




OPEN

DATA DESCRIPTOR

# An enhanced national-scale urban tree canopy cover dataset for the United States

Lucila M. Corro<sup>1</sup> , Kenneth J. Bagstad<sup>1</sup>, Mehdi P. Heris<sup>2</sup>, Peter C. Ibsen<sup>1</sup>, Karen G. Schleeweis<sup>3</sup>, Jay E. Diffendorfer<sup>1</sup>, Austin Troy<sup>4</sup>, Kevin Megown<sup>5</sup> & Jarlath P. M. O'Neil-Dunne<sup>6,7</sup>

Moderate-resolution (30-m) national map products have limited capacity to represent fine-scale, heterogeneous urban forms and processes, yet improvements from incorporating higher resolution predictor data remain rare. In this study, we applied random forest models to high-resolution land cover data for 71 U.S. urban areas, moderate-resolution National Land Cover Database (NLCD) Tree Canopy Cover (TCC), and additional explanatory climatic and structural data to develop an enhanced urban TCC dataset for U.S. urban areas. With a coefficient of determination ( $R^2$ ) of 0.747, our model estimated TCC within 3% for 62 urban areas and added 13.4% more city-level TCC on average, compared to the native NLCD TCC product. Cross validations indicated model stability suitable for building a national-scale TCC dataset (median  $R^2$  of 0.752, 0.675, and 0.743 for 1,000-fold cross validation, urban area leave-one-out cross validation, and cross validation by Census block group median year built, respectively). Additionally, our model code can be used to improve moderate-resolution TCC in other parts of the world where high-resolution land cover data have limited spatiotemporal availability.

## Background & Summary

Urban areas contain heterogeneous mixes of land cover types, which are driven by and influence complex social-ecological processes<sup>1–3</sup>. The spatial dynamics of urban trees and forests directly affect multiple urban ecosystem services, such as cooling<sup>4,5</sup>, pollution mitigation<sup>6</sup>, access to green space<sup>7</sup>, water regulation<sup>8</sup>, carbon sequestration<sup>7</sup>, and noise mitigation<sup>9</sup>. Many analyses in urban systems rely on the use of remotely sensed products depicting tree canopy cover (TCC). However, the accuracy of these products to represent the size and shape of urban forest patches varies by geography<sup>7,10</sup>.

Increasing the resolution, consistency, accuracy, and availability of multi-city to national-scale land cover and TCC data is of great interest to researchers, policy makers, and urban planners. Multi-city to continental approaches quantifying urban phenomena for hundreds of cities simultaneously have become quite prominent in the past decade, for example, examining land cover drivers of urban heat in Europe<sup>11</sup> and tree canopy-derived ecosystem services in U.S. cities<sup>12,13</sup>. Some multi-city studies have been completed using high-resolution land cover data to analyze distributions of ecosystem services<sup>14</sup>, but due to limited accessibility and standardization of the various high-resolution datasets, most recent work incorporating high-resolution data has focused on individual cities or regions. The greater precision provided by higher-resolution data is critical for multi-city to continental-scale studies as finer-grained data can often provide better estimates of land cover or tree canopy-derived urban ecosystem services compared to coarser satellite data<sup>10</sup>.

However, researchers generally face tradeoffs between moderate resolution (e.g., 10- to 30-m) TCC products suitable for long-term monitoring and high-resolution (e.g., 1- to 3-m or finer) products typically available for limited spatiotemporal extents or produced by commercial providers (i.e., not in the public domain). In this

<sup>1</sup>U.S. Geological Survey, Geosciences and Environmental Change Science Center, Denver, CO, 80225, USA. <sup>2</sup>Hunter College, Urban Policy & Planning, New York, NY, 10065, USA. <sup>3</sup>Forest Inventory and Analysis, U.S. Forest Service, Rocky Mountain Research Station, Riverdale, UT, 84405, USA. <sup>4</sup>College of Architecture and Planning, University of Colorado Denver, University of Colorado Denver, Denver, CO, 80202, USA. <sup>5</sup>Geospatial Technology and Applications Center, U.S. Forest Service, National Forest System, Salt Lake City, UT, 84138, USA. <sup>6</sup>Spatial Analysis Laboratory, Rubenstein School of Environment & Natural Resources, University of Vermont, Burlington, VT, 05405, USA.

<sup>7</sup>Deceased: Jarlath P. M. O'Neil-Dunne. ✉e-mail: [lcorro@usgs.gov](mailto:lcorro@usgs.gov)

study, we aim to bridge this gap, using limited-availability, high-resolution TCC data, more widely available moderate-resolution TCC data, ancillary data, and machine learning methods to substantially improve the quality of moderate-resolution TCC data in the United States.

Key moderate-resolution TCC products include those derived from the European Space Agency's Copernicus Sentinel-2 satellite and the U.S. Landsat program. The U.S. National Land Cover Database (NLCD), produced by the interagency Multi-Resolution Land Characteristics consortium, is a national Landsat-derived dataset for the United States, and includes both thematic land cover and continuous TCC and impervious cover<sup>15,16</sup>. These data encompass the conterminous United States, coastal Alaska, Hawaii, Puerto Rico, and the U.S. Virgin Islands. The thematic and percent impervious data are available in 2- to 3-year intervals from 2001–2021 and percent TCC data are available annually from 2011–2021, both at 30-m resolution. Substantial work has improved these products and their time series<sup>17,18</sup>. Meanwhile, Sentinel-2 data offer higher spatial resolution than Landsat (10 versus 30 m), and support a growing number of global thematic and continuous land cover products<sup>19</sup>. However, Sentinel-2 products have a shorter time series (dating to 2017) than Landsat-derived NLCD products (dating to 2001). This shorter historical time series is a major limitation in monitoring long-term land cover changes. Moreover, although the 10-m Sentinel-derived data show improvements over recent Landsat sensors for continuous canopy metrics<sup>20</sup>, it remains unclear whether 10 m is an adequate spatial resolution for understanding fine-grained spatial patterns of land and tree cover in urban areas<sup>21</sup>.

Overall, studies highlight how urban developed areas have an inherent spatial pattern whose fine scale cannot adequately be captured by 30-m-resolution spatial grain<sup>22</sup>. Because canopy cover does not scale linearly, 30-m scale observations measure different phenomena than 10-m or 1-m observations, leading to one potentially substantial source of uncharacterized error when urban TCC maps are generated from only 30-m imagery<sup>10,23</sup>. An optimal resolution for urban areas mapping has not been quantified, but the consensus is the higher the spatial resolution, the better to capture the inherent heterogeneity<sup>24</sup>. Forest fragmentation pattern metrics calculated using high-resolution input data demonstrate improved spatial precision of pattern indices in complex heterogeneous forests, including both urban and non-urban forests<sup>17</sup>.

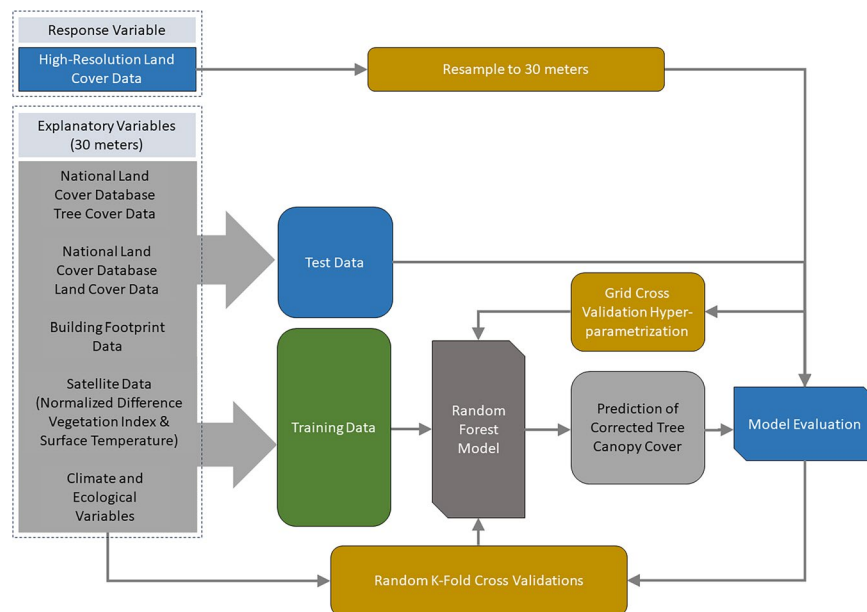
By contrast, high-resolution (e.g., 1- to 3-m) urban land cover and TCC products are increasingly available for individual cities, produced by government agencies<sup>25</sup>, academic researchers<sup>26,27</sup>, and private companies<sup>28,29</sup>. Consequently, there is no central repository for such data, and some are not publicly available. Along with uneven access, these datasets are inconsistent in their production methodology, metadata, time of measurement, and the land cover classification schema. Important advances are also being made in public high-resolution ( $\leq 5$  m) regional-scale land cover and TCC products, such as the National Aeronautics and Space Administration's (NASA) Carbon Monitoring System<sup>30</sup>, the Chesapeake Bay Program<sup>31</sup>, and National Oceanic and Atmospheric Administration's (NOAA) high-resolution Coastal Change Analysis Program<sup>32</sup>. Recent work has produced a TCC dataset for U.S. urban areas at 2-m spatial resolution, but this dataset is less accurate than previously mentioned products and intended to support Census block-scale, but not pixel-scale, analyses<sup>33</sup>. Other recent examples include a nationwide 0.5-m TCC dataset for 472 cities in Brazil<sup>34</sup> and global-scale 1-m tree canopy height data<sup>35</sup>. Such novel approaches currently remain constrained by limited spatial and/or temporal coverage. This reduces their value for multi-city monitoring over time, a limitation that can be addressed by pairing high-resolution TCC data with limited spatiotemporal coverage with moderate-resolution data with broad spatiotemporal coverage.

