200, 111477. <https://doi.org/10.1016/j.envres.2021.111477>.
- Carrión, D., Rush, J., Colicino, E., Just, A.C., 2024. Residential segregation and summertime air temperature across 13 northeastern U.S. states: potential implications for energy burden. *Environ. Res. Lett.* 19, 084005. <https://doi.org/10.1088/1748-9326/ad5b77>.
- Center For International Earth Science Information Network-CIESIN-Columbia University, 2018. Gridded population of the World, version 4 (GPWv4): population count, revision 11. Gridded population of the World. <https://doi.org/10.7927/H4JW8BX5>.
- Centers for Disease Control and Prevention, 2018. Agency for toxic Substances and Disease Registry/geospatial research, analysis, and Services program. CDC/ATSDR Social Vulnerability Index (SVI) 2018 Database US [WWW Document]. URL. <https://www.atsdr.cdc.gov/placeandhealth/svi/index.html>. (Accessed 17 June 2022).
- Crosson, W.L., Al-Hamdan, M.Z., Insaf, T.Z., 2020. Downscaling NLDAS-2 daily maximum air temperatures using MODIS land surface temperatures. *PLoS One* 15, e0227480. <https://doi.org/10.1371/journal.pone.0227480>.
- 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. *Intl Journal of Climatology* 28, 2031–2064. <https://doi.org/10.1002/joc.1688>.
- Dee, D.P., Ruhkovets, L., Todling, R., Da Silva, A.M., Larson, J.W., 2001. An adaptive buddy check for observational quality control. *Quart J Royal Meteorol Soc* 127, 2451–2471. <https://doi.org/10.1002/qj.49712757714>.
- Dewitz, J., 2021. National land cover Database (NLCD) 2019 products. <https://doi.org/10.5066/P9KZCM54>.
- Didan, Kamel, 2021. MODIS/Aqua vegetation indices monthly L3 global 1km SIN grid V061. <https://doi.org/10.5067/MODIS/MYD13A3.061>.
- Gutiérrez-Avila, I., Arfer, K.B., Wong, S., Rush, J., Kloog, I., Just, A.C., 2021. A spatiotemporal reconstruction of daily ambient temperature using satellite data in the Megalopolis of Central Mexico from 2003 to 2019. *Int. J. Climatol.* 41, 4095–4111. <https://doi.org/10.1002/joc.7060>.
- Hulley, Glynn, 2021. MODIS/Aqua land surface temperature/3-band emissivity daily L3 global 1km SIN grid day V061. <https://doi.org/10.5067/MODIS/MYD21A1D.061>.
- Just, A.C., Arfer, K.B., Rush, J., Lyapustin, A., Kloog, I., 2024. XIS-PM2.5: a daily spatiotemporal machine-learning model for PM2.5 in the contiguous United States. *Earth Space Sci. Open Arch.* <https://doi.org/10.1002/essoar.10512861.2>.
- Karl, T.R., Koscielny, A.J., 1982. Drought in the United States: 1895–1981. *J. Climatol.* 2, 313–329. <https://doi.org/10.1002/joc.3370020402>.
- Kloog, I., Chudnovsky, A., Koutrakis, P., Schwartz, J., 2012. Temporal and spatial assessments of minimum air temperature using satellite surface temperature measurements in Massachusetts, USA. *Sci. Total Environ.* 432, 85–92. <https://doi.org/10.1016/j.scitotenv.2012.05.095>.
- Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2, 56–67. <https://doi.org/10.1038/s42256-019-0138-9>.
- Miller, P., Barth, M., Benjamin, L., Helms, D., Campbell, M., Facundo, J., O'Sullivan, J., 2005. The meteorological assimilation data ingest system (MADIS) - providing value-added observations to the meteorological community. In: Presented at the 21st Conference on Weather Analysis and Forecasting/17th Conference on Numerical Weather Prediction.
- NASA, 2022. NLDAS-2 forcing dataset information. <https://ldas.gsfc.nasa.gov/nldas/v2/forcing>. (Accessed 23 June 2022).
- NOAA, 2013. U.S. Climate Regions | Monitoring References. National Centers for Environmental Information (NCEI) [WWW Document]. URL. <https://www.ncdc.noaa.gov/monitoring-references/maps/us-climate-regions.php>. (Accessed 10 July 2020).
- Oyler, J.W., Ballantyne, A., Jencso, K., Sweet, M., Running, S.W., 2015. Creating a topoclimatic daily air temperature dataset for the conterminous United States using homogenized station data and remotely sensed land skin temperature. *Int. J. Climatol.* 35, 2258–2279. <https://doi.org/10.1002/joc.4127>.
- Oyler, J.W., Dobrowski, S.Z., Holden, Z.A., Running, S.W., 2016. Remotely sensed land skin temperature as a spatial predictor of air temperature across the conterminous United States. *J. Appl. Meteorol. Climatol.* 55, 1441–1457. <https://doi.org/10.1175/JAMC-D-15-0276.1>.
- PRISM Climate Group, 2024. Oregon state U. <https://www.prism.oregonstate.edu/>. (Accessed 11 March 2022).
- State Climate Extremes Committee (SCEC), 2022. National centers for environmental information (NCEI). Record Past [WWW Document]. URL. <http://web.archive.org/web/20220812134705id>. (Accessed 15 September 2022).
- 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>.
- Tuholske, C., Caylor, K., Funk, C., Verdin, A., Sweeney, S., Grace, K., Peterson, P., Evans, T., 2021. Global urban population exposure to extreme heat. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2024792118. <https://doi.org/10.1073/pnas.2024792118>.
- US Geological Survey, 2017. 1 arc-second digital elevation models (DEMs) - USGS national map 3DEP downloadable data collection. <https://www.sciencebase.gov/catalog/item/4f70aa71e4b058caae3f8de1>. (Accessed 15 June 2022).
- Vandal, T., Kodra, E., Ganguly, A.R., 2019. Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. *Theor. Appl. Climatol.* 137, 557–570. <https://doi.org/10.1007/s00704-018-2613-3>.
- Yu, W., Yang, J., Sun, D., Ren, J., Xue, B., Sun, W., Xiao, X., Xia, J., Li, X., 2024. How urban heat island magnifies hot day exposure: global unevenness derived from differences in built landscape. *Sci. Total Environ.* 945, 174043. <https://doi.org/10.1016/j.scitotenv.2024.174043>.