Environmental Research 200 (2021) 111477

Contents lists available at ScienceDirect

Environmental Research

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

A 1-km hourly air-temperature model for 13 northeastern U.S. states using remotely sensed and ground-based measurements

Daniel Carrión [a,*], Kodi B. Arfer [a], Johnathan Rush [a], Michael Dorman [b], Sebastian T. Rowland [c], Marianthi-Anna Kioumourtzoglou [c], Itai Kloog [a,b], Allan C. Just [a,d]

[a] Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[b] Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[c] Department of Environmental Health Sciences, Columbia University, New York, USA
[d] Institute for Exposomic Research, Icahn School of Medicine at Mount Sinai, New York, USA

ARTICLE INFO

Keywords:
Air temperature
Remote sensing
MODIS
XGBoost
NLDAS-2
Social vulnerability

ABSTRACT

Background: Accurate and precise estimates of ambient air temperatures that can capture fine-scale within-day variability are necessary for studies of air temperature and health.
Method: We developed statistical models to predict temperature at each hour in each cell of a 927-m square grid across the Northeast and Mid-Atlantic United States from 2003 to 2019, across ~4000 meteorological stations from the Integrated Mesonet, using inputs such as elevation, an inverse-distance-weighted interpolation of temperature, and satellite-based vegetation and land surface temperature. We used a rigorous spatial cross-validation scheme and spatially weighted the errors to estimate how well model predictions would generalize to new cell-days. We assess the within-county association of temperature and social vulnerability in a heat wave as an example application.
Results: We found that a model based on the XGBoost machine-learning algorithm was fast and accurate, obtaining weighted root mean square errors (RMSEs) around 1.6 K, compared to standard deviations around 11.0 K. We found similar accuracy when validating our model on an external dataset from Weather Underground. Assessing predictions from the North American Land Data Assimilation System-2 (NLDAS-2), another hourly model, in the same way, we found it was much less accurate, with RMSEs around 2.5 K. This is likely due to the NLDAS-2 model's coarser spatial resolution, and the dynamic variability of temperature within its grid cells. Finally, we demonstrated the health relevance of our model by showing that our temperature estimates were associated with social vulnerability across the region during a heat wave, whereas the NLDAS-2 showed a much weaker association.
Conclusion: Our high spatiotemporal resolution air temperature model provides a strong contribution for future health studies in this region.

1. Introduction

There is growing interest in the association between temperature and health. Extreme temperatures are associated with adverse pregnancy outcomes (Zhang et al., 2017), cardiovascular events (Lin et al., 2009), and mortality (Gasparrini et al., 2015), among other adverse events. These associations are nonlinear, with both low and high temperatures (compared to moderate temperatures) relating to higher risk of an adverse event (Gasparrini et al., 2015). Most studies have focused on daily mean or maximum temperature, but some studies have shown that hourly temperature and diurnal temperature variation are related to heart attacks (Rowland et al., 2020) and mortality (Shi et al., 2015). Consequently, highly spatially and temporally resolved temperature models are useful for refined exposure assessment in health studies.

Urban areas are often warmer than their suburban or rural surroundings, a phenomenon known as the urban heat island effect. The difference can be on the order of 10 K on calm, clear nights, depending on factors such as vegetation, building geometry, construction materials, surface permeability, and anthropogenic heat sources (Oke, 1982). Since these factors vary in space, even within urban areas, near-surface

[*] Corresponding author.
E-mail address: daniel.carrion@mssm.edu (D. Carrión).

https://doi.org/10.1016/j.envres.2021.111477
Received 15 December 2020; Received in revised form 1 June 2021; Accepted 1 June 2021
Available online 12 June 2021

air temperature need not decrease linearly with distance from a city center, and can vary substantially between neighborhoods of a city. Thus, a city can be described as a heterogeneous mosaic with hotspots, forming an "urban heat archipelago" (Buyantuyev and Wu, 2010). For example, in Gothenburg, Sweden, from 1988 to 1990, mean temperature varied by up to 2.7 K between dense buildings and a nearby open area, and the 4 K mean temperature difference between the city center and a large park was similar to the mean temperature difference between the city center and rural surroundings (Eliasson, 1996). In Portland, Oregon, in the summer of 2006, a large park was often 2–4 K cooler than the city's rural surroundings, while the city's built areas were 2–5 K warmer (Hart and Sailor, 2009).

Another important finding is that temperature is related to social disadvantage. A study in Phoenix, Arizona, found that remotely sensed daytime land surface temperatures were 0.36 K cooler for each additional $10,000 in median family income of a census block group (Buyantuyev and Wu, 2010). Similarly, across the US, neighborhoods with historic housing discrimination (redlining) have been found to have higher summertime intra-urban land surface temperatures (Hoffman et al., 2020). Better exposure assessment, then, is necessary for heat-vulnerability and health-disparities research. Furthermore, many studies have found that within-day temperature variation can be large in any given location. While many health studies use mean daily temperature, it is possible that extrema or diurnal variation are also relevant to human health (Vicedo-Cabrera et al., 2016).

Many data products offer measures of temperature, but they often have substantial drawbacks for health applications. For example, many satellite instruments estimate land surface temperature (LST), but near-surface air temperature, measured at ground-based monitors, is more relevant for human thermoregulatory capacity. Those that do provide air temperature have a coarse resolution that are consequently inadequate for human health; for example, the AIRS temperature profile product has a 50-km horizontal resolution (Teixeira, 2013). While there are weather stations throughout the US that measure temperature, they are unevenly placed with regard to human populations (Kloog et al., 2014). Policy actions are often made using weather reports from meteorological stations (e.g., airport weather stations) that may not represent finer-scale variation in temperature where people live. When stations are tens of kilometers away from people, there is substantial uncertainty in exposure assignment, especially in urban areas where temperature varies dramatically over small distances. Many different temperature-prediction models have arisen as a result of such challenges. For example, the North American Land Data Assimilation System (NLDAS-2) model has an hourly temperature product at a coarse 0.125° grid (e.g., each cell covers about an 11 km × 14 km rectangle in New York State) throughout the contiguous US (Rui and Mocko, 2019). Kloog et al. developed daily 1-km mean temperature models across the northeastern US (2014) and France (2017). Oyler et al. developed a daily 800-m model with mean, minimum, and maximum temperature predictions (2015). Most recently, Crosson et al. developed a daily 1-km maximum and minimum temperature model (2020). These models all have notable strengths and limitations. Most of these models assume linear relationships between variables, but allowing for nonlinear relationships may be important to account for vast changes in temperature across space and time. Furthermore, the models have limited temporal resolution. For example, the models that estimate maximum and minimum temperatures do not estimate when during the day these values were attained. High temporal resolution is important for health studies, which may have highly time-resolved outcome data. Finally, researchers may overestimate the performance of models if they do not account for variation in the density of ground observations.

In this study, we created and compared the performance of five different statistical temperature models that integrate satellite and ground-based observations, and we identified the best model based on predictive accuracy and computational efficiency. We included both linear and non-linear models and we evaluated our models with spatial cross-validation (CV) and spatial weighting. We demonstrated the utility of the best model by comparing predictions across areal measures of social vulnerability, which is of particular interest for studies of socioeconomic status and health.

2. Materials and methods

We fit models predicting ground-based air-temperature measurements per combination of year and hour; for example, we fit one distinct model for midnight UTC in 2013, another for 1 AM UTC in 2013, and so on. Our goal was to compare five types of models and select the best performer. To select the model type, we first constructed four models, considering only 2 h of the day in each of two years. We used ten-fold CV to compare the accuracy of the models and selected a single type of model. We compared the selected model to NLDAS-2, validated it with a separate monitoring network, and examined how individual predictors contributed to the predictions. Finally, we made predictions across the study region and time period. Our summarized approach and procedures are depicted in Fig. 1 and detailed below.

2.1. Study region and time period

We modeled temperature for each year from 2003 through 2019. Our models covered the Northeast and Mid-Atlantic US, namely the states of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Delaware, Pennsylvania, Maryland, West Virginia, and Virginia, plus Washington, DC (Fig. 2). We represented the study region using the same grid as our satellite inputs, which consists of square cells, approximately 927 m on a side (nominally 1 km), in a sinusoidal projection. Restricted to our study region, the grid had 750,808 cells and covered a total of 644,670 km^2.