In this paper, we develop an approach to substantially improve the quality of moderate-resolution TCC data through the use of high-resolution TCC data and other variables. The product described in this paper built on recent work using decision tree machine learning models to improve a moderate-resolution national TCC product, the NLCD TCC dataset, using NLCD land cover and TCC, high-resolution land cover data for 27 U.S. cities and counties, and moderate-resolution environmental correlate data describing urban form and climate<sup>10</sup>. Our work had three goals. First, we developed methods to further improve the performance of this previous U.S. urban TCC correction model using more urban areas that span wider gradients of climate and city sizes, using random forest models (Fig. 1). Second, we produced an enhanced moderate-resolution (i.e., 30-m), single-year (circa 2011) national urban TCC product that can be used as a stand-alone data product or incorporated into future NLCD TCC products to generate time series data across U.S. urban areas. Specifically, as NLCD products continue to evolve, these methods could be integrated into the future NLCD TCC production process within urban areas (similar work could investigate the possibility to improve TCC data for non-urban forests, using different predictor variables). Third, we provided model code to enable replication of the approach in the United States or other parts of the world where high-resolution land cover data can be combined with coarser resolution TCC data (e.g., Landsat or Sentinel-derived) and additional explanatory variables to produce an enhanced urban TCC product. Such data can support improved modelling of various urban socio-environmental phenomena that rely on an understanding of TCC and its changes over time.

## Methods

We applied random forest models to TCC data derived from high-resolution land cover products, along with additional explanatory variables to model urban TCC for 71 U.S. urban areas at 30-m resolution (Fig. 1). We then validated the model and applied its predictions to produce an enhanced urban TCC dataset for all urban areas in the conterminous United States. We describe the data used, how we defined and selected urban areas, the random forest model, and validation approaches below. Throughout this paper, we refer to comparisons between (1) native (i.e., unaltered) NLCD TCC data, (2) upscaled TCC data derived from high-resolution data as described below, and (3) enhanced TCC data, the output of our random forest model.

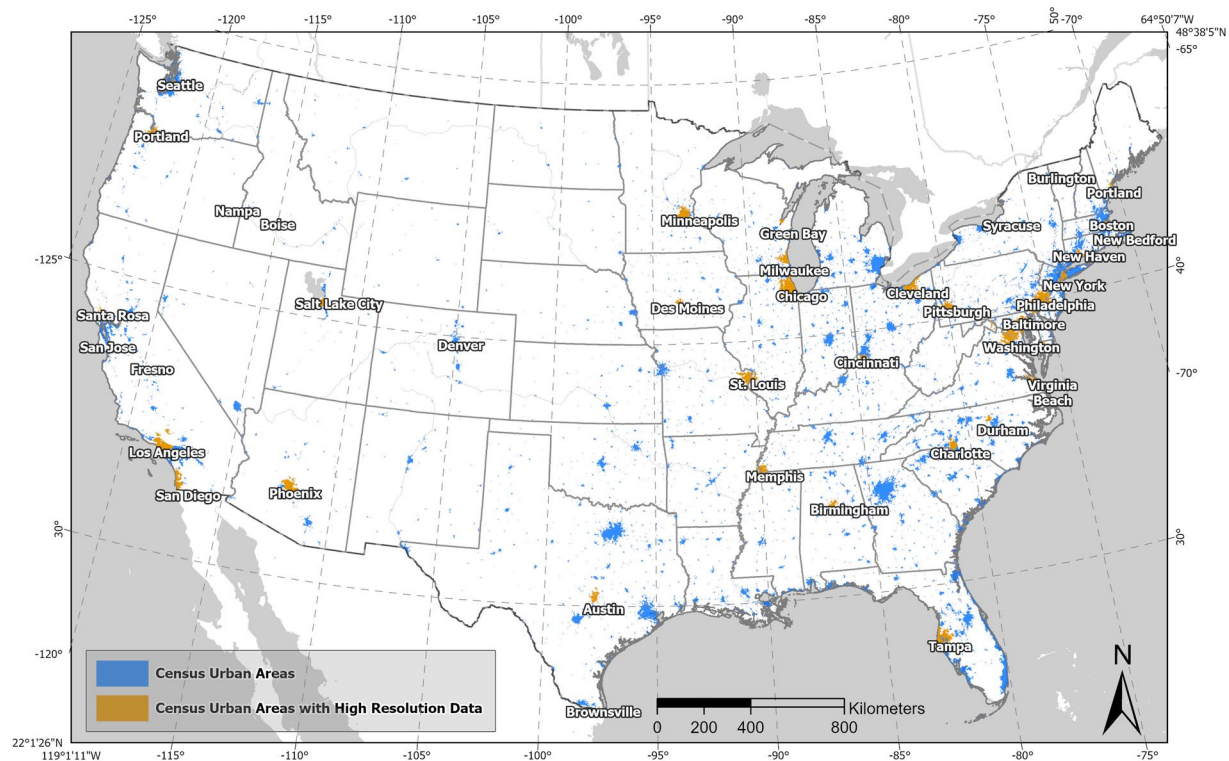
**Data sources.** We used two different sources of high-resolution land cover data for 71 U.S. urban areas as a response variable in the random forest model. The high-resolution data used in this paper cover a range of



**Fig. 1** Conceptual diagram of analytical methods for corrected tree canopy cover modelling, incorporating the high-resolution land cover response variable, explanatory variables, random K-fold cross-validation for training data, grid search cross-validation for hyperparameter tuning, and continuous model evaluation.

geographies from small towns to the nation's largest cities, as well as countywide data for several urban or suburban counties, and some data covering multi-county urban areas. For simplicity and consistency, we refer to these all as urban area datasets. High-resolution land cover data for 39 urban areas came from the U.S. Environmental Protection Agency (EPA) EnviroAtlas<sup>25</sup> with a spatial resolution of 1 m. Data for the remaining 32 urban areas was sourced from the University of Vermont's (UVM) Spatial Analysis Laboratory with a spatial resolution of 1 m or less<sup>36</sup>. Formal accuracy assessments for UVM New York City and Philadelphia tree canopy data had 98% and 97% user's accuracy, respectively. While accuracy assessments were completed only for selected cities, because a similar procedure was used to produce data for all cities, including manual correction, tree cover data for other cities are expected to have similar levels of accuracy. Prior to running the random forest model, we resampled the categorical 1-m land cover data to represent 30-m continuous upscaled TCC data and reprojected all data into a common coordinate reference system (ESRI:102039). Any 1-m pixel of tree canopy, woody wetland, and/or orchard was considered "tree canopy" and aggregated to 30 m to estimate percent TCC. Given our focus on urban TCC, we clipped all high-resolution datasets to the U.S. Census Bureau's 2020 urban area boundaries<sup>37,38</sup>. For five counties – Los Angeles and Sonoma, California, Anne Arundel and Montgomery, Maryland, and Jefferson, West Virginia – this meant that we split a single county-level dataset into multiple polygons aligned with Census urban areas. EPA data typically extend well beyond the boundaries of a single city, whereas UVM data typically cover smaller extents matching individual city boundaries. Although data for 27 cities and counties similarly analyzed by Heris *et al.*<sup>10</sup> were clustered in the Northeast and Midwestern United States, our high-resolution data provided coverage for an additional 23 urban areas from other parts of the United States, spanning wider gradients of climate, city size, and city age. We reached a total of 71 urban areas as the sample size in this study because (1) there were some duplicate cities/urban areas provided by both EPA and UVM data and (2) some of the original county/multicounty high-resolution datasets were split into multiple urban areas when clipping them using Census urban area boundaries. These high-resolution data covered a total area of 143,273 km<sup>2</sup>, which represents about 52.3% of Census urban areas.

We used a total of 14 explanatory variables (Fig. 1) that act as biophysical and structural drivers of urban tree canopy in our machine learning model, with all datasets assembled for each urban area, aligned, and snapped to a common 30-m raster grid. To represent high temporal resolution (i.e., annual) changes in urban structure, we used (1) thematic land cover, and continuous (2) impervious surface, (3) tree canopy data, and (4) tree canopy standard error, all from NLCD for map year 2011 (2019 NLCD release)<sup>39</sup>. Other structural influences on tree canopy included building footprint data for (5) total building footprint coverage per cell, (6) number of buildings that intersect each cell, and (7) area of the average building per cell<sup>40</sup>. We sourced building data from Microsoft, which were developed using aerial imagery from varying time intervals, so are not associated with a specific year<sup>40</sup>. Additionally, we incorporated (8) Census block group-scale data on median year built of structures to incorporate the effect of neighborhood age on TCC. Collectively, these structural parameters represent how urban form can either increase (through plantings) or decrease (through mortality) TCC<sup>41–43</sup>. Six additional explanatory variables represent biophysical and structural drivers of urban tree canopy, including remotely sensed data for (9) surface temperature and (10) Normalized Difference Vegetation Index (NDVI), both derived from cloud-free summer Landsat 8 imagery for the years 2013–2015, mean urban area climate data from PRISM Gridded Climate Data<sup>44</sup> including (11) average high temperature of the month of August (°C), (12)



**Fig. 2** Distribution of urban areas across the contiguous United States, differentiated by standard Census Urban Areas (273,844 km<sup>2</sup>) and those with high-resolution data available (143,273 km<sup>2</sup>). Some small urban areas are not visible at the national scale, and due to the clustering of some urban areas, not all 71 urban areas are labelled on this map.

lowest temperature for the month of January (°C) and (13) annual precipitation (mm), and (14) EPA level II ecoregions for each urban area<sup>45</sup>. These data represent biophysical and climatic controls on tree growth and function.

