A Data-Driven Approach for Urban Heat Island Predictions: Rethinking the Evaluation Metrics and Data Preprocessing

Berk Kıvılcım and Patrick Erik Bradley

Contact information of corresponding author. Patrick Erik Bradley
Institute of Photogrammetry and Remote Sensing
Karlsruhe Institute of Technology
Englerstr. 7
76131 Karlsruhe
Germany
Email: bradley@kit.edu
Telephone: ++49 721 608 47304

Acknowledgements. Our sincere gratitude goes to Martin Breunig for offering the opportunity to be part of this project, and to Markus Wilhelm Jahn for his excellent guidance of the first author during the initial stage of this work.

Declaration of conflicting interest.

Funding statement. This work is supported by the Deutsche Forschungsgemeinschaft under project number 469999674.

Ethical approval and informed consent statements.

Data availabilty statement. Data produced in this work can be provided upon request. The code for this work can be fund under the following link: https://github.com/BerkKivilcim/Urban-Heat-Modelling/tree/main

Any other identifying information.

A Data-Driven Approach for Urban Heat Island Predictions: Rethinking the Evaluation Metrics and Data Preprocessing

March 6, 2025

Abstract

A 2D raster data representing building volumes of each grids are derived from 3D vector-format urban data for use in machine learning applications. Since the task is to explore patterns, i.e. urban heat islands, Gaussian blurring is implemented on these generated 2D raster data before the training process. This strengthens the visual capturing of spatial relationships, and as a result the correlation rate between air temperature and building volume data is also increased. After the model training, the prediction results are not simply evaluated with most widely used shallow metrics like the Mean Square Error (MSE), but thanks to the raster format of input and output results, some image similarity metrics such as Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) that are able to detect and consider spatial relations are used during the evaluation and interpretation process, because of their higher usefulness in mimicking human visual judgements. The trained models with Random Forest and XGBoost methods which are capable of predicting the spatial distribution of air temperature by using building volume information are compared. By doing so, this research aims to assist urban planners in incorporating environmental parameters into their planning strategies, thereby facilitating more sustainable and inhabitable urban environments.

Keywords. [Building Volume, Air Temperature, Urban Planning, Urban Morphology, Machine Learning for Environment]

1 Introduction

Environmental parameters such as air temperature are critical determinants of human quality of life and energy efficiency management. Urban areas are densely populated and also highly correlated with some of these natural phenomena through urban morphology and landscape spatial patterns. Buildings are the most active areas of human activity and have a significant impact on the urban thermal environments by altering the heat exchange [1, 2, 3, 4, 5]. Consequently, predicting the effects of urban plans on environmental parameters is essential for proper decision making and planning to enhance the living conditions of cities. Moreover, a significant majority of the world's population is predicted to live in urban environments in the future [6]. Besides, temperature increases contribute to health issues with potential for heat-related illnesses and disrupt ecosystems and adversely affect biodiversity [7]. Therefore, it is essential to research the urban thermal environment containing buildings [8]. Previous studies have actually highlighted the strong correlation between urban morphology and air temperature, underscoring the importance of employing three-dimensional data in those analyses. Although it is recognised that climate issues still have limited impact on urban planning processes [9, 10], it is partly because of a gap between urban planners and climatologists [9]. To help urban planners and decision makers better understand and use the research findings, linking the climate issues to planning parameters can be more helpful than geographic or morphological parameters [2]. Therefore, in this article, a data-driven machine learning model is trained to predict urban near-surface air temperature by using building volumes to rethink about how to increase the correlations between the indicator parameters and perform better assessments for model accuracy evaluations.

The power of machine learning algorithms allows to evaluate environmental indicators on a large scale and to map urban air temperature [11, 12, 13, 14, 15], since those algorithms have the advantage of solving complex nonlinear problems with fewer computing sources and less time [16]. Thanks to machine learning methods, we can now better model the patterns of these urban forms to refine those of future cities to meet the needs of rapid urbanization [13]. Although the results obtained when using the machine learning models might differ from the actual measurements while exhibiting a similar trend as the measurements, this makes it still reasonable and acceptable [17]. This issue also occurred in the trained models of the present work. In the end, with the aid of trained models in this study, urban planners can manipulate the building volume values as they desire for the input of the machine learning model in order to observe how their plans will impact environmental factors. It is expected that changing the spatial arrangement of urban components may affect the land surface energy distribution [18]. This could allow them to balance and control their plans based on these impacts, potentially reducing the occurrence of flawed urban planning.

Many of the previous studies focused on investigating the close relation between temperature and buildings [8, 19, 20, 21, 22, 23, 24, 25], while many others implemented machine learning models to predict LST or air temperature by considering urban morphology [6, 16, 17, 26, 27, 28]. According to these studies, a higher building volume contributes to warmer environment within the city centre. Therefore, optimisation of the building volume should be seriously considered, especially during the urban planning decision-making [29]. However, most of these studies utilise the average building height of regions while focusing on large non-uniform selected regional blocks or local climate zones for predictions. On the other hand, our methodology incorporates building heights and footprints directly without averaging, and also associates volumes with two-dimensional air temperature raster data, enabling predictions not just for specific selected region blocks, but on an adjustable per-pixel basis. This approach will allow for very high-resolution predictions to be made swiftly as higher-resolution meteorological data become available in the future.

In addition, by performing pixel-level predictions, it is possible to express the results in raster format and test the model's accuracy by comparing them with ground truth data in the same format. This approach allows us to not just evaluate the models with shallow metrics like Mean Square Error (MSE), but also enables to capture spatial patterns and perform quantitative assessments similar to human visual perception using metrics such as SSIM (Structural Similarity Index) [30], and LPIPS (Learned Perceptual Image Patch Similarity) [31] which are widely adopted similarity metrics [32]. These two perceptual similarity metrics are also used in various different tools such as Nerfstudio [33] to evaluate trained models accuracy which makes them standardized metrics. Moreover, our model's process involved data from ten different cities, with seven used in the training phase and the remaining three for testing. For this reason, the present model still provides a generalised and robust prediction capability.

Besides, land surface temperature (LST) draws significant attention, as it modulates the air temperature of the lower layer of the urban atmosphere [34]. However, since air temperature data over 2 meters above the surface is already provided by the German Weather Service as an publicly available data, and LST is also sensitive to surface emissivity and reflectivity which might fool the trained model using many different cities, the air temperature data is used instead of LST data for the present methodology.