2.2. Ground-based air temperature

We obtained near-surface air-temperature data from the Meteorological Assimilation Data Ingest System (MADIS; NOAA, 2018), which is maintained by the US National Atmospheric and Oceanic Administration (NOAA). MADIS is a database assimilating weather observations from networks around the world, with the greatest data density in North America. It spans from 2001 to the present, includes automated quality control checks, and enforces uniformity in reported data, including metadata. The number of observations has grown over time as NOAA has continued to ingest data from new partner agencies. MADIS provides access to many datasets, but we used the meteorological surface observations from the National Mesonet dataset of the US, which is available in the MADIS research data archive to registered users. All MADIS stations we selected are shown in Fig. 2.

We used air temperature observations from Weather Underground to externally validate our models. Weather Underground is a private commercial network of over 250,000 personal weather stations around the world, many of which are in the US and Europe (The Weather Company, 2018). Thus, Weather Underground served as a useful independent dataset to test the performance of our models at new locations in the same region. For context, there are 705 1-km grid cells with co-located MADIS and Weather Underground stations, 3490 grid cells with only MADIS stations, and 1362 grid cells with only Weather Underground stations, while the remainder have neither station type (Figure S1).

We applied several filters and quality control checks to the MADIS and Weather Underground data, which are inspired by similar approaches taken by others (Dirksen et al., 2020; Gutiérrez-Avila et al., 2021; Napoly et al., 2018). Each dataset and year was processed independently:

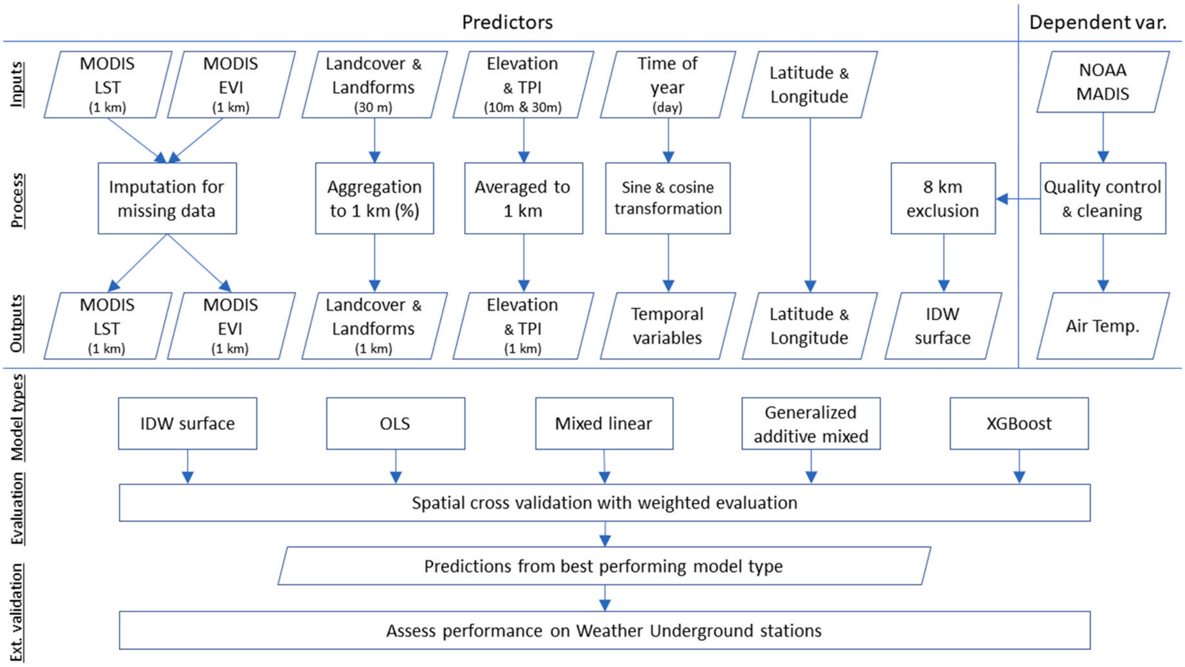

Fig. 1. Summarized flowchart of the near-surface air-temperature modeling strategy for 2003–2019 across the Northeast and Mid-Atlantic United States. Parallelograms represent datasets and rectangles are processes. All model types use the same predictors, except for the IDW surface model type, which only uses the IDW surface. LST = land surface temperature, EVI = enhanced vegetation index, IDW = inverse distance weighting, TPI = topographic position index.

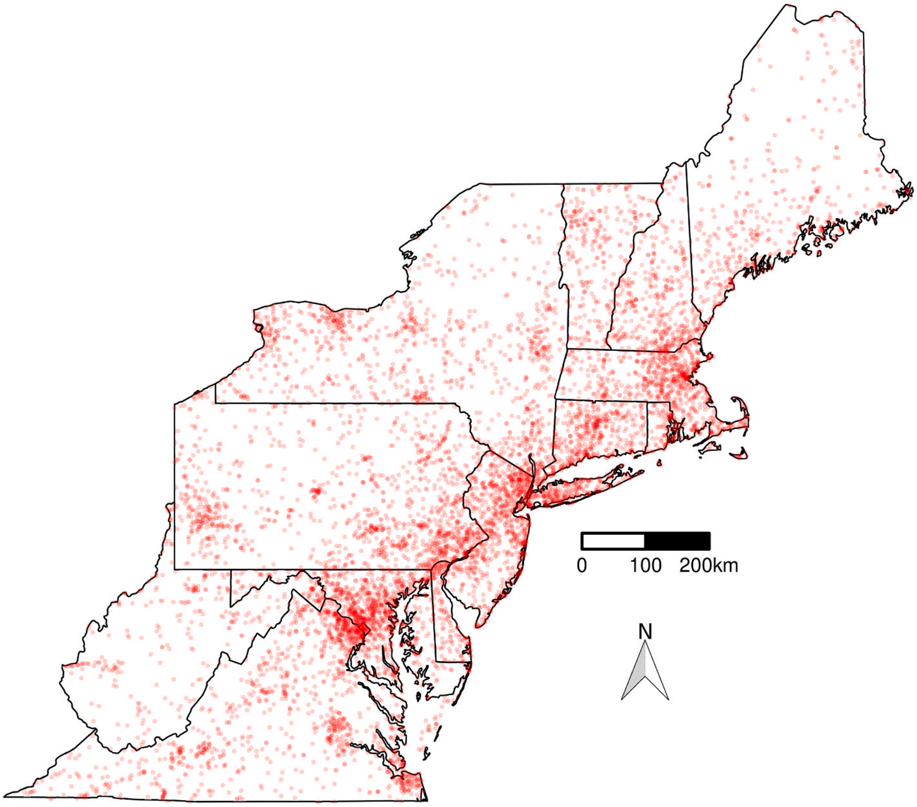

Fig. 2. The study region. The black lines are state borders, and the red points are grid cells with at least one MADIS station in at least one year of the study. Partial transparency is used to visualize areas with overplotting where multiple stations are sufficiently close together to make their points appear one on top of the other at this spatial resolution and for the scale represented. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

1. Select a single observation per named station ID and hour. Observations at more common locations (i.e., longitude–latitude pairs) are preferred.
2. Remove observations outside NOAA's official extreme-temperature records for the study region: −47 °C and 45 °C (NOAA, 2020).
3. Select a single observation per location and hour. Observations closer to the hour are preferred.
4. Compare each observation to its two nearest neighbors (no more than 100 km away) at the same hour. If the temperature differs from that of both neighbors by 20 K or more, drop all data for that station.

5. Select a single location for each 1-km Moderate Resolution Imaging Spectroradiometer (MODIS) grid cell. Locations that are more common are preferred.

2.3. Predictors

As predictors, we included 34 variables (Table 1), further details of which are given below.