**Urban areas definition.** The final urban TCC dataset follows the boundaries of the Census 2020 Urban Areas Dataset, an area covering 273,844 km<sup>2</sup>, or about 3.6% of the conterminous United States (Fig. 2)<sup>38</sup>. The Census Urban Areas Dataset provides a nationally consistent definition of urban form, incorporating both population density and land use characteristics, enabling us to avoid potential discrepancies that might arise from using multiple, locale-specific definitions.

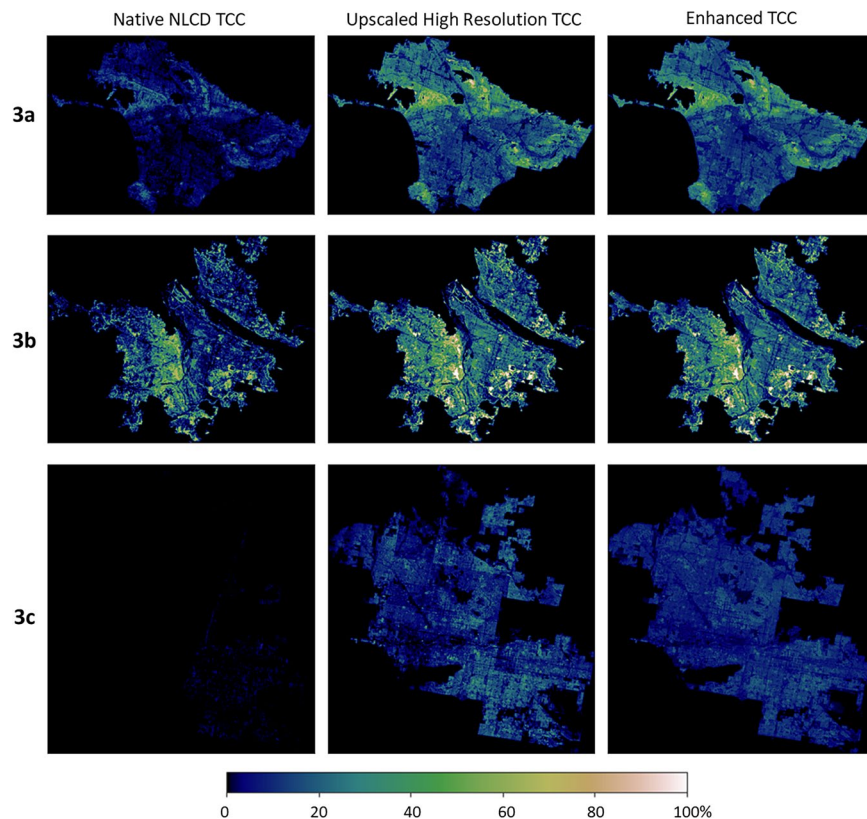
We used the updated Census urban areas dataset developed following the 2020 Census, which redefines urban areas, maintaining a focus on population density and developed land while incorporating several key changes<sup>38</sup>. The Bureau now defines urban areas as regions with a population of at least 2,500, a noteworthy reduction from the previous threshold of 50,000. Another important change is the introduction of a housing unit density threshold, replacing the previous impervious surface criterion for defining urbanized regions. The Bureau has also added a proximity criterion, which designates areas with at least 385 housing units per square mile near urban areas as part of an adjacent urban areas. These modifications offer a more nuanced perspective on urbanization, encompassing not only dense city centers but also less dense, yet substantially urbanized suburban and exurban regions.

**Machine learning algorithms and tuning.** We developed random forest models and predictions using the Python 'sklearn' package version 1.0.2 and the 'statsmodels' package version 0.13.2<sup>46,47</sup>. To assess the model's predictive performance, we allocated 70% of the data for training and 30% for testing. Our models built on existing random forest and decision tree models for urban TCC<sup>10</sup>, we used a common hyperparameter tuning workflow to improve random forest model predictions and include additional urban areas<sup>48</sup>. To optimize our random forest model, we used both Random Search and Grid Search cross-validation methods to create and test a series of models with varying hyperparameters to optimize model predictions. We then selected the optimal hyperparameters based on the results of the hyperparameter tuning to develop our best random forest model. We assessed models during tuning using the accuracy metrics Root Mean Squared Error and the proportion of variance (R<sup>2</sup>).

### Data Records

The final data product is an enhanced 30 m resolution raster dataset for urban TCC for all Census Urban Areas within the conterminous U.S., represented by values from 0–100, with a nominal 2011 date (i.e., built using NLCD map data year 2011, high-resolution land cover data built using imagery from 2004–2016, and ancillary data described above). The 2020 Census Urban Areas dataset, and our TCC data product which covers it, includes a total area of 273,844 km<sup>2</sup>. As an example, Fig. 3 shows the native 2011 NLCD TCC dataset, upscaled





**Fig. 3** Tree canopy cover (TCC) distribution in (3a) Los Angeles, California, (3b) Portland, Oregon, and (3c) Phoenix, Arizona, showing native National Land Cover Database (NLCD) tree canopy cover (left), upscaled high-resolution tree canopy cover (center), and enhanced tree cover using the random forest model (right).

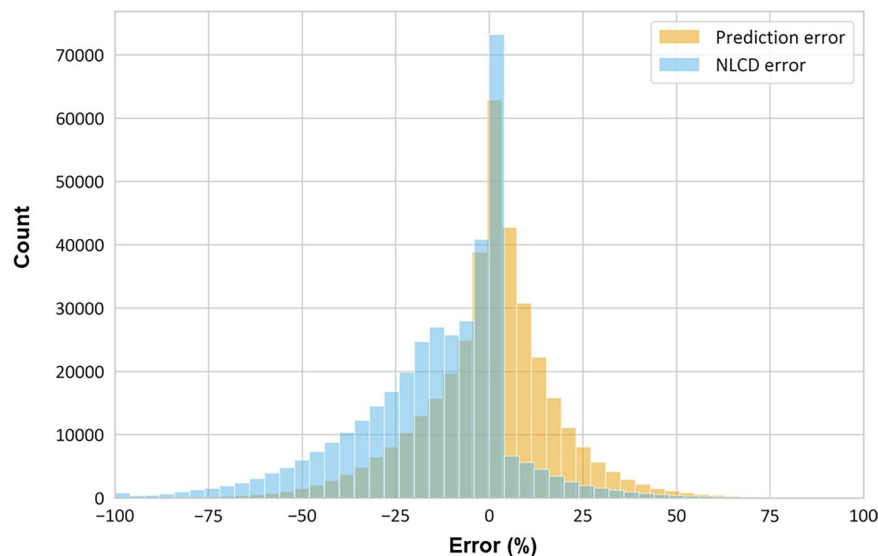
high-resolution TCC data, and our enhanced (predicted) NLCD TCC product for the urban areas of Los Angeles, CA, Portland, OR, and Phoenix, AZ. Similar figures for all 71 urban areas are included as supplementary information (Supplemental File 1). Our data are available as a U.S. Geological Survey data release at <https://doi.org/10.5066/P13LECKC><sup>49</sup>. Included in the data release are a cloud optimized GeoTiff, all supplemental files, training data, python code, and a metadata XML<sup>49</sup>.