One of the greatest challenges of urban planning today is to produce urban forms that meet the challenge of today's cities [13]. It is mentioned in the past studies that the landscape pattern of two-dimensional space alone is inadequate in explaining the complex thermal phenomena occurring in urban areas [18] and the correlation between thermal phenomena and the 3D-Volume-Index was higher than the 2D-Area-Index [8]. However, the role of urban morphology, such as building height is often overlooked in many cases [27]. Since height-related indicators have been typically chosen as the major parameters to characterise the three-dimensional landscape morphology [35, 36], one of the major limitations has been the difficulty of obtaining high-resolution 3D information about the scale of entire metropolitan areas [18] and accurately estimating the height of buildings on a large scale to obtain the 3D structure of buildings. This is a tough challenge [37]. In addition to that, many analyses from previous studies rely on official urban datasets provided by governments or profit-making organizations, which include building information [17]. Therefore, available and accessible 3D urban morphology data have become essential for extensive academic research on the built environment and urban climate, and a rapid methodology for extracting urban morphology information is urgently needed [38].

It is considered that voxel data is suitable for volumetric calculations in order to generate a 2D raster dataset that contains building volume information for each grid. This choice was made because the capability for volumetric calculations is a key advantage of voxel models, which is absent in other model types [39]. However, since city-scale voxel data were not available for our study areas, we generated voxel models by applying a set of extremely simple steps onto publicly accessible CityGML data. This approach allowed us to rapidly obtain voxel representations, though with lower precision. Unlike previous complex and comprehensive studies that have extensively focused on the voxelization issues of CityGML or CityJSON data, the method used in this study was intentionally simplified to accelerate our experimental processes, as we were not concerned with fine building details while focusing on city-scale since the voxel models only be used to generate 2D raster data.

Some of the comprehensive CityGML-to-voxel conversion algorithms introduced in previous studies [40, 41, 42, 43] implement only simulations on the scale of individual buildings instead of a whole city scale [39, 44, 45]. Cf. also [46, 47, 48] for extraction methods of watertight volumetric models from wireframe data like CityGML, and their robustness. Those methodologies are based on geometric intersection procedures in 3D spaces, since we just only used intersections in 2D spaces to swiftly create city-scale data for multiple urban regions. The other conventional methodologies for obtaining the voxel data makes the need for a large amount of laser scanning, and other sensor data are necessary [49].

Furthermore, the indicator parameters are not just sufficient for a reliable model training since temperature is also affected by the surroundings on a larger spatial scale [50]. Some of the previous studies shows that implementing some convolutional kernels especially the Gaussian Kernel reduced the noise in the raster values [51, 52]. In this study, before the training process a Gaussian blurring algorithm is implemented on 2D raster building volume data which increases the correlation rate between air temperature.

Some of the contributions of this study to literature includes: 1) The implementation of Gaussian blurring on building volume data increases the correlation between air temperature and building volume across all study regions, 2) It is observed that shallow metrics, such as Mean Square Error (MSE), which do not account for spatial relationships, may be misleading when evaluating models for urban heat island predictions. Instead, metrics such as Learned Perceptual Image Patch Similarity (LPIPS) and Structural Similarity Index Measure (SSIM) provide more valuable insights by incorporating spatial dependencies.

Consequently, this study might help future studies for forecasting various other natural phenomena in the future by enhancing the insights about data processing steps and result evaluations. The findings of this study are intended to provide a foundational framework for future research, in particular the ongoing research project *Distributed Simulation of Processes in Buildings and City Models*, funded by the German Research Foundation (DFG), where they can provide a basis for testing mathematical simulation models.

2 Methodology

This study exclusively utilized open-access data and open-source software tools. The employed datasets encompass CityGML data pertaining to the Thuringia state in Germany [53], coupled with hourly air temperature measurements provided by the German Weather Service. The temperature datasets present air temperature at a height of 2 meters above ground level and feature a 1 km spatial resolution [54]. The air temperature data can be accessed in [55]. Data processing procedures were mainly conducted using Python. Additionally, Paraview [56] was used for some visuzalitazion tasks, while QGIS [57] and several of its plugins such as CityJSON plugin [58] and GDAL rasterize tool [59], is needed for data preparation steps to create 2D building volume raster data. In addition, another open-source tool named 'citygml-tools' [60] was used for converting CityGML data into the JSON format.

2.1 Study Area

In this study, ten cities from the state of Thuringia in Germany, varying in size and population density, were selected for the analysis. Seven were used for training and three for testing. The dimensions of the selected areas for these cities are as follows: Erfurt (10km x 8km), Jena (8km x 10km), Weimar (10km x 8km), Suhl (8km x 10km), Altenburg (10km x 8km), Sondershausen (12km x 6km), Gotha (10km x 8km), Sonneberg (8km x 10km), Schmalkalden (10km x 8km), and Gera (8km x 10km). When associating building volume data with air temperature measurements, the air temperature datasets used were those recorded in the same year that the CityGML datasets were created for each respective city. This approach was implemented to minimize inconsistencies arising from temporal resolution discrepancies. To present the results more clearly and to better highlight the investigations of this study, instead of using multi-temporal datasets, we utilised air temperature data by averaging the values obtained specifically at 01:00 AM during the month of July. The rationale for selecting this particular time is based on previous studies, which demonstrated that the correlation between air temperature and urban morphology reaches its maximum at 01:00 AM during the summer [2]. This high correlation has made it possible to present and interpret the results in a clearer and more comprehensible manner.

The coordinate system of the CityGML data, which was used in this research, is EPSG:25832, while the air temperature data possesses latitude and longitude coordinates under the EPSG:4326 system, in addition to X and Y coordinates under the EPSG:3034 system.

The air temperature data, covering all of Germany, was cropped using the upper right and lower left coordinates of the CityGML data to align two datasets accurately. The consistency of data alignment following the crop ping process was assessed by converting the air temperature data into raster format from netCDF format and subsequently loading it into QGIS. The alignment was checked through qualitative comparisons between CityGML region and cropped air temperature region within a shared coordinate system in QGIS to confirm consistency. The visual representation of this comparisons given in Figure 1.