2.3.1. Satellite-based land surface temperature

The MODIS is an instrument aboard NASA's Aqua and Terra satellites that provides LST data across the study region at 1-km (more precisely, 927-m) resolution. Each satellite provides a daytime and a nighttime retrieval, for a total of up to 4 LST retrievals per day, with Terra overpass times at approximately 10:30 AM/PM and Aqua at 1:30 AM/PM local solar times. We matched each of our hourly observations to the temporally closest overpass of each of the four types. The LST data from the Collection 6 processing were downloaded for both satellites (Wan et al., 2015a, 2015b). Observations whose quality-control codes indicated an average error greater than 2 K were treated as missing. The rates of missingness per year, aggregating across days, satellites, and overpasses, ranged from 41% to 47%.

2.3.2. Topography

We used a series of topographical measures to account for important physical processes related to the warming and cooling of Earth. Land cover was derived from the 2011 National Land Cover Dataset, which provides 16 categorical land cover classes at a 30-m spatial resolution (Homer et al., 2015). Given that our ultimate model has a 1-km resolution, we computed the proportion of each class in each of our 1-km grid cells. This dataset also has a 30-m impervious surface layer, of which we similarly computed the mean for each grid cell. Finally, we used focal-window processing to compute the proportion of surface water within a 15-km radius of each grid cell.

We incorporated two time-invariant measures of topography: void-filled elevation from the NASA Shuttle Radar Topography Mission at 1-arc-second spatial resolution (Farr et al., 2007), and multi-scale topographic position index (Theobald et al., 2015) derived from the USGS 1/3-arc-second digital elevation model. The topographic position index is a continuous measure of relative topography, that is, whether a point location is a peak or is in a valley. Higher values indicate a greater peak. The elevation is at approximately 30-m resolution, and the topographic position index is at approximately 10-m resolution, so we computed the mean for each of our grid cells.

2.3.3. Other predictors

A monthly enhanced vegetation index (EVI) was included in models as a measure of vegetative cover and was derived from Collection 6 of Terra satellite retrievals (Didan, 2015). The rate of missingness was 2.6% in 2003, but ranged from 0.030% to 0.092% per year in the remaining years.

Table 1
All the variables used as inputs to the predictive models.

Variable types	Variables
Spatial gradients	1) Longitude; 2) Latitude; 3) Inverse-distance weighted air temperature; 4) Elevation; 5) Topological position index;
Temporal gradients	6) Sine and 7) Cosine terms for seasonality;
MODIS LST	8) Terra day; 9) Terra night; 10) Aqua day; 11) Aqua night;
MODIS Vegetation	12) Enhanced vegetation index;
Land cover	13) Water; 14) Developed - open space; 15) Barren; 16) Deciduous forest; 17) Evergreen forest; 18) Mixed forest; 19) Shrub/scrub; 20) Grassland/herbaceous; 21) Pasture/hay; 22) Cultivated crops; 23) Woody wetlands; 24) Emergent herbaceous wetland; 25) Impervious surface; 26) Water within buffer;
Landforms	27) Peak/ridge; 28) Upper slope; 29) Upper slope - flat; 30) Lower slope; 31) Lower slope - warm; 32) Lower slope - flat; 33) Valley; 34) Valley - narrow

The longitude and latitude coordinates of ground observations were included to allow for spatial variation along east-west and north-south axes. Seasonality was included using trigonometric time terms: sin(2π (d - 1)/n) and cos(2π(d - 1)/n), where d is the day of the year and n is the total number of days in the year.

3.4. Imputation for missing data

To impute missing values in the four LST variables and EVI, we used a simplified similar-pixels method inspired by Yu et al. (2019). First, all of the grid cells in the study area were split with 30-means clustering, specifically with the MacQueen algorithm (MacQueen, 1967). We used continuous heat-insolation load index (CHILI) (Theobald et al., 2015) to help with imputation. CHILI was derived from the National Elevation Dataset (Gesch et al., 2002). We used CHILI to estimate the relative impact of incident radiation in an area, considering attributes such as latitude, aspect, and slope. The clustering variables were elevation, topographic position index, CHILI, percent impervious surface, percent water (in a 15-km buffer), and 2015 population density. These variables are all temporally invariant, so we could use a single set of clusters for all years and hours. Then, for each cluster, year, and satellite variable (Terra day temperature, Terra night temperature, Aqua day temperature, Aqua night temperature, or EVI), we fit a linear model with ordinary least squares (OLS). The outcome was the satellite variable, and the predictors were the clustering variables, longitude, latitude, and one dummy variable for each time unit (day for LST, month for EVI) with more than 100 non-missing values of the outcome. In the case of LST, these models were fit separately within months. We trained each model on the non-missing values and used its predictions to replace missing values. Hence, we had observed or imputed LST and EVI variables for all of our air temperature-prediction models.

3.5. Predictive modeling

3.5.1. Hot and cold hours for initial evaluation

For model development, we restricted attention to 2 hours in each of two years, 2004 and 2013, both for computational speed and to ensure we had data for testing models that we had not already used for selecting models. For each station and UTC-based day of hourly temperature observations, using all years of data, we recorded the hottest and coldest hour. We found that the most frequent hottest hour of the day was 8 PM UTC (3 PM Eastern Standard Time) and the most frequent coldest hour was 10 AM UTC (5 AM Eastern Standard Time). We refer to these times as the "hot hour" and "cold hour", respectively.

3.5.2. Model performance and selection

3.5.2.1. Spatial cross-validation. Prediction modeling often uses CV for model evaluation. A standard CV approach might involve splitting the observations randomly into 10 groups (folds) and training the model 10 times, once without each fold, using the remaining data. We have found that without further adjustments, this approach can yield overly optimistic results in spatiotemporal modeling (Just et al., 2020a, 2020b), so we used a spatial CV scheme to prevent our models from being trained on stations close to the stations for which they were making predictions. For each year, we randomly split stations into 10 folds; for each fold, we excluded all stations from training that were within 8164 m of a station in the test set. The quantity 8164 m was the median distance of all grid cells from their nearest station in 2018. Thus, when making predictions in 2018 to arbitrary grid cells, one makes predictions 8 km, on average, away from the nearest station. Setting exclusion zones of this size gave us a CV scheme that is representative of how the model will be used.

3.5.2.2. Weighted evaluation. An issue with CV, given the non-random spatial arrangement of ground monitors, is that it would tend to emphasize model performance in areas that are dense with monitors. This emphasis is problematic because the purpose of our model is to predict temperatures in areas that do not already have good monitor coverage. Hence, we created weighted metrics by dividing the 750,808 cells of the master grid into 100 approximately equally sized evaluation regions (Figure S2), and assigning a total weight of 1 to each group of observations in a single region at a single time. Within these groups, observations were weighted equally. We chose the evaluation regions with an algorithm that iteratively assigns cells in contiguous square rings.

3.5.2.3. Selection of model type and comparisons. Our model-selection process was based on minimizing the weighted prediction error in CV, being mindful of diminishing returns as computation time increases. When comparing models, we first present the standard deviation (SD) of the dependent variable as a measure of the overall variability that we seek to explain. We focus our model evaluation and model comparison on the root mean square error (RMSE), which can be compared to the SD to see the decrease in error attributable to the model. In this way, the SD serves as a useful benchmark to contextualize our prediction error, e.g. a RMSE of 2 K might be considered strong performance if the SD is 10 K, but not as strong if the SD is 4 K. We also present our results alongside the prediction error from NLDAS-2 estimates of air temperature at 2-m height. In this case, the prediction for a given station at a given hour is simply the NLDAS-2 temperature for the NLDAS-2 grid cell in which the station falls. This comparison is provided since NLDAS-2 estimates have been used in human health studies (Rowland et al., 2020; Wu et al., 2018) and public health tracking regarding heat impacts (Centers for Disease Control & Prevention, 2020).

3.5.3. Statistical model types

We considered five types of models, varying from simple to complex, and compared their performance using CV. Except for the simplest (IDW), all models used the same predictors.