### Technical Validation

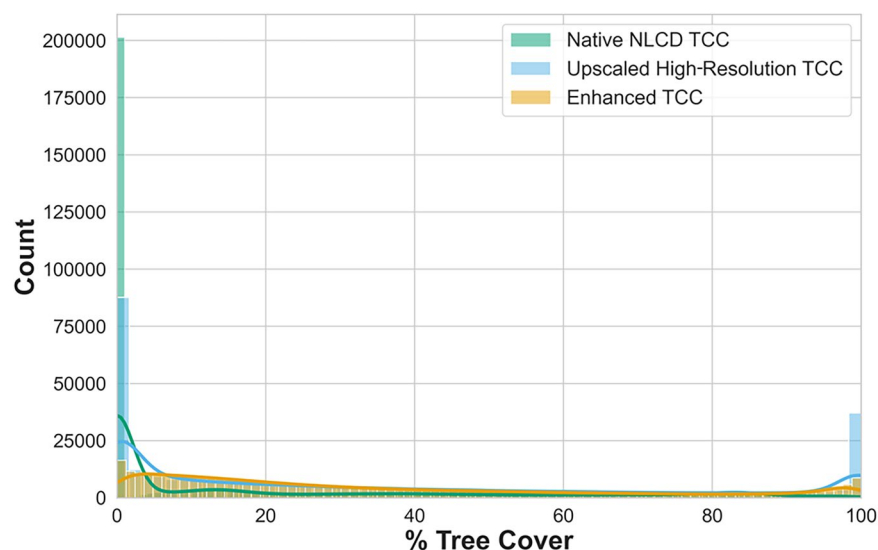
**Model performance.** We evaluated our model's performance using  $R^2$ , Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Our enhanced TCC model's overall  $R^2$  value of 0.747 showed that it explained variation in TCC relatively well. Additionally, our model substantially normalized the distribution of error (the difference between upscaled TCC derived from high-resolution data and native NLCD TCC), Fig. 4). Our model consistently improved TCC estimates relative to the native NLCD TCC dataset, because NLCD's 30 m resolution inherently misses finer-grained tree canopy that is detected by higher-resolution data, which we upscaled. Heris *et al.* (2022) previously evaluated how climate and urban structure influence the inherent underestimates in NLCD TCC<sup>10</sup>. Except at extremely low and high values, our modelled TCC values (smoothed orange line), were generally closer to the high-resolution values (green line) than the uncorrected NLCD TCC (blue line; Fig. 5).

At the level of individual urban areas,  $R^2$  values ranged from 0.292 (Phoenix-Mesa, AZ) to 0.813 (Portland, ME); 52 of 71 urban areas had  $R^2$  values of 0.6 or greater (Supplemental File 2 and in the U.S. Geological Survey data release<sup>49</sup>). The native NLCD TCC product underestimated TCC in all urban areas – by as little as 3.5–3.7% in Shannondale, West Virginia, and Durham, North Carolina, by 13.4% on average across the urban areas, and by as much as 29.4% in Forestville, California. Our TCC model added TCC to correct these underestimates (3.6–5.1% added TCC in Durham and Charlotte, North Carolina, 12.5% on average across the urban areas, and 20.7% and 24.7% for Tampa-St. Petersburg, Florida, and Forestville, California, respectively).

Model performance was generally better in urban areas with higher TCC (coefficient of determination ( $R^2$ ) = 0.52, Fig. 6). Three small humid-region urban areas with extensive forested cover within or just outside the cities had TCC above 60% but  $R^2$  values below 0.7 - Williamsburg, Virginia; Shady Side-Deale, Maryland; and Shannondale, West Virginia. In fitting a quadratic relationship to Fig. 6, we assumed that the relationship between TCC and  $R^2$  for these three small cities is atypical and was not indicative of national-scale patterns. By contrast, urban area-level model  $R^2$  was weaker in areas with the least TCC, typically found in the arid regions of the western United States. Using Analysis of Variance (ANOVA) with Tukey's Honestly Significant Difference post-hoc analysis, we found statistically significant differences in  $R^2$  values by EPA Level I ecoregions ( $p < 0.01$ , Fig. 7), with urban areas in more arid ecoregions tending to have lower  $R^2$  values. However, urban area  $R^2$  values



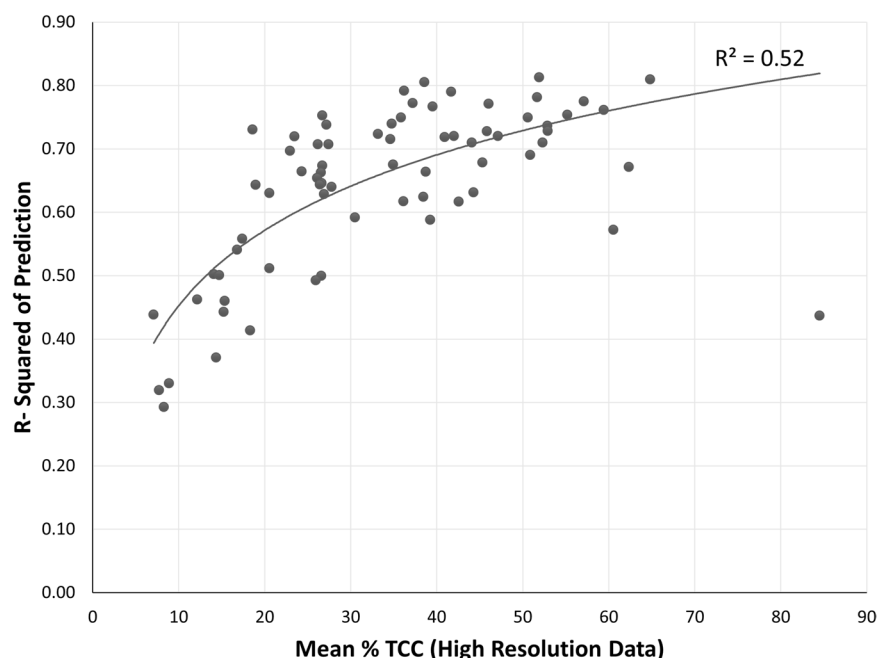
**Fig. 4** Summary of the distribution of error in National Land Cover Database (NLCD) tree canopy cover (TCC) and our Predicted TCC (deviation by % when native NLCD TCC are compared to upscaled high-resolution TCC data versus enhanced TCC).



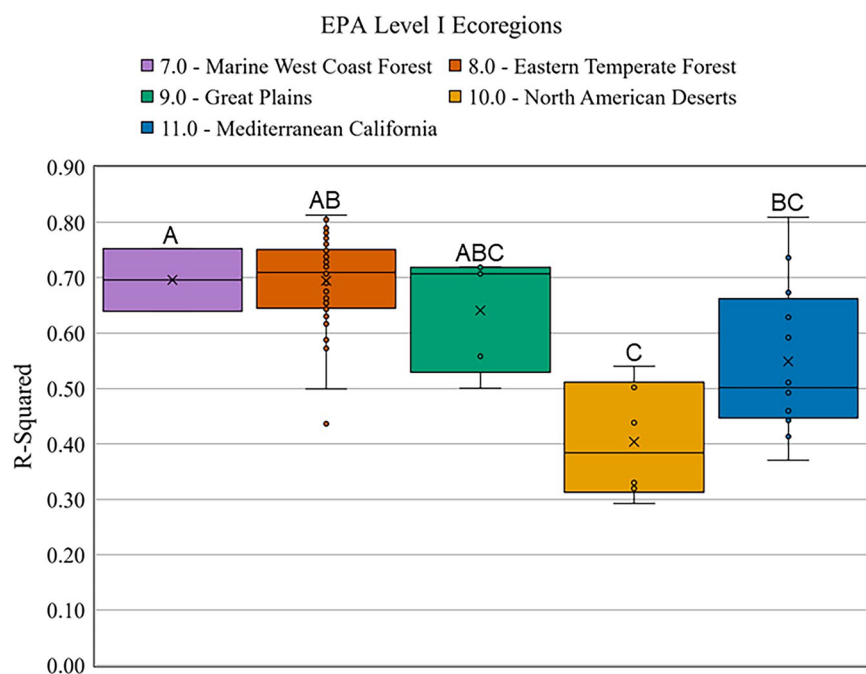
**Fig. 5** Comparative distributions of tree canopy cover (TCC) data from three data sources: (1) native National Land Cover Database (NLCD) tree canopy data (blue); (2) upscaled high-resolution tree canopy data (green); and (3) enhanced tree canopy data (red).

had negligible relationships with climate variables, including mean annual precipitation, January minimum temperature, and August maximum temperature, with  $R^2$  values for linear regression models of 0.040, 0.004, and 0.014, respectively.

Prediction error (difference between upscaled and native NLCD TCC versus enhanced and native NLCD TCC) was substantially reduced – to 0.9% on average – compared to the existing error in the native NLCD TCC dataset – 13.4% on average – although it was overestimated by 2.1–3.2% in four urban areas and underestimated by 2.0–7.1% in 15 urban areas. When evaluating prediction errors using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), our enhanced TCC model consistently outperformed the native NLCD TCC product. Urban area RMSE values for the enhanced TCC ranged from 8.3% to 24.7%, substantially lower than the native NLCD TCC's RMSE range of 13.0% to 37.0%. Similarly, MAE values for the enhanced TCC varied between 5.2% and 17.6%, compared to the native NLCD TCC's MAE range of 7.1% to 29.9% (Supplemental File 2 and data release<sup>49</sup>). On average across all urban areas, the enhanced TCC reduced RMSE by 32% and MAE by 31%, indicating improved accuracy and precision.



**Fig. 6** Scatter plot of coefficient of determination ( $R^2$ ) values for each urban area versus their respective mean high-resolution tree canopy cover (TCC), with a quadratic line of fit.



**Fig. 7** Boxplot depicting the distribution of coefficient of determination ( $R^2$ ) values across different U.S. Environmental Protection Agency (EPA) Level I Ecoregions. The boxes represent the interquartile range (IQR) of the  $R^2$  values, the horizontal line within each box indicates the median, and the 'x' marks represent the mean. Outliers are shown as individual points outside the whiskers, which extend to 1.5 times the IQR from the box. Letters correspond to statistically significant differences in mean  $R^2$  values between the Ecoregions using the Tukey's Honestly Significant Difference post-hoc analysis.