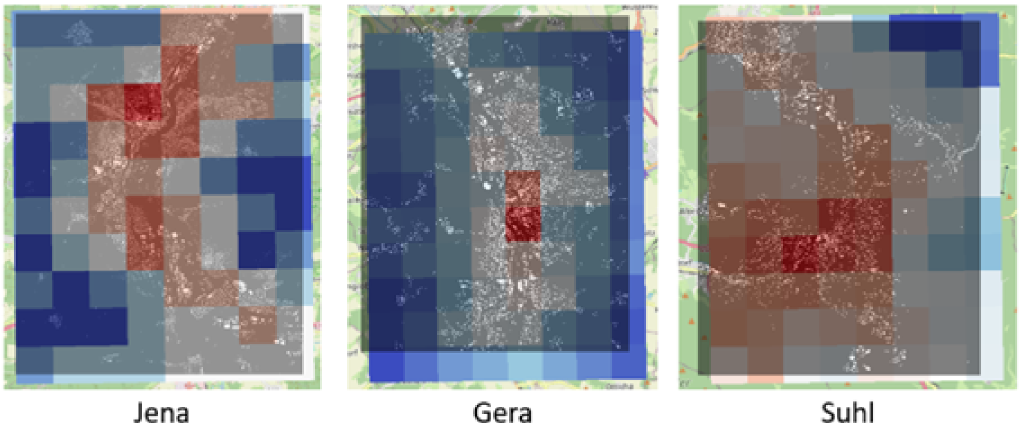


Figure 1: The overlapped visualization of air temperature raster region and rasterized CityGML region for the cities of Jena, Gera and Suhl

2.2 Creating 2D Building Volume Data in Raster

Traditional techniques for converting CityGML data into voxels operate by calculating intersections between CityGML and potentially billions of grid points for high-resolution and extensive areas. Although existing methods can model the many details of buildings and produce complex building voxels, it requires substantial computational power and time. In this study, this complex methods are just avoided to conduct our experiments swiftly and a simplified method that focusing solely on regions with buildings within a two-dimensional plane, assigning a single height value to each building thereby enabling the rapid construction of less detailed buildings. Consequently, the aim of this process is derive a 2D raster building footprint data. A brief workflow illustration is given in Figure 2.

2.2.1 Retrieving 2D Building Footprint Areas

Initially, the CityGML data, downloaded via [53], covered an area of 2 km × 2 km. Therefore, these data were merged to create a single comprehensive

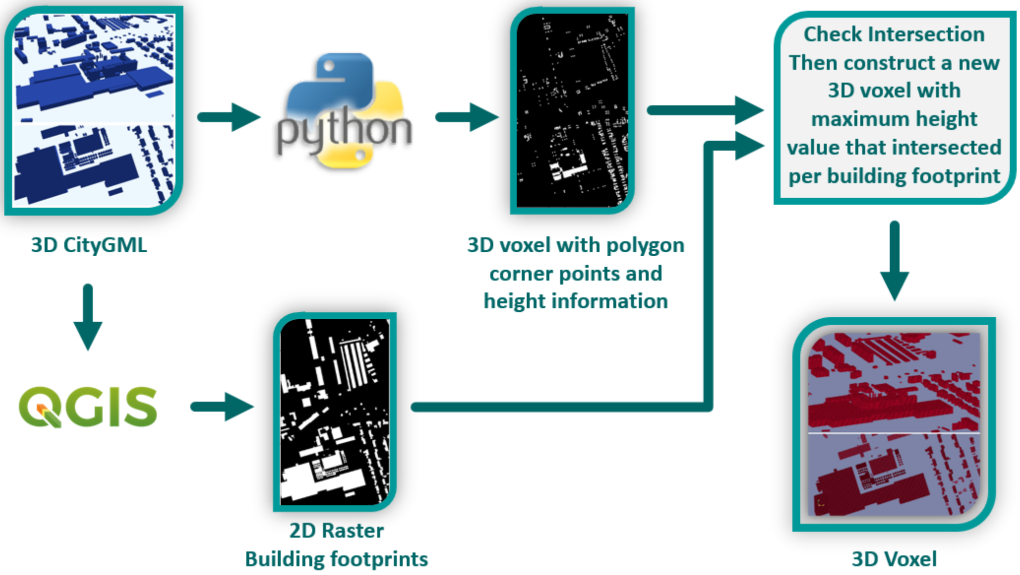

3D CityGML
python
3D voxel with polygon corner points and height information
Check Intersection Then construct a new 3D voxel with maximum height value that intersected per building footprint
QGIS
2D Raster Building footprints
3D Voxel

Figure 2: Workflow of the voxelisation process

CityGML file for each city. This resulting CityGML file was then converted into the CityJSON format and imported into QGIS using the "CityJSON Loader" plugin. Subsequently, GDAL's rasterisation tool was employed to produce raster data for any desired region at any specified resolution. In this study, the raster resolution was set at 1 metre, and the necessary boundary regions for the raster image were extracted from the CityJSON data. Additional parameters selected during the use of the tool included: *"A fixed value to burn: 1", "Assign a specified no data value to output bands: -999", "Output data type: Int16", "Pre-initialize the output image with value: 0".* The output of this process is extracted as GeoTiff format in the coordinate system of EPSG:25832.

2.2.2 Real-world coordinate system to voxel coordinate system

This step involves calculating which location indices in our voxel system correspond to each building's polygon vertices that possess EPSG:25832 coordinate data. In addition to the horizontal plane coordinates of these corner points, the data also includes building height information. Consequently, this allows us to utilise the heights of buildings to combine them with 2D building footprints from the raster image.

For instance, when the region encompassing the city of Erfurt with an area of 10 km × 8km, and the voxel resolution is selected as 1 metre, the number of voxels in the horizontal plane should be 10000 x 8000. Consequently, the indices of the voxel array range from 0 to 9999 for width and from 0 to 7999 for height in Python indexing. Considering all these factors, a normalisation method was employed to transform data from the EPSG:25832 coordinate system to the local voxel coordinate system. The formulas used for the X and Y axes are provided below, cf. eq. (1) and (2). The reason for employing different formulas for the X and Y axes is the orientation of arrays in Python, where the origin (0,0) index is at the top-left corner, whereas the real-world coordinate system of the study area places the origin at the bottom-left. This discrepancy causes a flip along the Y-axis, leading to inconsistencies. The formula used for the Y-axis adjusts this issue. Furthermore, after the normalisation process in the formula, the resulting values between 0 and 1 are multiplied by the width or height values using $width - 1$ or $height - 1$. This minus 1 subtraction adjustment is made because Python indexing starts at 0. In the end, as voxels constitute discrete grids, the new coordinates derived from the formula must be integers. Therefore, a rounding operation is applied to the computed values to ensure they conform to this requirement.

$$X_{\text{voxel}} = \text{round}\left(\frac{(X_{\text{i}} - X_{\text{min}})}{(X_{\text{max}} - X_{\text{min}})} \cdot (width - 1) \right) \tag{1}$$

$$Y_{\text{voxel}} = \text{round}\left(\frac{(Y_{\text{max}} - Y_{\text{i}})}{(Y_{\text{max}} - Y_{\text{min}})} \cdot (height - 1) \right) \tag{2}$$

2.2.3 Assigning the height information to building footprints

In the final stage of the voxelisation process, a voxel is generated based on the normalised building polygon vertex coordinates positioned within the voxel grid and 2D building footprint raster image. If these corner points on the voxels align with a building depicted in a two-dimensional raster image, then the height of the building is derived from the highest height value among the matching vertex coordinates. This method produces buildings in the voxel space that are highly detailed and accurate in terms of their footprint, yet adopt a simplified approach for height representation by assigning a single height value per building. A result of this voxelisation process is demonstrated in the Figure 3.