3.5.3.1. Inverse distance weighting procedures. The simplest model, IDW, was both considered alone and used as a predictor in all other models. To calculate the IDW surface, we predicted the temperature at a grid cell as the mean of temperatures at all other stations (except those withheld for CV) at the same time, weighted by the reciprocal of the squared distance ($1/d^2$, where d is distance) (Quagliolo et al., 2020; Samanta et al., 2012). IDW surfaces are typically constructed using all other stations to make predictions at a given station, withholding only the value at the station being predicted. However, such an IDW scheme, when used with the dependent variable in CV, leads to leakage of test data into the training set, with commensurate overfitting and overly optimistic assessment of model performance. To avoid this leakage, we developed an IDW method that draws on differing subsets of the stations for predictions made to each location, based on our CV folds. When considering IDW alone as a model, each point calculated in the IDW used only those points not in the test set, so there were 10 sets of allowed stations. When incorporating IDW as a predictor for other models, it was necessary to withhold sets of stations not just for the observations in the test fold to avoid leakage, but also points within the same fold as each training station to avoid overfitting of the IDW. We sped up the computation of IDW by caching distance weights and sets of allowed stations, which saved time because we needed to make IDW interpolations for many different times at the same locations.

3.5.3.2. Regression models. Another set of models was based on regression. We considered 1) OLS, 2) a mixed-effects linear model with a set of per-day random intercepts in place of the trigonometric time terms, and 3) a generalized additive mixed model (GAMM) that took the mixed-effects model and added a three-way tensor-product smooth of longitude, latitude, and the integer day of the year (Wood, 2006). Each dimension of the penalized tensor-product smooth was allocated an upper limit of 10 degrees of freedom.

3.5.3.3. XGBoost models. Our final model type used a machine-learning algorithm called extreme gradient boosting (XGBoost; Chen and Guestrin, 2016). XGBoost grows a sequence of regression trees, fitting each tree to the residual of the assemblage of all prior trees, and uses several kinds of tunable regularization, allowing it to strike a balance between flexibility, avoidance of overfitting, and computation time. We increased parsimony by fixing the number of trees at 100 and using Dropouts meet Multiple Additive Regression Trees (DART; Vinayak and Gilad-Bachrach, 2015), a dropout method that ignores a randomly selected subset of existing trees during the construction of each new tree.

We tuned six of XGBoost's hyperparameters with our data for the cold hour in 2013 and the hot hour in 2004. We first randomly selected 50 sets of hyperparameters with a Latin hypercube sampling technique, ensuring broad coverage of the six-dimensional hyperparameter space (Just et al., 2020a, 2020b; Stein, 1987). Then we performed CV in each of the two data slices to assess each hyperparameter set. On the basis of combined performance, we chose the following hyperparameters: eta (learning rate) = 0.24; gamma (minimum split loss) = 0.023; lambda (L_2 regularization) = 0.021; alpha (L_1 regularization) = 5.9; max_depth (maximum tree depth) = 9; and dropout rate = .01, as defined in the XGBoost documentation (XGBoost developers, 2020).

3.6. External validation

After model selection, we assessed the model's performance using a monitoring network on which the model was never trained: Weather Underground. We assessed predictive performance, via RMSE, for all available hourly observations at the 2067 Weather Underground personal weather stations in the study region in 2013.

3.7. Interpreting variable importance and model predictions

All of our predictors have physiographic or topographic reasons for inclusion because they capture various sources of spatial and temporal variability in air temperature. Because the empirical relationships fit by XGBoost and the relative contribution of each predictor is not easily summarized, we used Shapley additive explanations (SHAPs; Lundberg et al., 2020) to interpret our XGBoost models. The method uses game theory to decompose each prediction into a sum of real numbers (SHAPs), one for each predictor, plus a bias term that applies to all predictions. Thus, a SHAP can be interpreted like the product of a coefficient and predictor value in OLS; for example, a SHAP of −2 means the model attributes a 2-unit decrease in its prediction to that predictor. The distribution of SHAPs can be used to understand variable contributions across the distribution of the predictor, aiding in model interpretability. The mean absolute SHAP of a given predictor is a summary measure of feature importance.

3.8. Model application to social vulnerability

Recent literature has shown that extreme summertime temperatures are associated with area-level measures of racism and socioeconomic status (Buyantuyev and Wu, 2010; Hoffman et al., 2020). This is an area of interest for social scientists, resulting in constructs such as energy burden and energy insecurity (Hernández, 2016; Ito et al., 2018), as well as public health researchers, due to downstream influences on the health of vulnerable populations (Madrigano et al., 2015). We explored the association between social vulnerability and the temperature at midnight EDT on July 22, 2011, which has the hottest observed mean temperature at midnight (when most of the population would be home) of all days in the study period.

Our measure of social vulnerability was the Centers for Disease Control and Prevention's social vulnerability index for 2010 (Flanagan et al., 2011, 2018). Each census tract has a vulnerability score, derived from 15 census indicators of social disadvantage, and computed as a quantile rank across the US. Scores range from 0 (least vulnerable) to 1 (most vulnerable). We focus on the association of temperature and vulnerability within counties to avoid comparisons across vastly different physical and human geographies. For each of two kinds of temperature estimates (our XGBoost model vs. the NLDAS-2 estimate), we ran a mixed-effects linear-regression model where the dependent variable was predicted temperature and the predictors were a fixed intercept, a fixed slope of vulnerability score, and per-county random intercepts and random slopes of vulnerability score (with no assumption of correlation between the random intercepts and random slopes). The unit of analysis was the tract, and the dependent variable was a spatially-weighted average of predicted temperature of the intersecting grid cells.

4. Results

4.1. Model comparisons and selection

We compared five different temperature models, each with a spatial resolution of 1 km and a temporal resolution of 1 h. Table 2 shows the accuracy of each model based on the RMSE, assessed with our spatial CV scheme, to predict temperature at the hot and cold hours in 2004 and 2013. As expected, all models have larger weighted error than unweighted error, since unweighted error emphasizes performance in data-rich areas. Models also tended to perform better in 2013 than 2004, despite similar baseline variability (as seen in the SD of the original observations), perhaps because more observations are available in later years.

Comparing the models to each other, we saw improvement with increasing complexity from IDW alone to OLS with many predictors, and likewise from OLS to the mixed model. The GAMM and XGBoost were best overall, and their RMSEs were little different from each other. However, the GAMM was much slower, taking 2 to 6 times as long to run as the equivalent XGBoost model. Given these results we used XGBoost for all subsequent analyses.

4.2. Full cross-validation

4.2.1. All years and hours

Table 3 shows the performance of XGBoost for every year-hour, aggregated by year. In every year, and in both weighted and unweighted evaluation, the RMSE was much smaller than the SD of the original observations, indicating high accuracy. Overall, our predictions demonstrated an average error less than 2 K. Performance improved over time, likely due to a tenfold increase in the number of observations available; the weighted RMSE decreased from 1.8 K in 2003 to 1.4 K in 2018. The mean of the yearly weighted RMSEs was 1.58 K R^2 values were consistently around 0.98.

Table 3 also shows the result of assessing NLDAS-2 predictions with the same observations as our model. We see that with or without weighting, in all years, the RMSEs are substantially higher than those of our model, failing to go below 2.3 K. The mean of the yearly weighted RMSEs was 2.46 K. Yearly weighted MSEs of NLDAS-2 were on average 2.5 times that of XGBoost. We also tried including the NLDAS-2 air temperature as a predictor in our XGBoost model, but found no meaningful improvement in RMSE, so we left it out to avoid the dependency (results not shown).

4.2.2. Hottest and coldest days

We also assessed performance during the hottest and coldest days of the study period, which may be particularly useful data for health studies of temperature extremes. Municipalities generally use temperature and humidity thresholds for heat-related action plans. For example, Washington, DC, has a threshold of 92 °F (33.3 °C; Homeland Security and Emergency Management Agency District of Columbia, 2020) for the air temperature or heat index. In our data, among the 2.3% of station-days (according to local time) in which an hourly temperature of at least 33.3 °C occurred, the mean yearly weighted SD was 5.55 K, while the mean weighted RMSE was 1.80 K for our model and 2.77 K for NLDAS-2 predictions. Among the 29% of station-days with an hourly temperature of at most 0 °C, the mean SD was 7.03 K and the mean RMSEs were 1.72 K for our model and 2.62 K for NLDAS-2 prediction. These results are largely consistent with the performance across all days.