In summary, while our model fits were generally good for most urban areas, there was a more consistent pattern of underestimation than overestimation in the predictions, as indicated by the negative prediction error for many urban areas. While the enhanced TCC predictions underestimate tree canopy, we have more effectively

Variable	Importance with Building Data	Importance without Building Data
NLCD <sup>a</sup> Tree Canopy Cover (TCC)	0.38	0.40
Normalized Difference Vegetation Index	0.16	0.18
NLCD TCC Standard Error	0.09	0.09
NLCD Impervious	0.08	0.08
Surface Temperature	0.08	0.10
NLCD Land Cover	0.06	0.06
Median Year Built	0.04	0.04
Total Building Footprint Coverage	0.02	—
Average Area of Buildings Intersecting Each Cell	0.02	—
Average January Low Temperature	0.02	0.02
Average Annual Precipitation	0.02	0.02
Number of Unique Buildings Intersecting Each Cell	0.01	—
EPA <sup>b</sup> Level II Ecoregion	0.01	0.01
Average August High Temperature	0.01	0.01

**Table 1.** Variable importance for each explanatory variable in the model. <sup>a</sup>NLCD: National Land Cover Database. <sup>b</sup>EPA: Environmental Protection Agency.

addressed the underestimations inherent in the native NLCD TCC data, as indicated by the substantial reductions in RMSE and MAE.

Four temporally dynamic explanatory variables derived from NLCD and two from Landsat imagery (i.e., NDVI and surface temperature) were the most influential predictors of TCC (Table 1). These six explanatory variables had a collective importance of 0.85, with the largest contributions from NLCD TCC (0.38) and NDVI (0.16). Other explanatory variables had importance values of 0.04 (Census block group year built, which will change slowly over time in neighborhoods experiencing changes in their building stock), 0.06 (collective value for three long-term climate average variables plus static EPA ecoregions) and 0.05 (collective value for three building datasets for which we lack temporally dynamic data). When rerunning the model without the building variables, the  $R^2$  value decreased slightly from 0.747 to 0.728, indicating a minimal reduction in the model's explanatory power. Variable importances for the model without building variables remained consistent relative to the model with the building variables (e.g., NLCD TCC: 0.40, NDVI: 0.18, surface temperature: 0.10).

The low importance of static building datasets and minimal impacts of their exclusion implies that our model predominantly relies on variables that can be readily updated for different years (e.g., NLCD, NDVI, surface temperature, median year built) and for which the use of static data is appropriate (multi-decade climate averages and ecoregions). This indicates that our model can be appropriately used to produce time series of enhanced TCC data by using updated versions of its most influential variables. Further, while temporally dynamic building data are to our knowledge not yet available in the United States, their recent release in the Global South portends the likely greater availability and accuracy of dynamic building data in the future<sup>50</sup>.

**Cross validation.** We used randomized K-fold cross-validation using the Python 'sklearn' package version 1.0.2<sup>47</sup> to measure model performance and validate the accuracy of the predicted TCC data. This process meant that the division of the data into K folds was performed randomly for each repetition, mitigating potential bias associated with non-random division and enhancing the robustness of model validation<sup>51</sup>. We assessed the performance of the model in each of the K iterations by computing the adjusted  $R^2$  score for each iteration prediction. The average performance across all K iterations provided a more robust estimate of the model's ability to predict TCC accurately compared to a single random train-test split. We did not perform spatial cross validation because recent work indicates it lacks a robust theoretical basis and standard cross validation approaches often have less bias than spatial cross validation<sup>52</sup>.

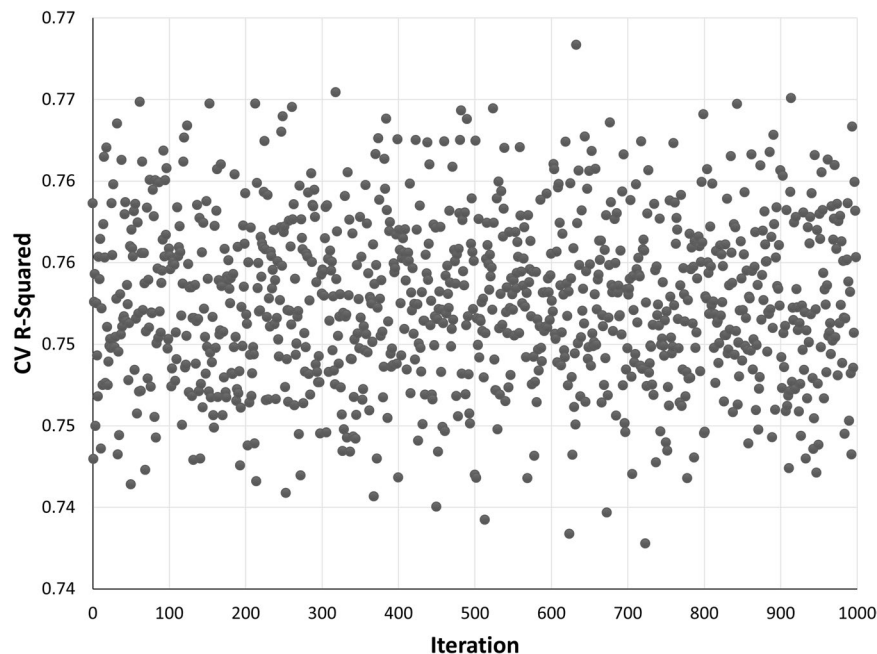
In addition to a random K-fold cross validation, we performed a similar cross validation using individual urban areas ( $k = 71$ ). In each fold, a single urban area was left out as the validation set, and the model was trained on the remaining 70 urban areas. This procedure was iteratively performed 71 times, ensuring that each urban area was used as a validation set once. This method ensures that during training, the model is never exposed to data from the urban area being used for validation, providing a rigorous evaluation of its generalization capability across different urban regions.

Finally, to address the possibility that older neighborhoods may contain larger and more mature trees<sup>41</sup>, i.e., greater TCC, we performed a cross validation on unique Census block group-scale median year built (1939 to 2016;  $k = 76$ ). For each fold, one unique median year-built subset is held out for validation, while the remaining 75 subsets are used for training. This process is repeated 76 times, with each median year-built subset serving as the validation set once.

The 1,000-fold random cross-validation results indicated a high degree of stability in predicting outcomes from diverse data splits. Across the 1,000 iterations,  $R^2$  values ranged from 0.738 to 0.768, with a mean  $R^2$  of 0.753 (Fig. 8). Such consistency indicated that the model exhibits strong generalization capabilities when applied to unseen data, making it reliable as a basis for producing a national-scale data product.

$R^2$  scores from the urban area leave-one-out cross-validation ranged from 0.287 to 0.808 when Charles Town-Ranson, West Virginia (Jefferson County urban area) and New York, New York (city) were left out of training, respectively (Fig. 9). The median value indicated that for a typical urban area, roughly 67.5% of the variation in the outcome was explained by the model. This exercise identified some cities that were relatively





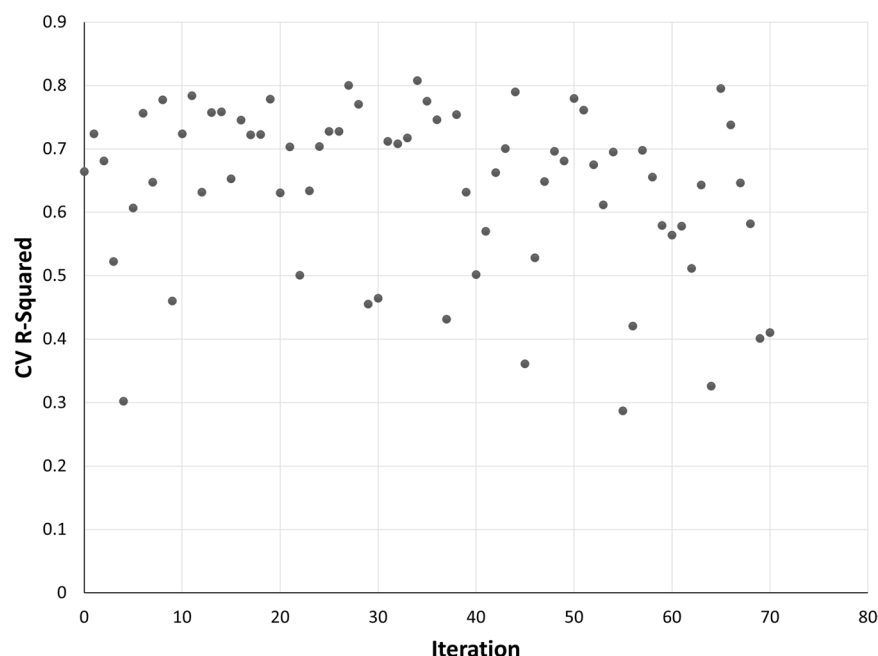
**Fig. 8** Scatter plot representing the results from a  $k = 1,000$ -fold cross validation (CV). Each point indicates the coefficient of determination ( $R^2$ ) value obtained from a single fold of the cross validation. The y-axis shows variation in model performance ( $R^2$  values) across the 1,000 iterations.

influential in the overall model, i.e., having city-specific models that performed well, but when holding out that city's data and rerunning the model, had a reduced  $R^2$  value (primarily small cities in the eastern United States and northern California, but also Baltimore, Maryland; Durham North Carolina; Pittsburgh, Pennsylvania; and Seattle, Washington). Other cities'  $R^2$  values increased during the leave-one-out cross validation, relative to the city-specific results. By leaving out data for these cities, model performance improved, which was the case in some arid cities with low initial  $R^2$  values (e.g., some cities in the intermountain west and southern California) but also in a few larger cities with higher initial  $R^2$  values (e.g., New York City, New York; Milwaukee, Wisconsin; Minneapolis-St. Paul, Minnesota-Wisconsin).