2.3 Machine Learning Training

In the machine learning model, the phenomenon targeted for prediction is air temperature. Therefore, air temperature data with a resolution of 1km

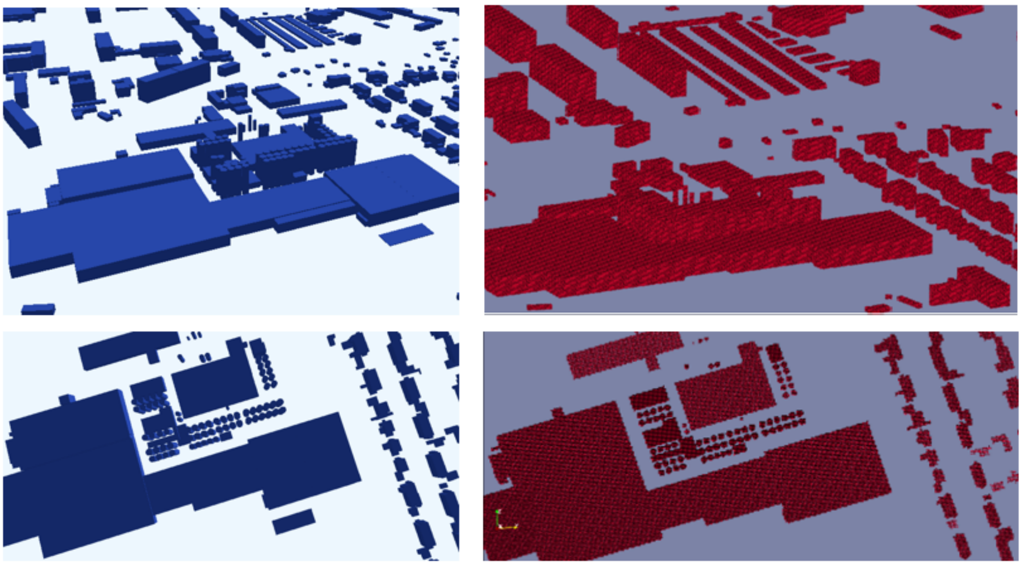


Figure 3: The left column represents the CityGML visualization of Gotha from different viewing angles, the right column represents the voxel visualisation of the same region of Gotha

× 1km has been employed as ground truth data in the training phase. Given that the resolution of these air temperatures is 1km × 1km, the area covered by the voxels is divided into a grid commensurable with the voxel numbers. For example, the region with size of 10km × 8km is divided into a 10 × 8 grid. Afterwards, new two-dimensional building volume data, containing the total building volume for each grid are generated. Consequently, the adjustable resolution of the voxels allows to associate them with higher resolution air temperature data, if available. Thereby our methodology also ensures that the voxelised methods can adapt to improved or different meteorological data resolutions. After the rasterization steps, the other implemented steps before the machine learning training is presented in Figure 4.

Incorporating spatial neighbourhood characteristics during the training phase is crucial, especially when considering urban environments with varying spatial patterns since even if a region has a high building volume but is isolated without many surrounding buildings, it may exhibit a lower urban heat island effect than expected. To account for this, a Gaussian blurring method has been applied to the 2D building volume input data. This approach has demonstrated an increase in the correlation between the building volume in each city and their air temperature. The selected Gaussian kernel parameters are; *sigma value = 0.85* and *radius = 1*. Furthermore, this correlation

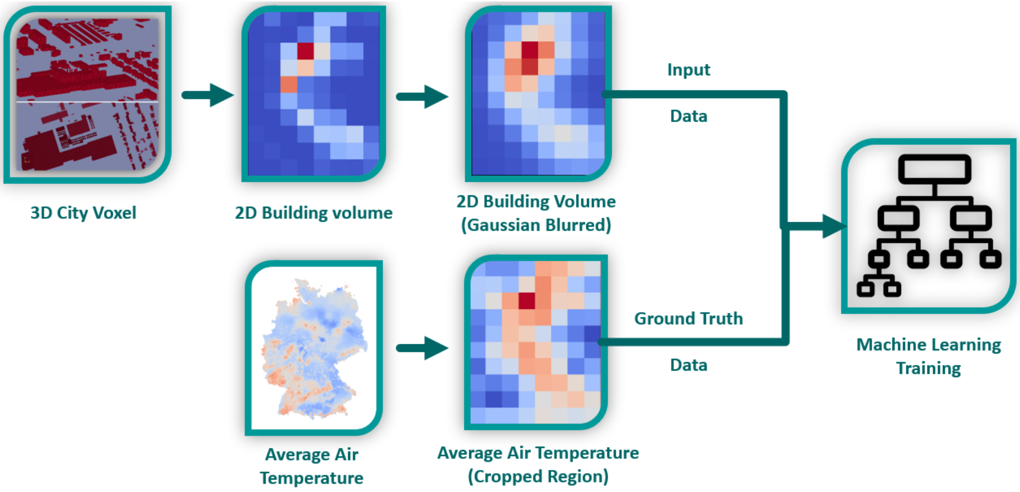

3D City Voxel
2D Building volume
2D Building Volume (Gaussian Blurred)
Input
Data
Average Air Temperature
Average Air Temperature (Cropped Region)
Ground Truth
Data
Machine Learning Training

Figure 4: A brief workflow demonstration of data pre-processing steps for the machine learning training

tends to rise in cities with higher population densities, indicating a significant interplay between urban morphology and thermal behavior. The amount of correlation of original data and Gaussian blurred data with air temperature is given in Table 1, while the visualisations of those datasets are presented in Table 2, which clearly shows the effect of the Gaussian blur.

The Random Forest (RF) and Extreme Gradient Boosting (XGBoost) techniques are selected with hyper-parameter optimisation conducted via a trial and error method for machine learning training. The reason for using the more primitive trial-and-error method instead of systematic hyper-parameter optimization approaches such as Bayesian optimization or grid search was to prioritize not only achieving the best results in quantitative metrics but also to qualitatively assess the models. The goal was to identify models that provided the most generic outcomes based on human visual judgment. Moreover, as evidenced in the conclusion of this study, we demonstrated that when evaluating the trained models using image similarity metrics designed to mimic human visual judgment, the metric results do not always align with those obtained from traditional evaluation metrics such as Mean Square Error, which is also supports the idea of using trial-and-error method in this study. A previous study demonstrated the effectiveness of XGBoost compared to other techniques for predicting urban heat island effects [61]. For the RF model, the hyper-parameters were established as follows: *number of trees* = 100,000, *maximum depth of trees* = 3, *minimum number of samples required to split a node* = 4, *minimum number of samples per leaf* = 2, and *max_features* set to 'sqrt'. For the XGBoost model, the parameters were set to: *number of trees* = 300,000, *maximum depth of trees* = 3, and *learning rate* = 0.000003. Additionally, to mitigate overfitting, augmented data were utilized during training, with parameters specified as *number of samples* = 100 and *noise level* = 0.01."