4.2.3. Example time series

Fig. 3 is an example of how our hourly models capture within-day variation of temperature on a hot day. Shown are the observed temperature and the CV-predicted temperature for each hour of July 22, 2011, Eastern Daylight Time, at a MADIS station near Rochester, New York (longitude −75.4145, latitude 43.1122). We chose this station for having the median per-station RMSE for that day, namely 1.25 K, among all 1148 stations with an observation for every hour on that day.

4.2.4. Feature contributions

Fig. 4 shows feature contributions as SHAPs for all hours in 2013. Fig. 4(a) shows the distribution of SHAPs and the corresponding feature values (linearly rescaled to have minimum 0 and maximum 1) for the ten features with the greatest mean absolute SHAP (excluding the IDW feature, which was by far the most important feature, with a mean absolute SHAP of 8.65). Each vertical bar of each density plot is colored according to the mean of all feature values that have the corresponding range of SHAPs, while the height of the bar indicates the relative frequency of that range. We see several relationships that are to be expected, such as lower elevations (Fig. 4(b)) and lower latitudes having higher SHAPs. Fig. 4(c) shows that impervious surfaces contribute to predictions more during nighttime than daytime, with greater imperviousness having more positive SHAPs. Similarly, nighttime LST values appear to be more informative than daytime LST values. Table S1 lists the mean absolute SHAPs for all 34 variables, and Figure S3 shows the relationship between SHAPs and values for every predictor.

Table 2

Comparison of spatially cross-validated predictive accuracies (RMSEs) of models for the hot and cold hours in 2004 and 2013, spatiotemporally unweighted (Unw.) and weighted (W.). SDs and RMSEs are in kelvins.

	2004, hot hour		2004, cold hour		2013, hot hour		2013, cold hour	
Observations	151,231		147,500		1,090,722		1,070,886	
Stations	782		767		4153		4127	
	Unw.	W.	Unw.	W.	Unw.	W.	Unw.	W.
SD	10.85	11.24	9.93	10.30	10.74	11.25	9.45	9.85
IDW only	1.78	2.15	1.67	2.03	1.57	2.01	1.50	1.94
Linear regression	1.62	1.91	1.52	1.82	1.41	1.69	1.33	1.65
Mixed model	1.57	1.82	1.48	1.75	1.34	1.59	1.28	1.56
Generalized additive model	1.56	1.77	1.50	1.77	1.29	1.46	1.25	1.47
XGBoost	1.56	1.80	1.46	1.71	1.29	1.48	1.23	1.46

Table 3
Cross-validation results (temperature SD and RMSE, in kelvins) from XGBoost models (along with the RMSE from NLDAS-2 for comparison) for all year-hours.

			Unweighted				Weighted		
	Observations	Stations	R^2	SD	RMSE (XGBoost)	RMSE (NLDAS-2)	SD	RMSE (XGBoost)	RMSE (NLDAS-2)
2003	2,429,994	575	0.979	11.00	1.59	2.47	11.13	1.82	2.50
2004	3,599,525	796	0.980	10.75	1.50	2.42	11.11	1.74	2.46
2005	4,583,722	1003	0.981	11.05	1.53	2.45	11.31	1.81	2.52
2006	6,324,191	1267	0.975	9.64	1.52	2.42	9.99	1.76	2.49
2007	4,350,445	905	0.979	11.08	1.60	2.54	11.37	1.80	2.57
2008	6,942,368	1382	0.982	10.29	1.38	2.34	10.61	1.65	2.43
2009	10,616,378	1820	0.983	10.27	1.33	2.34	10.74	1.58	2.45
2010	12,373,631	2123	0.985	10.90	1.33	2.34	11.15	1.54	2.40
2011	13,508,410	2314	0.984	10.51	1.32	2.38	10.82	1.56	2.45
2012	22,167,011	4039	0.983	9.74	1.28	2.32	10.38	1.51	2.43
2013	25,994,855	4195	0.986	10.43	1.25	2.34	10.87	1.46	2.42
2014	25,225,773	4131	0.986	10.64	1.26	2.45	11.40	1.45	2.52
2015	26,962,107	4076	0.987	11.22	1.29	2.38	11.60	1.49	2.45
2016	28,198,019	4092	0.985	10.53	1.28	2.40	10.97	1.47	2.47
2017	28,086,992	4048	0.986	10.39	1.24	2.36	10.74	1.40	2.42
2018	26,025,033	3936	0.988	11.04	1.21	2.34	11.35	1.36	2.40
2019	23,641,305	3932	0.987	10.88	1.25	2.32	11.11	1.41	2.39

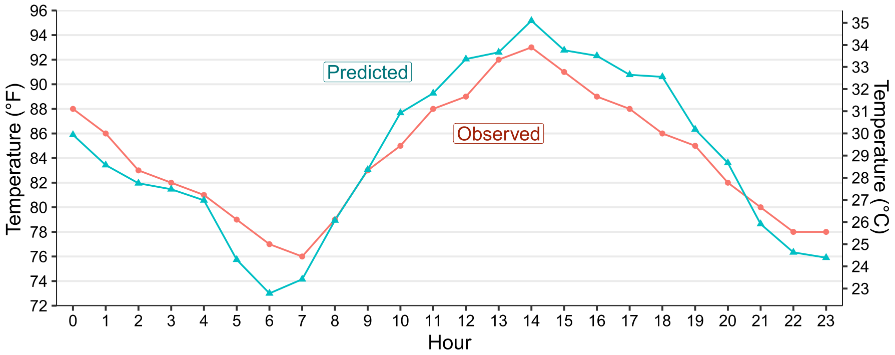

Fig. 3. Predicted and observed temperature for July 22, 2011 near Rochester, New York. Note that this station, like many others in MADIS, reports observed temperature in integer degrees Fahrenheit.

4.3. New air temperature predictions

4.3.1. Maps

Fig. 5 is an example of new air temperature predictions, made with the XGBoost model trained on all of a year's data. Fig. 5(a) depicts the entire study region at midnight EDT on July 22, 2011. Temperatures are clipped to [18 °C, 32 °C] for visibility; only 0.075% of the predictions fall outside this range. The coolest temperatures are in the northernmost part of the study area in Maine. Fig. 5(b) zooms in on Manhattan. An urban heat island is visible, particularly in the Bronx, Northern Manhattan, and parts of New Jersey. NLDAS-2 gridlines are shown in the zoomed-in map; our model predicts at a much finer resolution, capturing more spatial variability of temperature, which would have not been possible with NLDAS-2. Figure S4 and Figure S5 similarly show Boston and Washington, DC.

4.3.2. External validation

We predicted temperature at each of 14,257,658 observations closest to the hour at 2067 different Weather Underground stations in 2013. Aggregating across hours, the SD of the observations was 10.51 K unweighted and 10.73 K weighted, whereas the RMSE was 1.29 K unweighted and 1.39 K weighted. These results are similar to those obtained with our spatial CV with the 2013 MADIS data, supporting our stringent cross-validation strategy for model evaluation and providing evidence for the generalizability of our model. The results also support the quality of the Weather Underground data, so long as any data-quality problems in MADIS and Weather Underground are independent.

4.3.3. Application to social vulnerability

We used mixed models to characterize temperature at midnight EDT on July 22, 2011 as a function of social vulnerability in all 434 counties of the study area. When using our new temperature estimates, the fixed slope was 0.71 K (95% CI 0.60, 0.82) per unit difference in the social vulnerability index, compared to 0.18 K (95% CI 0.12, 0.25) when using NLDAS-2 temperatures. Thus, according to our XGBoost model, the most vulnerable census tracts were 0.71 K hotter at this time than the least vulnerable in a typical county. Fig. 6 shows the modeled associations for two counties in New York City and two in Upstate New York. The Upstate counties were chosen because they are distant from New York City and cover urban, suburban, and rural landscapes.