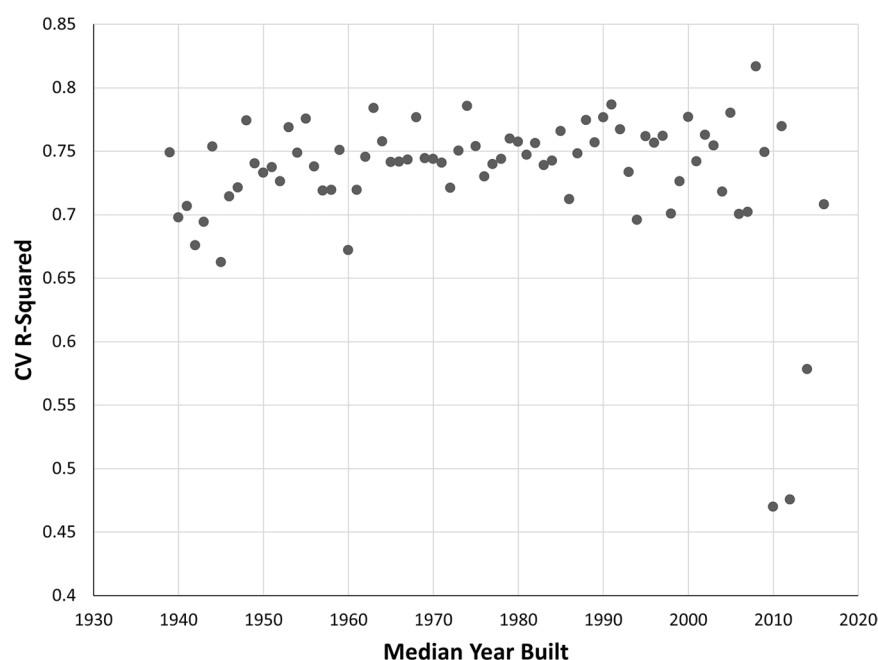
The Census block group median year-built leave-one-out cross-validation produced relatively consistent  $R^2$  values (mean = 0.73), indicating that the model was able to consistently predict TCC with reasonable accuracy across various median Census block group year built, even when each specific year was excluded from the training set. Recent years had both some of the lowest values (2010, the lowest, at 0.470, as well as 2012 and 2014); 2008 had the highest  $R^2$  of 0.817 (Fig. 10). From the mid-2000s onward, the number of cities with median year-built Census block groups declined rapidly (Fig. 11), most likely in response to the instability in the housing market brought on by the Great Recession. With fewer cities represented in more recent years, median TCC was highly variable by median year built (Fig. 12).

In many cases, the few Census block groups within these more recent years are characterized by densely built areas with very little to no TCC. Our random forest model tended to underperform when predicting areas with minimal tree cover because of its difficulty in accurately predicting 0% TCC. The lack of trees within a small number of highly built Census block groups likely contributed to poorer model performance in 2012 and 2014, as random forests require a full range of input values to make accurate predictions, and these areas lack the variability in TCC that the model needed to learn effectively. Additionally, the 2011 map year NLCD TCC product is built using multiyear imagery for 2007–2011, with most data being from 2010. This may introduce some localized error in our model estimates for Census block groups with median year built around these years, for example, if 2007 imagery indicated TCC levels that changed substantially due to subsequent construction.

**Spatial autocorrelation.** Positive spatial autocorrelation refers to the phenomenon where values of a variable observed at points near each other are, on average, more similar than those for points located more distantly from each other. In the context of large-scale geographic regression models, spatial autocorrelation can lead to biased parameter estimates, invalid standard errors, and ultimately Type 1 errors. This is because the number of reported samples will necessarily be higher than the number of spatially independent samples (independence is a requirement for Ordinary Least Squares regression), leading to inflation of test statistics in the absence of adjustment<sup>53</sup>. However, the degree of spatial autocorrelation in urban tree canopy can vary widely from one urban area to another due to factors including urban planning and zoning, demographic distribution, and environmental characteristics.



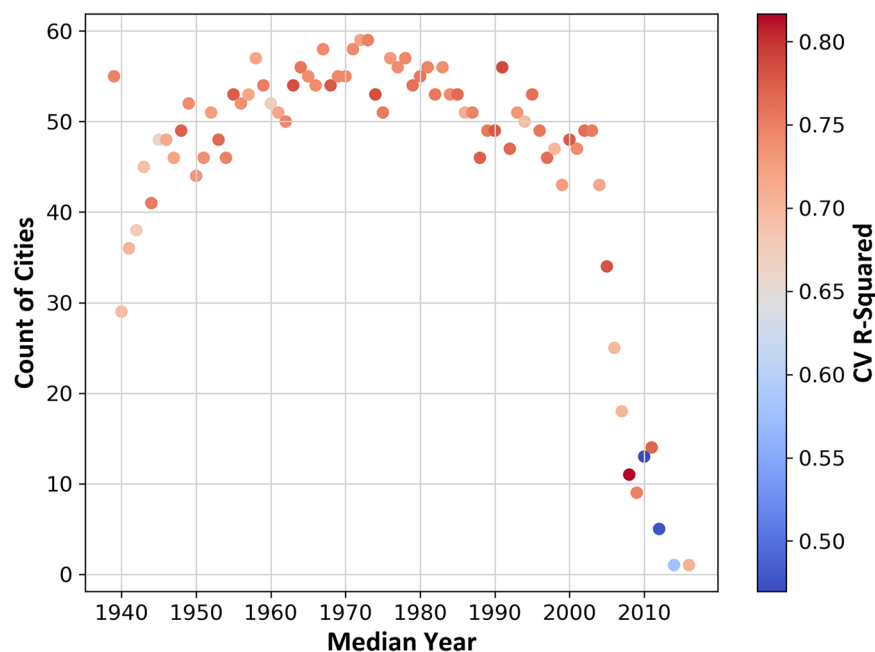
**Fig. 9** Leave-one-out cross-validation (CV) results of the 71-urban area tree canopy correction model, where each partition corresponds to a unique urban area. Each point indicates the coefficient of determination ( $R^2$ ) value.



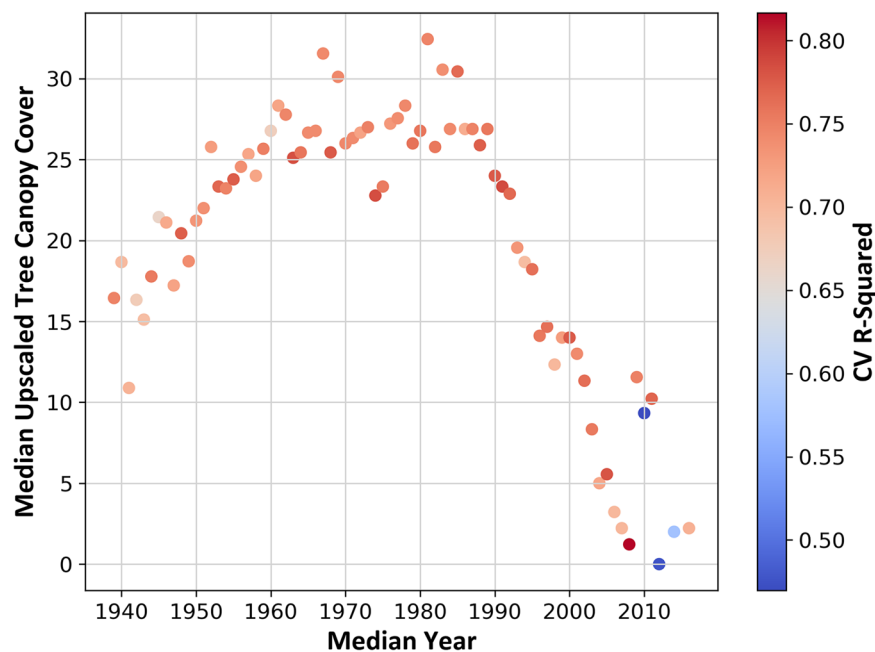
**Fig. 10** Cross-validation (CV) results across 76 iterations for different values of the median year built data by U.S. Census block group. Each point indicates the coefficient of determination ( $R^2$ ) value.

Spatial autocorrelation can be addressed by creating a spatial weights matrix that reflects spatial relationships specific to the area being analyzed, and subsequent application of spatial error models. Because our model aims to predict outcomes across a very diverse array of urban areas (Fig. 2), this would necessitate the development of a unique spatial weights matrix for each individual prediction area. Doing so would pose substantial challenges in terms of scalability and flexibility of the model, as each new prediction area would require the generation and validation of a new spatial weights matrix.