Table 1: The table represents the correlation values of the original 2D building volume data and the Gaussian blurred data with the air temperature

Datasets	Correlation with building volume and air temperature	Correlation with Gaussian blurred building volume and air temperature
Altenburg	0.78	0.93
Erfurt	0.85	0.90
Gera	0.76	0.87
Gotha	0.71	0.83
Jena	0.73	0.82
Schmalkalden	0.48	0.65
Sondershausen	0.62	0.71
Sonneberg	0.53	0.74
Suhl	0.52	0.66
Weimar	0.67	0.77

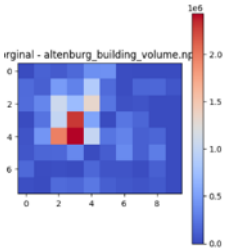
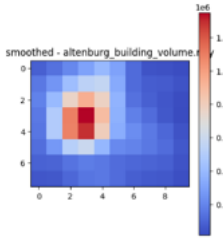
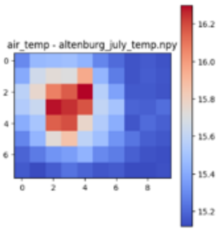
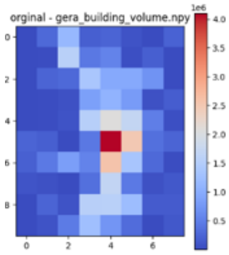
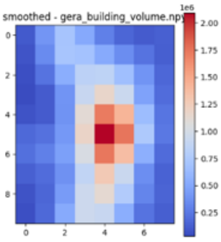
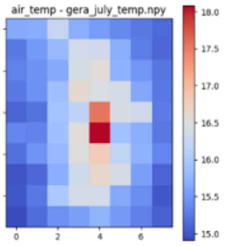
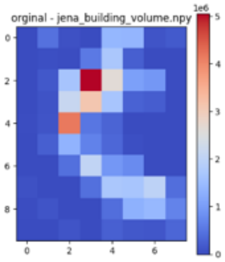
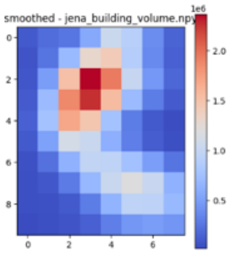
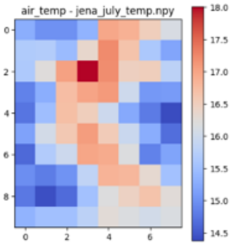
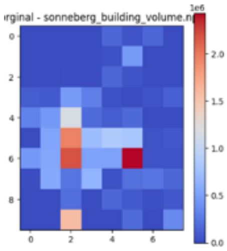
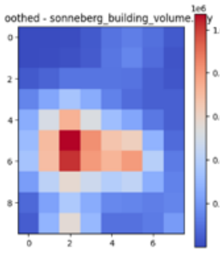
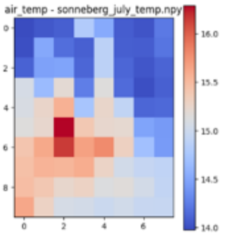
	Building Volume Data	**Gaussian Blurred Building Volume Data**	**Air Temperature Data**
Altenburg			
Gera			
Jena			
Sonneberg			

Table 2: This figure illustrates the spatial distribution of building volumes (expressed in units of m^3) and a monthly average air temperature of July at 01:00 AM (expressed in units of C°) in selected urban areas to show the effect of Gaussian blur, and the relation between urban morphology and air temperature.

3 Results

Since the accuracy of the trained model and the corresponding parameter selections were determined through a trial-and-error approach, multiple training processes were repeated. The resulting training outcomes were analysed both qualitatively and quantitatively. However, in this results section, the model that qualitatively provided the most accurate results and best represented the spatial patterns of the urban heat island effect is presented.

For this reason, rather than focusing solely on achieving a lower MSE (Mean Square Error), visual results were used as the primary basis for accuracy. For instance, in evaluations using deeper trees with both the XGBoost and Random Forest methods, MSE values as low as 0.20 $°C^2$ were achieved for XGBoost, while values around 0.45 $°C^2$ were observed for Random Forest during our experiments. However, upon reviewing the qualitative results, it became evident that these seemingly satisfactory quantitative outcomes were the result of overfitting. This was especially apparent in the visual outcomes, where the urban heat island patterns of cities were not predicted in a consistent and semantically correct manner, failing to capture the expected spatial patterns. On the other hand, some experiments provide appropriate and consistent visual patterns, despite having higher quantitative error values such as 0.92 $°C^2$ MSE for the Random Forest and 0.84 $°C^2$ for the XGBoost techniques.

In addition to the visual analysis of the predictions, the differences between the predicted results and ground truth data were also examined. Upon investigation, it was observed that the Random Forest model's error behavior did not follow any discernible spatial patterns. Instead, the error appeared to occur randomly across the spatial domain, suggesting that the model's inaccuracies were distributed without any systematic bias or spatial structure. On the other hand, when examining the error distribution of the model trained with XGBoost, it was observed that the highest error levels were concentrated in regions corresponding to urban areas. This phenomena can be seen in the following figures of this section.

The comparison between the prediction results and ground truth for the test dataset (Erfurt, Suhl, Sonneberg) obtained using the Random Forest model is presented in Table 4. Although the spatial patterns are well predicted and presented in that Table 4, some deviations in air temperature predictions were observed especially for the cities of Schmalkalden (in training dataset) and Suhl (in test dataset). For instance, the air temperature predictions for Schmalkalden ranged between 14.37°C and 15.64°C, whereas the ground truth data showed a range of 15.14°C to 16.84°C. Similarly, for Suhl, the predictions fell between 14.36°C and 16.02°C, while the actual measurements ranged from 11.37°C to 14.84°C. These discrepancies directly contribute to higher MSE values. However, for other cities, significant differences in air temperature predictions were not observed. For example, the prediction range for Altenburg was 14.72°C to 16.02°C, compared to the ground truth range of 15.12°C to 16.3°C. For Erfurt, the prediction range was 14.33°C to 17.61°C, while the ground truth range was 14.01°C to 17.05°C. Similarly, Gera's prediction interval was 14.79°C to 17.59°C, with a ground truth interval of 14.90°C to 18.08°C; Gotha's prediction interval was 14.33°C to 16.03°C, compared to the ground truth of 13.19°C to 14.88°C; Jena's prediction interval was 14.72°C to 17.61°C, with the ground truth range at 14.37°C to 18.01°C; Sondershausen's prediction interval was 14.33°C to 15.64°C, while the ground truth ranged from 13.94°C to 16.42°C; Sonneberg's prediction interval was 14.33°C to 16.02°C, compared to the ground truth interval of 13.98°C to 16.29°C; and finally, Weimar's prediction interval was 14.58°C to 16.49°C, while the ground truth ranged from 12.94°C to 16.29°C.