5. Discussion

We created a 1-km hourly air temperature model covering the Northeast and Mid-Atlantic US from 2003 to 2019 with the XGBoost machine-learning algorithm and a large quality-controlled dataset from ground stations, satellite remote sensing, and physiographic covariates. Our model performed best in 2018, with an unweighted RMSE of 1.21 K and a weighted RMSE of 1.36 K. The results from our novel spatially-weighted evaluation method demonstrate that failing to account for the non-uniform spatial distribution of observations can lead to overly optimistic estimates of model performance. Model performance was still good when subsetting to the hottest or coldest days in the study period, which are important for studies of the effects of extreme temperatures on health, and typically harder to reconstruct. The model performed well in an external validation at thousands of Weather Underground stations,

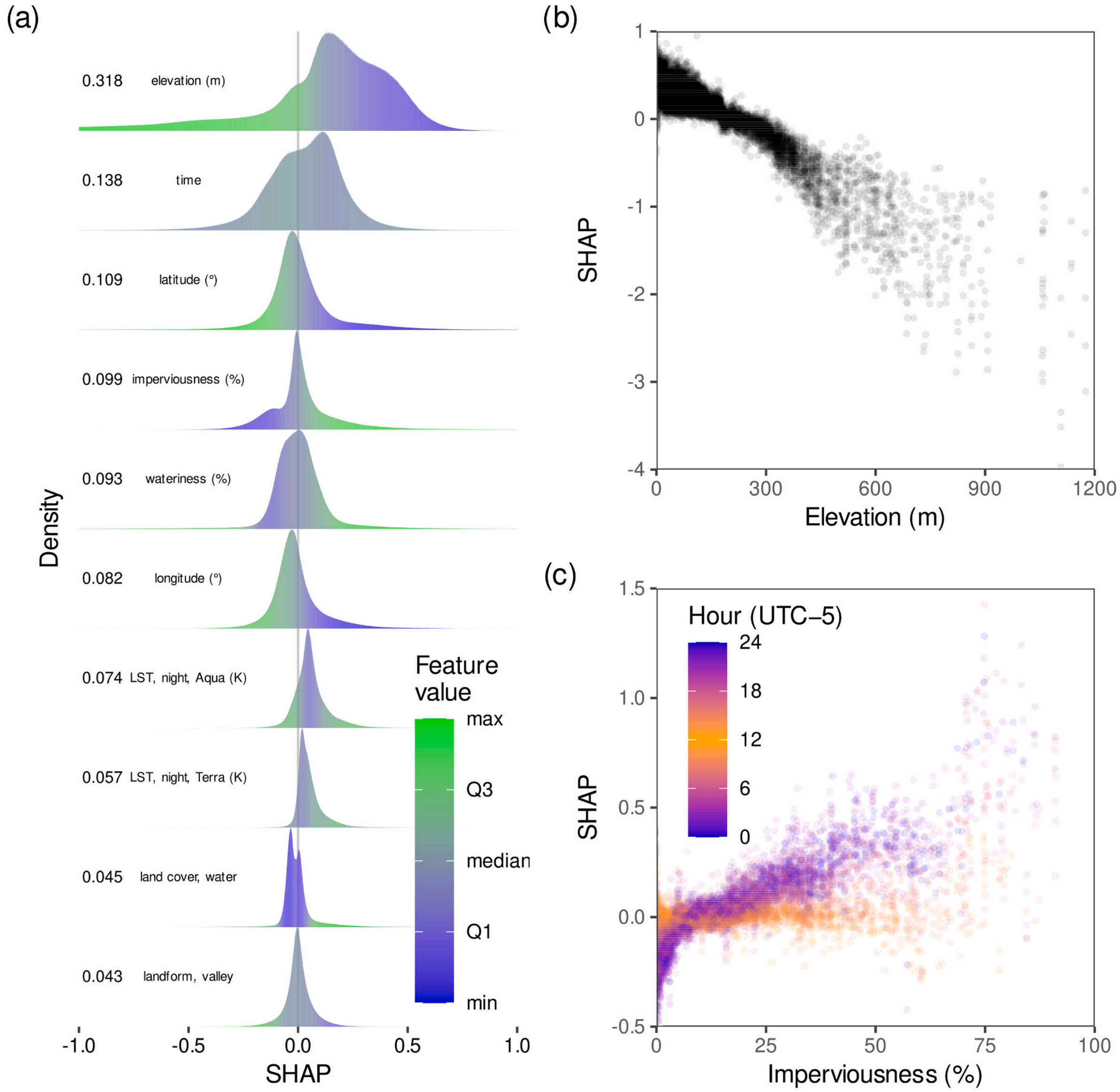

Fig. 4. Relationships between the SHAPs and the values of predictors for all hours in 2013. In our SHAP summary plot (a), predictors are sorted by mean absolute SHAP, with the greatest on top (except IDW, with a mean absolute SHAP of 8.65). The height of the bar indicates the relative frequency of that range and color indicates the feature value. In SHAP dependence plots for (b) elevation and (c) imperviousness, only 10,000 randomly selected predictions are plotted. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

providing further evidence for the generalizability of the model. We also used SHAP to show how various predictors contributed to the predictions, with IDW-interpolated temperature from other stations (carefully constructed to avoid leakage of testing data into model training) providing the largest SHAPs, followed by the elevation, time of year, and latitude. Finally, the example model application shows that midnight air temperature is positively associated with social vulnerability within counties during an extreme-heat event, demonstrating our model's potential for social science and human-health studies.

Several air-temperature models have been developed or used for health studies, including NLDAS-2 (Rowland et al., 2020; Wortzel et al., 2019; Wu et al., 2018). We found that our model's predictions were considerably more accurate than those of NLDAS-2. While NLDAS-2 is also hourly, its spatial resolution is coarse, so it misses substantial spatial heterogeneity, as shown in Fig. 5(b). If coarser resolution masks the association of higher temperatures with vulnerability, it may limit the utility of those models in public health planning and policymaking. And in fact, we found that NLDAS-2 was markedly less associated with social vulnerability within counties than our model was. Therefore, our model seems to reconstruct temperature heterogeneity associated with social deprivation and vulnerability, as seen in other studies (Buyantuyev and Wu, 2010; Hoffman et al., 2020). Accurate reconstruction of temperature profiles for vulnerable populations has important implications for energy insecurity (Hernández, 2013), heat action plans, and urban planning. By comparison, the Kloog et al. models operate at a finer 1-km spatial resolution like our model, but their temporal resolution is a daily average (Kloog et al., 2014, 2017). Consequently, they cannot capture within-day variation that may be important for certain acute health outcomes.

The high temporal and spatial resolutions of our model are among its many strengths. Our model balances accuracy and computational efficiency: XGBoost outperformed linear models including a random slopes approach used previously for daily temperature models (Kloog et al., 2014) and flexible penalized spatiotemporal smoothers. We took particular care to avoid leakage of test data (such as in our construction of IDW surfaces) and overfitting (with a spatial CV scheme). Ensemble

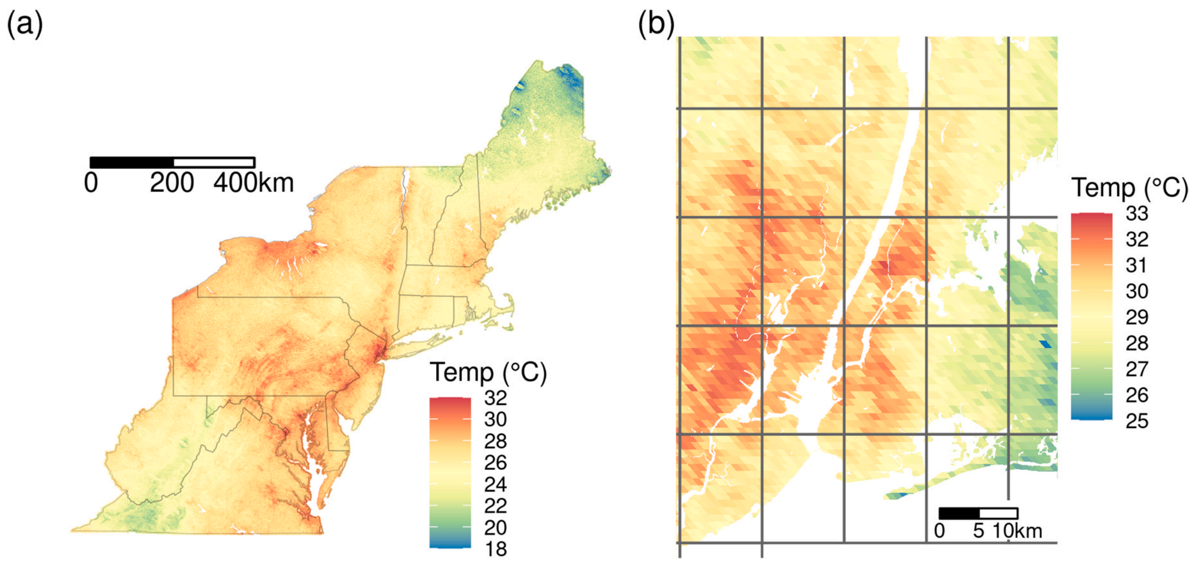

Fig. 5. Maps of 1-km temperature predictions at midnight EDT on July 22, 2011. Water bodies are masked out from the images. Grid cells appear nonrectangular because they have been reprojected from the MODIS sinusoidal projection to plate carrée. (a) Showing the entire study region of 750,808 cells. (b) Panned to Manhattan and enlarged. Overlaid gridlines show NLDAS-2 cells that are roughly 11 × 14 km in extent.