Moreover, the primary strength of random forest algorithms lies in their ability to handle large datasets with numerous variables and its robustness to overfitting, rather than in modelling spatial dependencies<sup>31</sup>. Being



**Fig. 11** Scatter plot showing the number of cities with Census block groups having median year built each year, color-coded by cross-validation coefficient of determination ( $R^2$ ) values.



**Fig. 12** Scatter plot showing the median tree canopy cover (TCC) for each median year built, color-coded by coefficient of determination ( $R^2$ ) values obtained from cross validation.

non-parametric, random forest models do not rely on assumptions about the distribution or structure of the data, including the independence of observations. This inherent flexibility allows them to better accommodate spatial autocorrelation by capturing complex spatial patterns and dependencies directly from the data, potentially mitigating the issues that biased parameter estimates and invalid standard errors commonly raise for parametric models. Finally, unlike when using traditional regression models, our focus is not on identifying which variables are most influential in predicting TCC or understanding the magnitude of their effects. Therefore, to maintain the model's generalizability and applicability across a wide range of urban settings, and to ensure its scalability and feasibility, we did not include spatial error matrices and spatial error models from our random forest model<sup>54–56</sup>.

## Usage Notes

The purpose of the corrected TCC dataset described in this paper is to (1) enhance the accuracy and spatial distribution of mapped mean TCC within U.S.-wide urban areas and (2) do so in a way suitable for monitoring change over time, by using NLCD products as inputs. Our random forest model can be applied across more years than just the current 2011 dataset, and thus helps address key spatiotemporal coverage limitations of current high-resolution TCC data. Leveraging refined methodology and advanced algorithms, the methodology improves common inaccuracies found in urban areas for the NLCD TCC product, yielding a more precise depiction of the TCC in urban landscapes nationwide. By refining the spatial distribution of TCC, the dataset better reflects TCC variability and pattern, facilitating an improved understanding of the urban forest structure. Additionally, the augmented accuracy of the random forest-generated TCC provides critical support for various analyses quantifying both ecological processes and the ecological and socio-economic importance of urban forests, which contribute to sustainable urban planning and development. However, despite its strengths and spatial coverage, this dataset is not intended to accurately represent the precise locations of individual trees. Consequently, it should not be used to identify site-specific locations for planting individual trees. Its main strengths lie in the broad-level assessment and planning of urban forestry<sup>57</sup>, particularly in supporting time-series assessments, greatly contributing to the study and conservation of urban green spaces, and enhancing our ability to quantify and optimize the ecosystem services benefits provided by urban TCC.

The model code enables users to produce their own time series of urban TCC in regions within or outside the United States as new data become available. Users seeking to produce their own urban TCC datasets should be aware of the timing of all model inputs. The use of inputs that are temporally misaligned by more than a few years will increase the error of subsequent products. Greater error owing to temporal misalignment of inputs is most likely in urban areas or parts of urban areas that are being quickly developed or redeveloped, are affected by pest or disease outbreaks that may cause widespread tree mortality, or potentially in humid tropical to subtropical climates where longer growing seasons and greater water availability can lead to more rapid tree growth than in cooler and drier climates. We suggest that those reusing our code to produce time-series urban TCC datasets benchmark their data first, as we did, against the year(s) best matching underlying high-resolution TCC data, then proceed to compare modeled results to those of time-series moderate-resolution TCC products (i.e., the native NLCD TCC product, in our case).

## Code availability

The code used to produce the data available is in the form of Jupyter notebooks. All code is publicly available at <https://doi.org/10.5066/P13LECKC49>.

Received: 6 May 2024; Accepted: 13 March 2025;

Published online: 24 March 2025

## References

1. Cadenasso, M. L., Pickett, S. T. A. & Schwarz, K. Spatial heterogeneity in urban ecosystems: Reconceptualizing land cover and a framework for classification. *Frontiers in Ecology and the Environment* **5**, 80–88, <https://doi.org/10.1890/1540-9295> (2007).
2. Pickett, S. T. A. *et al.* Dynamic heterogeneity: a framework to promote ecological integration and hypothesis generation in urban systems. *Urban Ecosystems* **20**, 1–14, <https://doi.org/10.1007/s11252-016-0574-9> (2017).
3. Alberti, M. *et al.* The Complexity of Urban Eco-evolutionary Dynamics. *BioScience* **70**, 772–793, <https://doi.org/10.1093/biosci/biaa079> (2020).
4. Alonzo, M., Baker, M. E., Gao, Y. & Shandas, V. Spatial configuration and time of day impacts the magnitude of urban tree canopy cooling. *Environmental Research Letters* <https://doi.org/10.1088/1748-9326/ac12f2> (2021).
5. Jiao, M., Zhou, W., Zheng, Z., Wang, J. & Qian, Y. Patch size of trees affects its cooling effectiveness: A perspective from shading and transpiration processes. *Agricultural and Forest Meteorology* **247**, 293–299, <https://doi.org/10.1016/j.agrformet.2017.08.013> (2017).
6. Escobedo, F. J., Kroeger, T. & Wagner, J. E. Urban forests and pollution mitigation: Analyzing ecosystem services and disservices. *Environmental Pollution* **159**, 2078–2087, <https://doi.org/10.1016/j.envpol.2011.01.010> (2011).
7. Yang, Y. *et al.* Spatial Heterogeneity analysis of urban forest ecosystem services in Zhengzhou City. *PLoS ONE* **18**, e0286800, <https://doi.org/10.1371/journal.pone.0286800> (2023).
8. Dowtin, A. L., Cregg, B. C., Nowak, D. J. & Levia, D. F. Towards optimized runoff reduction by urban tree cover: A review of key physical tree traits, site conditions, and management strategies. *Landscape and Urban Planning* **239**, 104849, <https://doi.org/10.1016/j.landurbplan.2023.104849> (2023).
9. Fletcher, D. H. *et al.* Location, Location, Location: Modelling of Noise Mitigation by Urban Woodland Shows the Benefit of Targeted Tree Planting in Cities. *Sustainability* **14**, 7079, <https://doi.org/10.3390/su14127079> (2022).
10. Heris, M., Bagstad, K. J., Troy, A. R. & O'Neil-Dunne, J. P. M. Assessing the Accuracy and Potential for Improvement of the National Land Cover Database's Tree Canopy Cover Dataset in Urban Areas of the Conterminous United States. *Remote Sensing* **14**, 1219, <https://doi.org/10.3390/rs14051219> (2022).
11. Venter, Z. S., Chakraborty, T. & Lee, X. Crowdsourced air temperatures contrast satellite measures of the urban heat island and its mechanisms. *Sci. Adv.* **7**, eabb9569, <https://doi.org/10.1126/sciadv.abb9569> (2021).
12. Nowak, D. J., Ellis, A. & Greenfield, E. J. The disparity in tree cover and ecosystem service values among redlining classes in the United States. *Landscape and Urban Planning* **221**, 104370, <https://doi.org/10.1016/j.landurbplan.2022.104370> (2022).
13. Heris, M. *et al.* Piloting urban ecosystem accounting for the United States. *Ecosystem Services* **48**, 101226, <https://doi.org/10.1016/j.ecoser.2020.101226> (2021).
14. Locke, D. H. *et al.* Residential housing segregation and urban tree canopy in 37 US Cities. *npj Urban Sustainability* **1**, 15, <https://doi.org/10.1038/s42949-021-00022-0> (2021).
15. Homer, C., Huang, C., Yang, L., Wylie, B. & Coan, M. Development of a 2001 National Land-Cover Database for the United States. *Photogrammetric Engineering and Remote Sensing* **70**, 829–840, <https://doi.org/10.14358/PERS.70.7.829> (2004).
16. Homer, C. *et al.* Completion of the 2011 National Land Cover Database for the Conterminous United States – Representing a Decade of Land Cover Change Information. *Photogrammetric Engineering*, (2015).
17. Wickham, J. & Riitters, K. H. Influence of high-resolution data on the assessment of forest fragmentation. *Landscape Ecol* **34**, 2169–2182, <https://doi.org/10.1007/s10980-019-00820-z> (2019).
18. Housman, I. *et al.* National Land Cover Database Tree Canopy Cover Methods. 26 [https://data.fs.usda.gov/geodata/rastergateway/treecanopycover/docs/TCC\\_v2021-4\\_Methods.pdf](https://data.fs.usda.gov/geodata/rastergateway/treecanopycover/docs/TCC_v2021-4_Methods.pdf) (2023).