The similar comparison between the predictions obtained using the XGBoost model and ground truth data for the test cities (Sondershausen, Schmalkalden, Erfurt) is presented in Table 5. The air temperature prediction ranges for each city using the model trained with the XGBoost method are as follows: Altenburg: 14.58°C to 15.43°C, Erfurt: 14.44°C to 16.08°C, Gera: 14.72°C to 16.08°C, Gotha: 14.44°C to 15.60°C, Jena: 14.68°C to 16.08°C, Schmalkalden: 14.01°C to 15.10°C, Sondershausen: 14.46°C to 15.10°C, Sonneberg: 14.01°C to 15.43°C, Suhl: 14.01°C to 15.43°C, and Weimar: 14.58°C to 16.08°C. The ground truth air temperature intervals for comparison have been provided in the previous paragraph.

Considering the experiments conducted and the quantitative and qualitative results obtained, it has been demonstrated that metrics such as MSE do not fully reflect the accuracy of the model. It was also observed that hyperparameter selection plays a critical and direct role in model performance. Additionally, the Random Forest technique was shown to accurately predict overall spatial patterns even on unseen data (test data) during training.

In addition to the MSE comparisons, the similarities between the predicted and ground truth images were analyzed using SSIM (Structural Similarity Index), and LPIPS (Learned Perceptual Image Patch Similarity) metrics. The SSIM value of 1 indicates two images are exactly the same while 0 means no similarity. LPIPS is specifically designed to evaluate the similarities like human visual perception through deep learning techniques to overcome the shallowness of SSIM [31]. If a LPIPS value is closer to 0 that means two images are very similar to each other. The kernel size for the SSIM calculation was set to 5x5. For the LPIPS computation, all prediction and ground truth data were resized to 64x64, and the AlexNet architecture [62] was used as the multi-layer perceptron model within the deep neural network. That means both of these methods uses kernels to detect the spatial relationships during the evaluations. The results of these metrics are given in the Table 3. While the model trained using the XGBoost method outperformed the Random Forest model in terms of the MSE metric, both qualitative assessments and other quantitative metrics like SSIM and LPIPS indicated that the Random Forest-based method outperformed the XGBoost model across all cities.

Table 3: The table represents SSIM and LPIPS values for Random Forest and XGBoost models across different cities

Dataset	**Random Forest**		**XGBoost**	
	SSIM	**LPIPS**	**SSIM**	**LPIPS**
Altenburg	0.89	0.0000059	0.74	0.0000183
Erfurt	0.82	0.0000504	0.71	0.0002
Gera	0.88	0.0000377	0.72	0.0001
Gotha	0.80	0.0000479	0.77	0.0000430
Jena	0.77	0.00014	0.61	0.0003
Schmalkalden	0.70	0.0000479	0.47	0.0001
Sondershausen	0.65	0.0001	0.45	0.0002
Sonneberg	0.65	0.0001	0.46	0.0001
Suhl	0.62	0.0001	0.46	0.0002
Weimar	0.80	0.0000924	0.71	0.0001

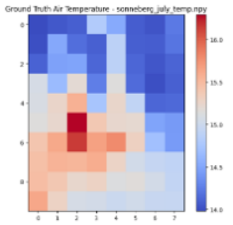
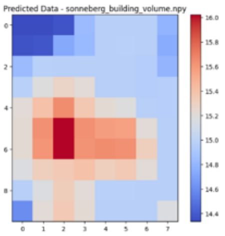
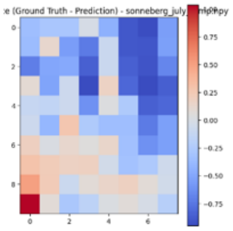
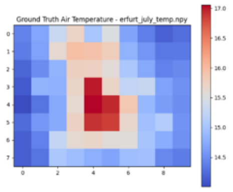
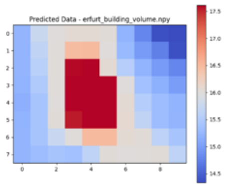
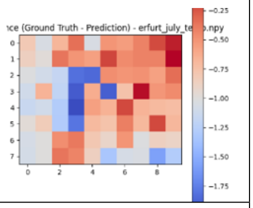
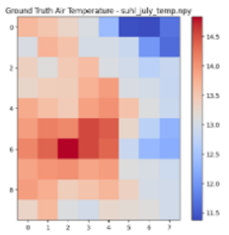
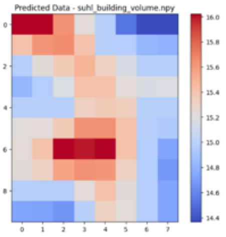
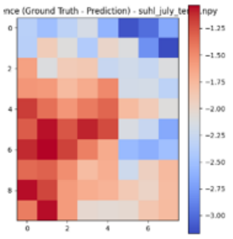
	Ground Truth Air Temperature Map	**Predicted Air Temperature Map**	**Difference Between Ground Truth and Predictions**
Sonneberg			
Erfurt			
Suhl			

Table 4: Comparison of the air temperature prediction results obtained from the model trained with Random Forest technique and ground truth data for the test dataset which is not included during the training process. The difference map between prediction and ground truth images given in the bottom row.