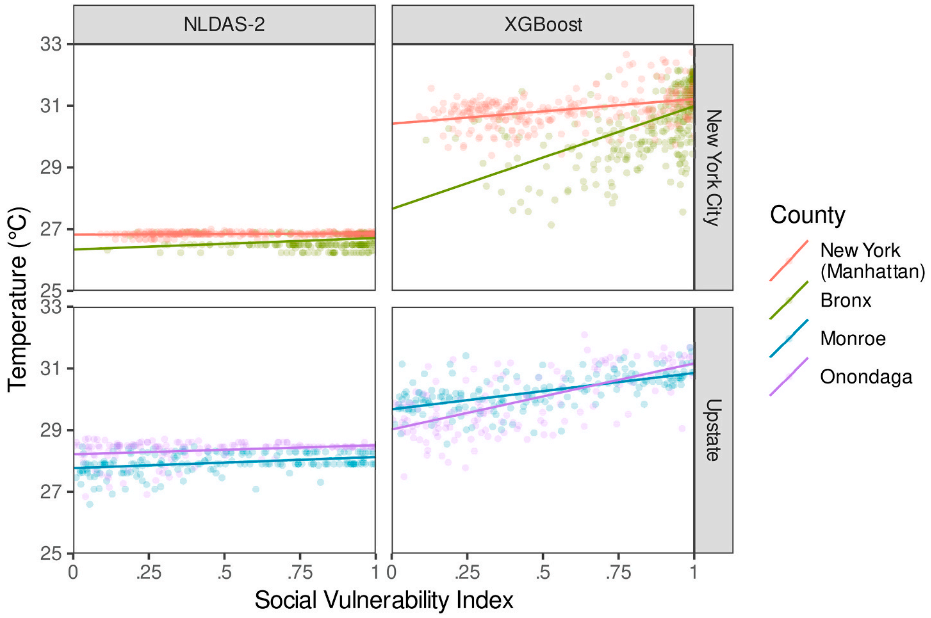

Fig. 6. Predicted temperature per census tract versus social vulnerability index in four counties. Each point represents a census tract and its temperature at midnight EDT on July 22, 2011 based on NLDAS-2 or our XGBoost model.

methods offer a way to combine the outputs of different prediction models to improve predictive performance, but ensemble modeling for spatiotemporal data requires yet more complexity to avoid leakage. We employed a spatial CV scheme with weighted model evaluation that, to our knowledge, is novel for air-temperature models. The spatial CV is particularly important to ensure that no prior information from sites and nearby exclusion zones is used in model training. This approach is relevant for health studies, since we aim to predict temperatures at times and places for which we do not have data, such as unmonitored residential addresses. When we compared unweighted and weighted results, we found that weighted RMSEs were consistently higher. This implies that the standard unweighted approach yields overly optimistic performance metrics for out-of-sample predictions due to the assumption that all observations are equally informative for assessment. Nonetheless, our RMSEs are quite small overall with an unweighted mean square error (MSE) from our XGBoost predictions that is 1/3 of the MSE of NLDAS-2 temperatures, averaging across all years. Our hourly 1-km model may be particularly useful in cities because the spatial and temporal resolution can capture the heterogeneity of the urban heat archipelago. We found that the contribution of imperviousness to our model's predictions is greatest at night, the period when other studies have found the urban heat island effect is most pronounced (Oyler et al., 2016). The high spatial resolution of our model makes it ideal for risk assessment applications, including assessment of exposure disparities, and in environmental health studies. Finally, the model's annual construction makes it easy to update for new years: models for previous years do not have to be refit.

A limitation of annual fitting is that our model cannot learn patterns across years. We utilize LST from four daily overpasses from the two MODIS instruments but are limited by data availability to the period from 2003 onward. Therefore, our model is not ideal for long-term trends. Another limitation is that our model uses area-level predictors to make predictions at point locations, the MADIS weather stations. This design assumes that the station is representative of the mean temperature of the 1-km grid cells, which may contribute to our modest prediction error. Additionally, our model measures temperature alone, and many health studies are interested in assessing the role of apparent temperature in adverse health outcomes. We are extending the methods of this approach and adding satellite-based retrievals of column water vapor to construct ambient humidity-prediction models for future health applications (Just et al., 2020a, 2020b). Lastly, our analysis of social vulnerability was only conducted for the temperature at a single hour of a single heat wave. It was only intended for expository purposes. Future analyses should consider longer-term relationships with social vulnerability and health.

6. Conclusion

We compared five approaches to create an extensible, flexible, and accurate air temperature model. Ultimately, we used the XGBoost algorithm to create hourly 1-km predictions across the Northeast and Mid-Atlantic US from 2003 to 2019. These predictions can play a pivotal role in future applications to social science and human health.

Funding sources

This work was supported by National Institute of Environmental Health Sciences grants P30 ES023515, P30 ES009089, R00 ES023450, R01 ES028805, R01 ES030616, and US-Israel Binational Science Foundation (BSF) grant 2017277. DC was funded by NICHD T32HD049311. The funders had no role in the design, analysis, or reporting of this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2021.111477.

References

Buyantuyev, A., Wu, J., 2010. Urban heat islands and landscape heterogeneity: linking spatiotemporal variations in surface temperatures to land-cover and socioeconomic patterns. Landsc. Ecol. 25, 17–33.

Centers for Disease Control & Prevention, 2020. Heat & health tracker [WWW document]. About the data. https://ephtracking.cdc.gov/Applications/heatTracker/. (Accessed 12 January 2020).

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794. https://doi.org/10.1145/2939672.2939785.

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.

developers, XGBoost, 2020. XGBoost Parameters — Xgboost 1.3.0 [WWW Document]. URL (accessed 11.25.2020). https://xgboost.readthedocs.io/en/latest/parameter.html.

Didan, K., 2015. MOD13A3 MODIS/Terra Vegetation Indices Monthly L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD13A3.006.

Dirksen, M., Knap, W.H., Steeneveld, G.-J., Holtslag, A.A., Tank, A.M.K., 2020. Downscaling daily air-temperature measurements in The Netherlands. Theor. Appl. Climatol. 142, 751–767.

Eliasson, I., 1996. Urban nocturnal temperatures, street geometry and land use. Atmos. Environ. 30, 379–392.

Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., Alsdorf, D., 2007. The Shuttle radar topography mission. Rev. Geophys. 45 https://doi.org/10.1029/2005RG000183.

Flanagan, B.E., Gregory, E.W., Hallisey, E.J., Heitgerd, J.L., Lewis, B., 2011. A social vulnerability index for disaster management. J. Homel. Secur. Emerg. Manag. 8 https://doi.org/10.2202/1547-7355.1792.

Flanagan, B.E., Hallisey, E.J., Adams, E., Lavery, A., 2018. Measuring community vulnerability to natural and anthropogenic hazards: the centers for Disease control and Prevention's social vulnerability index. J. Environ. Health 80, 34.

Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., Tobias, A., Tong, S., Rocklöv, J., Forsberg, B., Leone, M., De Sario, M., Bell, M.L., Guo, Y.-L.L., Wu, C.-f., Kan, H., Yi, S.-M., de Sousa Zanotti Stagliorio Coelho, M., Saldiva, P.H.N., Honda, Y., Kim, H., Armstrong, B., 2015. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. Lancet 386, 369–375. https://doi.org/10.1016/S0140-6736(14)62114-0.