19. Venter, Z. S., Barton, D. N., Chakraborty, T., Simensen, T. & Singh, G. Global 10 m Land Use Land Cover Datasets: A Comparison of Dynamic World, World Cover and Esri Land Cover. *Remote Sensing* **14**, 4101, <https://doi.org/10.3390/rs14164101> (2022).
20. Korhonen, L., Hadi, Packalen, P. & Rautiainen, M. Comparison of Sentinel-2 and Landsat 8 in the estimation of boreal forest canopy cover and leaf area index. *Remote Sensing of Environment* **195**, 259–274, <https://doi.org/10.1016/j.rse.2017.03.021> (2017).
21. Rioux, J.-F., Cimon-Morin, J., Pellerin, S., Alard, D. & Poulin, M. How Land Cover Spatial Resolution Affects Mapping of Urban Ecosystem Service Flows. *Front. Environ. Sci.* **7**, 93, <https://doi.org/10.3389/fevns.2019.00093> (2019).
22. Zhang, B. *et al.* The scale effects of the spatial autocorrelation measurement: aggregation level and spatial resolution. *International Journal of Geographical Information Science* **33**, 945–966, <https://doi.org/10.1080/13658816.2018.1564316> (2019).
23. Nowak, D. J. & Greenfield, E. J. Evaluating The National Land Cover Database Tree Canopy and Impervious Cover Estimates Across the Conterminous United States: A Comparison with Photo-Interpreted Estimates. *Environmental Management* **46**, 378–390, <https://doi.org/10.1007/s00267-010-9536-9> (2010).
24. Zhu, Z. *et al.* Understanding an urbanizing planet: Strategic directions for remote sensing. *Remote Sensing of Environment* **228**, 164–182, <https://doi.org/10.1016/j.rse.2019.04.020> (2019).
25. Pilant, A., Endres, K., Rosenbaum, D. & Gundersen, G. US EPA EnviroAtlas Meter-Scale Urban Land Cover (MULC): 1-m Pixel Land Cover Class Definitions and Guidance. *Remote Sensing* **12**, 1909, <https://doi.org/10.3390/rs12121909> (2020).
26. MacFaden, S. W., O’Neil-Dunne, J. P. M., Royar, A. R., Lu, J. W. T. & Rundle, A. G. High-resolution tree canopy mapping for New York City using LIDAR and object-based image analysis. *Journal of Applied Remote Sensing* **6**, 063567, <https://doi.org/10.1117/1.JRS.6.063567> (2012).
27. Zhang, Y. *et al.* UrbanWatch: A 1-meter resolution land cover and land use database for 22 major cities in the United States. *Remote Sensing of Environment* **278**, 113106, <https://doi.org/10.1016/j.rse.2022.113106> (2022).
28. Earth Define. *Earth Define* <https://www.earthdefine.com/>.
29. Impact Observatory 3m Land Cover. *Impact Observatory* <https://www.impactobservatory.com/3m-land-cover/>.
30. Huang, W. *et al.* High-resolution mapping of aboveground biomass for forest carbon monitoring system in the Tri-State region of Maryland, Pennsylvania and Delaware, USA. *Environ. Res. Lett.* **14**, 095002, <https://doi.org/10.1088/1748-9326/ab2917> (2019).
31. Ahmed, L., Claggett R. P. & McDonald, S. Chesapeake Bay Land Use and Land Cover (LULC) Database 2022 Edition. <https://doi.org/10.5066/P981GV1L> (2023).
32. National Oceanic and Atmospheric Administration, Office for Coastal Management. C-CAP High-Resolution Land Cover. (2024).
33. McDonald, R. I. *et al.* The tree cover and temperature disparity in US urbanized areas: Quantifying the association with income across 5,723 communities. *PLOS ONE* **16**, e0249715, <https://doi.org/10.1371/journal.pone.0249715> (2021).
34. Guo, J., Xu, Q., Zeng, Y., Liu, Z. & Zhu, X. X. Nationwide urban tree canopy mapping and coverage assessment in Brazil from high-resolution remote sensing images using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing* **198**, 1–15, <https://doi.org/10.1016/j.isprsjprs.2023.02.007> (2023).
35. Tolan, J. *et al.* Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sensing of Environment* **300**, 113888, <https://doi.org/10.1016/j.rse.2023.113888> (2024).
36. O’Neil-Dunne, J. P. M., MacFaden, S. W., Royar, A. R. & Pelletier, K. C. An object-based system for LiDAR data fusion and feature extraction. *Geocarto International* **28**, 227–242, <https://doi.org/10.1080/10106049.2012.689015> (2013).
37. Ratcliffe, M., C Burd, Holder, K. & Fields, A. Defining Rural at the U.S. Census Bureau. <https://doi.org/10.13140/RG.2.2.16410.64969> (2016).
38. U.S Census Bureau. Census Urban Area Criteria. *Federal Register* **87** (2022).
39. Ruefenacht, B. *et al.* Forest Service Tree Canopy Cover Mapping: 2016 Product Suite and Methods. 47 [https://data.fs.usda.gov/geodata/rastergateway/treecanopycover/docs/TCC\\_2016\\_MethodsReport\\_2022-08-16.pdf](https://data.fs.usda.gov/geodata/rastergateway/treecanopycover/docs/TCC_2016_MethodsReport_2022-08-16.pdf) (2022).
40. Heris, M. P., Foks, N. L., Bagstad, K. J., Troy, A. & Ancona, Z. H. A rasterized building footprint dataset for the United States. *Sci Data* **7**, 207, <https://doi.org/10.1038/s41597-020-0542-3> (2020).
41. Lowry, J. H., Ramsey, R. D. & Kjølgren, R. K. Predicting urban forest growth and its impact on residential landscape water demand in a semiarid urban environment. *Urban Forestry & Urban Greening* **10**, 193–204, <https://doi.org/10.1016/j.ufug.2011.05.004> (2011).
42. Manfra, R., Dos Santos Massoca, M., Martins Cerqueira Uras, P., Cavalari, A. A. & Maselli Locosselli, G. Average height of surrounding buildings and district age are the main predictors of tree failure on the streets of São Paulo/Brazil. *Urban Forestry & Urban Greening* **74**, 127665, <https://doi.org/10.1016/j.ufug.2022.127665> (2022).
43. Hilbert, D., Roman, L., Koeser, A., Vogt, J. & Van Doorn, N. Urban Tree Mortality: A Literature Review. *AUF* **45** <https://doi.org/10.48044/jauf.2019.015> (2019).
44. PRISM Group. PRISM Climate Group. <http://www.prism.oregonstate.edu/> (2007).
45. Omernik, J. M. & Griffith, G. E. Ecoregions of the Conterminous United States: Evolution of a Hierarchical Spatial Framework. *Environmental Management* **54**, 1249–1266, <https://doi.org/10.1007/s00267-014-0364-1> (2014).
46. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
47. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in 92–96 <https://doi.org/10.25080/Majora-92bf1922-011> (2010).
48. Prober, S. M. *et al.* Climate-adjusted provenancing: a strategy for climate-resilient ecological restoration. *Frontiers in Ecology and Evolution* **3** <https://doi.org/10.3389/fevo.2015.00065> (2015).
49. Corro, L. M. *et al.* An enhanced national-scale urban tree canopy cover dataset for the United States. *ScienceBase Catalog* <https://doi.org/10.5066/P13LECKC> (2025).
50. Sirko, W. *et al.* High-Resolution Building and Road Detection from Sentinel-2. Preprint at <http://arxiv.org/abs/2310.11622> (2024).
51. Coulston, J. W., Blinn, C. E., Thomas, V. A. & Wynne, R. H. Approximating Prediction Uncertainty for Random Forest Regression Models. *Photogrammetric Engineering & Remote Sensing* **82**, 189–197, <https://doi.org/10.14358/PERS.82.3.189> (2016).
52. Wadoux, A. M. J.-C., Heuvelink, G. B. M., De Bruin, S. & Brus, D. J. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling* **457**, 109692, <https://doi.org/10.1016/j.ecolmodel.2021.109692> (2021).
53. Dale, M. R. T. & Fortin, M.-J. Spatial autocorrelation and statistical tests in ecology. *Écoscience* **9**, 162–167, <https://doi.org/10.1080/11956860.2002.11682702> (2002).
54. Hurlbert, S. H. Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs* **54**, 187–211, <https://doi.org/10.2307/1942661> (1984).
55. Legendre, P. Spatial Autocorrelation: Trouble or New Paradigm? *Ecology* **74**, 1659–1673, <https://doi.org/10.2307/1939924> (1993).
56. Davies, G. M. & Gray, A. Don’t let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). *Ecology and Evolution* **5**, 5295–5304, <https://doi.org/10.1002/ece3.1782> (2015).
57. Nelson, M. L. *et al.* Existing Vegetation Classification, Mapping and Inventory Technical Guide, Version 2.0. (2015).

## Acknowledgements

Funding for this work was provided by the U.S. Geological Survey’s Climate Research and Development and Land Change Science programs. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. Late in the writing of this manuscript, our co-author, Jarlath O’Neil-Dunne tragically and unexpectedly passed away. Beyond his valuable contribution to this paper, we acknowledge his profound contribution to the field of land cover and urban forest mapping. One of the most influential pioneers and thought leaders in this field, Jarlath refined and perfected some of the most sophisticated



methods for accurately classifying urban tree canopy at very fine resolutions, using object-based image analysis and data fusion. His skill with these methods, combined with the amazing team of spatial analysts he built and led at the University of Vermont, made him one of the most sought-after urban forest mapping experts in the country and resulted in his lab conducting urban forest assessments for dozens of cities. We will miss him greatly as a colleague, collaborator, and close friend.

### Author contributions

L.M.C. - study design, analysis, writing. K.J.B. - study design, writing. M.P.H. - study design, writing. P.C.I. - study design, writing. K.G.S. - study design, reviewing and editing. J.E.D. - study design, reviewing and editing. A.T. - study design, reviewing and editing. K.M. - study design, reviewing and editing. J.P.M.O. - study design, reviewing and editing.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41597-025-04816-0>.

**Correspondence** and requests for materials should be addressed to L.M.C.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2025