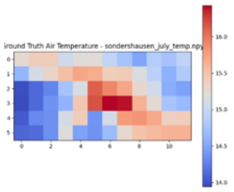
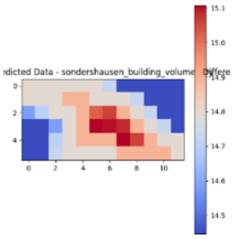
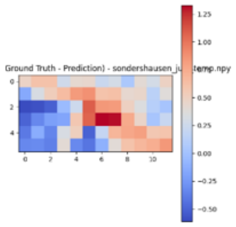
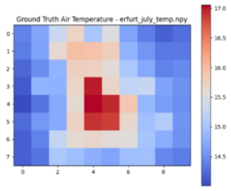
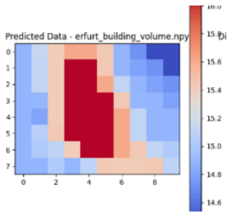
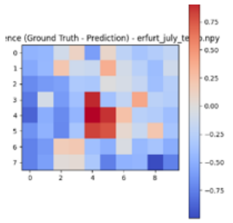
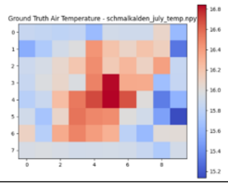
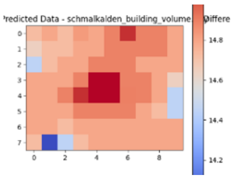
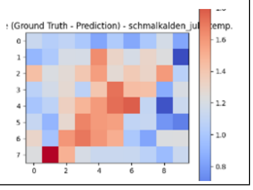
	Ground Truth Air Temperature Map	**Predicted Air Temperature Map**	**Difference Between Ground Truth and Predictions**
Sondershausen			
Erfurt			
Schmalkalden			

Table 5: Comparison of the air temperature prediction results obtained from the model trained with XGBoost technique and ground truth data for the test dataset which is not included during the training process. The difference map between prediction and ground truth images given in the bottom row.

4 Discussion

Although the voxelization methodology presented in this study is designed to yield quick results, it is previously mentioned that the results compromise the level of detail of buildings. Since only a single height value is assigned per building, this approach can present challenges with more complex building structures. In such cases, the voxelization process may lack sufficient precision, leading to some degree of compromise in accurately capturing the architectural complexity of certain buildings.

While generating voxels, the intersection between 2D footprints of the buildings and the building polygon corner points in the voxel coordinate system is used. To increase the possibility of intersection, instead of treating these building polygon corner points as single pixels, 3x3 patches are used. It is important to note that the choice of the patch size is related to the resolution being used. For instance, at lower voxel resolutions, the use of patches may not be necessary, whereas at higher resolutions, larger patches might be required to ensure intersection. On the other hand, using larger patches reduce the precision in cases where buildings are densely situated and have varying heights. However, since the voxel resolution is selected as 1 meter in this study, the use of 3×3 meter patches does not significantly cause the issues in this context. Moreover, buildings that are closely positioned in many urban areas often share the similar height values which minimize the potential impact on precision. Nevertheless, even working with 1 meter resolution voxels and using 3×3 patches, some buildings were not generated due to the failure to achieve intersections. For example, 24990 out of 26642 buildings were successfully matched and generated in Erfurt, it is 19154 out of 20180 in Jena, 17313 out of 18353 in Weimar, 17551 out of 18626 in Suhl, 11526 out of 13001 in Altenburg, 7803 out of 8454 in Sondershausen, 12782 out of 13620 in Gotha, 12002 out of 12814 in Sonneberg, 9847 out of 10670 in Schmalkalden, 19693 out of 22255 in Gera.

Since the correlation between air temperature and building volume data is not perfect, it is not expected to achieve entirely accurate prediction results. However, given that cities with high population densities tend to exhibit stronger correlations with air temperature data, focusing the training process only by using cities such as Tokyo, New York, Istanbul or other cities with high population density may lead to more consistent model training. This approach could improve the performance of models designed for future analyses of metropolitan urban environments. Also, it is possible to obtain significantly different model outcomes even when using the same hyperparameters. Therefore, it remains feasible to train models that outperform or underperform the ones presented in this study by utilizing the same hyperparameters through trial-and-error processes across multiple trainings.

Besides, the methodology presented in this study not just solely focuses on the quantitative accuracy of predictions, but also observed the spatial patterns. This revealed that low MSE values alone are not sufficient to understand model accuracy. Therefore, the use of additional metrics (SSIM, LPIPS) for comparing prediction and ground truth images after post-training are useful. Among the other metrics, LPIPS stands as one of the most state-of-the-art but there are still some drawbacks of this metric. A previous study indicates that LPIPS is susceptible to such imperceptible adversarial perturbations where the LPIPS values are significantly affected by adding some noise or manipulating just a single pixel [32].

5 Conclusion

In this study, the voxel data allowed for calculating the amount building volumes and represent it in 2D raster format for each air temperature pixel region. Since these building volumes are represented in raster format, it becomes possible to apply image processing techniques such as Gaussian blurring. Gaussian blurring enabled the integration of spatial neighborhood relationships before the training process, as the value of a pixel is influenced by adjacent pixel values. Thereby, the correlation rate is also increased between air temperature data and building volume data after implementing Gaussian blur.

Additionally, the proposed CityGML-to-voxel conversion steps facilitated the rapid generation of city-scale 3D volumetric data, accelerating experiments by allowing quick acquisition of voxel data from different regions. This rapid conversion might enables urban planners to quickly implement their plans in digital applications and observe the impact on environmental indicators.

In addition, as demonstrated by previous studies, when model accuracy is evaluated using the MSE metric, the XGBoost method produces models with low error rates. However, thanks to the raster format of our rasterized methodology, these errors are observed to be systematic and the spatial distributions are not well captured. On the other hand, the Random Forest method has higher MSE values, it qualitatively demonstrates more consistent spatial patterns and the error distribution appears more randomly. Furthermore, the raster-based format of both predictions and ground truth data allowed for the implementations of image similarity metrics like SSIM or LPIPS that also capture and consider spatial relationships for evaluation process. Since Random Forest method qualitatively provided better results than XGBoost method, these image similarity metrics proves that qualitative finding in a quantitative manner. The SSIM and LPIPS scores for each city, indicating that the prediction results of the Random Forest method are better than XGBoost. Besides, using the testing data confirms that these accurate patterns are not due to overfitting but rather indicate good generalization.

Taking all of these into consideration, it has been demonstrated that previous studies relying solely on MSE to evaluate their models are inadequate and lack depth. The raster-based data driven approach presented here takes into account spatial neighborhood relationships both before and after training, resulting in more accurate outcomes and insightful analyses. The models used in this study kept simple without any multi-variable or multi-temporal sources to highlight the benefits of the data driven methodology, which is intended to serve as a foundation for future research and has potential applications in a variety of fields.

References

[1] Yilong Han, John E Taylor, and Anna Laura Pisello. "Toward mitigating urban heat island effects: Investigating the thermal-energy impact of bio-inspired retro-reflective building envelopes in dense urban settings". In: *Energy and Buildings* 102 (2015), pp. 380–389.

[2] Yuliang Lan and Qingming Zhan. "How do urban buildings impact summer air temperature? The effects of building configurations in space and time". In: *Building and Environment* 125 (2017), pp. 88–98.

[3] Lin Liu et al. "Climate-conscious spatial morphology optimization strategy using a method combining local climate zone parameterization concept and urban canopy layer model". In: *Building and Environment* 185 (2020), p. 107301.

[4] Zheng Ren, Bin Jiang, and Stefan Seipel. "Capturing and characterizing human activities using building locations in America". In: *ISPRS International Journal of Geo-Information* 8.5 (2019), p. 200.