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., Tyler, D., 2002. The national elevation dataset. Photogramm. Eng. Rem. Sens. 68, 5–32.

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 2003-2019. Int. J. Climatol. 1–17 https://doi.org/10.1002/joc.7060.

Hart, M.A., Sailor, D.J., 2009. Quantifying the influence of land-use and surface characteristics on spatial variability in the urban heat island. Theor. Appl. Climatol. 95, 397–406.

Hernández, D., 2013. Energy insecurity: a framework for understanding energy, the built environment, and health among vulnerable populations in the context of climate change. Am. J. Publ. Health 103, e32–e34. https://doi.org/10.2105/AJPH.2012.301179.

Hernández, D., 2016. Understanding 'energy insecurity'and why it matters to health. Soc. Sci. Med. 167, 1–10.

Hoffman, J.S., Shandas, V., Pendleton, N., 2020. The effects of historical housing policies on resident exposure to intra-urban heat: a study of 108 US urban areas. Climate 8, 12.

Homeland Security and Emergency Management Agency District of Columbia, 2020. Heat emergency plan information [WWW document]. https://hsema.dc.gov/page/heat-emergency-plan-information. (Accessed 12 March 2020).

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J., Megown, K., 2015. Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. Photogramm. Eng. Rem. Sens. 81, 345–354.

Ito, K., Lane, K., Olson, C., 2018. Equitable Access to air conditioning: a city health department's perspective on preventing heat-related deaths. Epidemiology 29, 749–752. https://doi.org/10.1097/EDE.0000000000000912.

Just, A.C., Arfer, K.B., Rush, J., Dorman, M., Shtein, A., Lyapustin, A., Kloog, I., 2020a. Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter (PM2. 5) using satellite data over large regions. Atmos. Environ. 239, 117649.

Just, A.C., Liu, Y., Sorek-Hamer, M., Rush, J., Dorman, M., Chatfield, R., Wang, Y., Lyapustin, A., Kloog, I., 2020b. Gradient boosting machine learning to improve satellite-derived column water vapor measurement error. Atmospheric measurement techniques 13, 4669–4681.

Kloog, I., Nordio, F., Coull, B.A., Schwartz, J., 2014. Predicting spatiotemporal mean air temperature using MODIS satellite surface temperature measurements across the Northeastern USA. Rem. Sens. Environ. 150, 132–139. https://doi.org/10.1016/j.rse.2014.04.024.

Kloog, I., Nordio, F., Lepeule, J., Padoan, A., Lee, M., Auffray, A., Schwartz, J., 2017. Modelling spatio-temporally resolved air temperature across the complex geo-climate area of France using satellite-derived land surface temperature data: modelling air temperature IN France. Int. J. Climatol. 37, 296–304. https://doi.org/10.1002/joc.4705.

Lin, S., Luo, M., Walker, R.J., Liu, X., Hwang, S.-A., Chinery, R., 2009. Extreme high temperatures and hospital admissions for respiratory and cardiovascular diseases. Epidemiology 20, 738–746. https://doi.org/10.1097/EDE.0b013e3181ad5522.

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. Nature Machine Intelligence 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Presented at the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. Oakland, CA, USA.

Madrigano, J., Ito, K., Johnson, S., Kinney, P.L., Matte, T., 2015. A case-only study of vulnerability to heat wave–related mortality in New York City (2000–2011). Environ. Health Perspect. 123, 672–678.

Napoly, A., Grassmann, T., Meier, F., Fenner, D., 2018. Development and application of a statistically-based quality control for crowdsourced air temperature data. Front. Earth Sci. 6, 118.

NOAA, 2018. Meteorological assimilation data ingest System (MADIS) [WWW document]. National oceanic and atmospheric administration (accessed 8.17.2020). https://madis.ncep.noaa.gov/.

NOAA, 2020. State climate extremes committee (SCEC) | extremes | national centers for environmental information (NCEI) [WWW document]. National oceanic and atmospheric administration (accessed 8.24.2020). https://www.ncdc.noaa.gov/extremes/scec/records.

Oke, T.R., 1982. The energetic basis of the urban heat island. Q. J. R. Meteorol. Soc. 108, 1–24.

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.

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. Journal of Applied Meteorology and Climatology 55, 1441–1457.

Quagliolo, C., Pezzoli, A., Ignaccolo, R., Davila, J.L.S., 2020. Time-lagged inverse-distance weighting for air temperature analysis in an equatorial urban area (Guayaquil, Ecuador). Meteorol. Appl. 27, e1938.

Rowland, S.T., Boehme, A.K., Rush, J., Just, A.C., Kioumourtzoglou, M.-A., 2020. Can ultra short-term changes in ambient temperature trigger myocardial infarction? Environ. Int. 143, 105910.

Rui, H., Mocko, D., 2019. Readme Document for North American Land Data Assimilation System Phase 2 (NLDAS-2) Products.

Samanta, S., Pal, D.K., Lohar, D., Pal, B., 2012. Interpolation of climate variables and temperature modeling. Theor. Appl. Climatol. 107, 35–45.

Shi, L., Kloog, I., Zanobetti, A., Liu, P., Schwartz, J.D., 2015. Impacts of temperature and its variability on mortality in New England. Nat. Clim. Change 5, 988–991. https://doi.org/10.1038/nclimate2704.

Stein, M., 1987. Large sample properties of simulations using Latin hypercube sampling. Technometrics 29, 143–151.

Teixeira, J., 2013. AIRS/Aqua L2 standard physical retrieval (AIRS-only) V006. Goddard Earth Sciences Data and Information Services Center. https://doi.org/10.5067/Aqua/AIRS/DATA202.

The Weather Company, 2018. Weather underground [WWW document]. Weather Underground. www.wunderground.com.

Theobald, D.M., Harrison-Atlas, D., Monahan, W.B., Albano, C.M., 2015. Ecologically-relevant maps of landforms and physiographic diversity for climate adaptation planning. PloS One 10, e0143619. https://doi.org/10.1371/journal.pone.0143619.

Vicedo-Cabrera, A.M., Forsberg, B., Tobias, A., Zanobetti, A., Schwartz, J., Armstrong, B., Gasparrini, A., 2016. Associations of inter-and intraday temperature change with mortality. Am. J. Epidemiol. 183, 286–293.

Vinayak, R.K., Gilad-Bachrach, R., 2015. Dart: Dropouts meet multiple additive regression trees. Presented at the Artificial Intelligence and Statistics. PMLR, pp. 489–497.

Wan, Z., Hook, S., Hulley, G.C., 2015a. MOD11A1 MODIS/Terra land surface temperature/emissivity daily L3 global 1km SIN grid V006. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD11A1.006.

Wan, Z., Hook, S., Hulley, G.C., 2015b. MYD11A1 MODIS/Aqua land surface temperature/emissivity daily L3 global 1km SIN grid V006. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MYD11A1.006.

Wood, S.N., 2006. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. Biometrics 62, 1025–1036.

Wortzel, J., Norden, J., Turner, B., Haynor, D., Kent, S., Al-Hamdan, M., Avery, D., Norden, M., 2019. Ambient temperature and solar insolation are associated with decreased prevalence of SSRI-treated psychiatric disorders. J. Psychiatr. Res. 110, 57–63.

Wu, C.Y., Zaitchik, B.F., Gohlke, J.M., 2018. Heat waves and fatal traffic crashes in the continental United States. Accid. Anal. Prev. 119, 195–201.

Yu, W., Tan, J., Ma, M., Li, X., She, X., Song, Z., 2019. An effective similar-pixel reconstruction of the high-frequency cloud-covered areas of southwest China. Rem. Sens. 11, 336. https://doi.org/10.3390/rs11030336.

Zhang, Y., Yu, C., Wang, L., 2017. Temperature exposure during pregnancy and birth outcomes: an updated systematic review of epidemiological evidence. Environ. Pollut. 225, 700–712. https://doi.org/10.1016/j.envpol.2017.02.066.