[5] Chen Zhong et al. "Revealing centrality in the spatial structure of cities from human activity patterns". In: *Urban Studies* 54.2 (2017), pp. 437–455.

[6] Alireza Attarhay Tehrani et al. "Predicting urban Heat Island in European cities: A comparative study of GRU, DNN, and ANN models using urban morphological variables". In: *Urban Climate* 56 (2024), p. 102061.

[7] A John Arnfield. "Two decades of urban climate research: a review of turbulence, exchanges of energy and water, and the urban heat island". In: *International Journal of Climatology: a Journal of the Royal Meteorological Society* 23.1 (2003), pp. 1–26.

[8] Zhiwei Yang et al. "Application of building geometry indexes to assess the correlation between buildings and air temperature". In: *Building and Environment* 167 (2020), p. 106477.

[9] Ingegärd Eliasson. "The use of climate knowledge in urban planning". In: *Landscape and urban planning* 48.1-2 (2000), pp. 31–44.

[10] Edward Ng. “Towards planning and practical understanding of the need for meteorological and climatic information in the design of high-density cities: A case-based study of Hong Kong”. In: *International Journal of Climatology* 32.4 (2012), pp. 582–598.

[11] Soheil Fathi et al. “Machine learning applications in urban building energy performance forecasting: A systematic review”. In: *Renewable and Sustainable Energy Reviews* 133 (2020), p. 110287.

[12] Lun Liu et al. “A machine learning-based method for the large-scale evaluation of the qualities of the urban environment”. In: *Computers, environment and urban systems* 65 (2017), pp. 113–125.

[13] Stephane Cedric Koumetio Tekouabou et al. “Reviewing the application of machine learning methods to model urban form indicators in planning decision support systems: Potential, issues and challenges”. In: *Journal of King Saud University-Computer and Information Sciences* 34.8 (2022), pp. 5943–5967.

[14] Zander S Venter et al. “Hyperlocal mapping of urban air temperature using remote sensing and crowdsourced weather data”. In: *Remote Sensing of Environment* 242 (2020), p. 111791.

[15] Sung J Yoo, Taeyong Kwon, and Young S Lyoo. “Challenges of influenza A viruses in humans and animals and current animal vaccines as an effective control measure”. In: *Clinical and experimental vaccine research* 7.1 (2018), pp. 1–15.

[16] Chengliang Fan et al. “Exploring the relationship between air temperature and urban morphology factors using machine learning under local climate zones”. In: *Case Studies in Thermal Engineering* 55 (2024), p. 104151.

[17] Tsz-Kin Lau and Tzu-Ping Lin. “Investigating the relationship between air temperature and the intensity of urban development using on-site measurement, satellite imagery and machine learning”. In: *Sustainable Cities and Society* 100 (2024), p. 104982.

[18] Zhong Zheng et al. “The higher, the cooler? Effects of building height on land surface temperatures in residential areas of Beijing”. In: *Physics and Chemistry of the Earth, Parts A/B/C* 110 (2019), pp. 149–156.

[19] Yunfeng Hu, Zhaoxin Dai, and Jean-Michel Guldmann. “Modeling the impact of 2D/3D urban indicators on the urban heat island over different seasons: A boosted regression tree approach”. In: *Journal of environmental management* 266 (2020), p. 110424.

[20] Huifang Li et al. “Quantifying 3D building form effects on urban land surface temperature and modeling seasonal correlation patterns”. In: *Building and Environment* 204 (2021), p. 108132.

[21] Timothy R Oke. “The heat island of the urban boundary layer: characteristics, causes and effects”. In: *Wind climate in cities* (1995), pp. 81–107.

[22] Ian D Stewart and Tim R Oke. “Local climate zones for urban temperature studies”. In: *Bulletin of the American Meteorological Society* 93.12 (2012), pp. 1879–1900.

[23] James A Voogt and Tim R Oke. “Thermal remote sensing of urban climates”. In: *Remote sensing of environment* 86.3 (2003), pp. 370–384.

[24] Jianguo Wu. Urban sustainability: an inevitable goal of landscape research. 2010.

[25] Weiqi Zhou, Ganlin Huang, and Mary L Cadenasso. “Does spatial configuration matter? Understanding the effects of land cover pattern on land surface temperature in urban landscapes”. In: *Landscape and urban planning* 102.1 (2011), pp. 54–63.

[26] Anqi Lin et al. “How does urban heat island differ across urban functional zones? Insights from 2D/3D urban morphology using geospatial big data”. In: *Urban Climate* 53 (2024), p. 101787.

[27] Biao Liu, Xian Guo, and Jie Jiang. “How urban morphology relates to the urban heat island effect: A multi-indicator study”. In: *Sustainability* 15.14 (2023), p. 10787.

[28] Tom Raaymakers. “Understanding Urban Temperature Differences through 2D/3D Urban Morphology”. In: (2024).

[29] Nurul Amirah Isa et al. “Building Volume Effects on Ambient Temperature In The Kuala Lumpur City”. In: *IOP Conference Series: Earth and Environmental Science*. Vol. 489. 1. IOP Publishing. 2020, p. 012011.

[30] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[31] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[32] Abhijay Ghildyal and Feng Liu. “Attacking perceptual similarity metrics”. In: *arXiv preprint arXiv:2305.08840* (2023).

[33] Matthew Tancik et al. "Nerfstudio: A modular framework for neural radiance field development". In: *ACM SIGGRAPH 2023 Conference Proceedings*. 2023, pp. 1–12.

[34] James A Voogt and TR Oke. "Effects of urban surface geometry on remotely-sensed surface temperature". In: *International Journal of Remote Sensing* 19.5 (1998), pp. 895–920.

[35] Zhi Cai, Guifeng Han, and Mingchun Chen. "Do water bodies play an important role in the relationship between urban form and land surface temperature?" In: *Sustainable cities and society* 39 (2018), pp. 487–498.

[36] Guanhua Guo et al. "Characterizing the impact of urban morphology heterogeneity on land surface temperature in Guangzhou, China". In: *Environmental Modelling & Software* 84 (2016), pp. 427–439.

[37] Chih-Da Wu, Shih-Chun Candice Lung, and Jihn-Fa Jan. "Development of a 3-D urbanization index using digital terrain models for surface urban heat island effects". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 81 (2013), pp. 1–11.

[38] Chao Ren et al. "Developing a rapid method for 3-dimensional urban morphology extraction using open-source data". In: *Sustainable Cities and Society* 53 (2020), p. 101962.

[39] Medhini Heeramaglore and Thomas H Kolbe. "Semantically enriched voxels as a common representation for comparison and evaluation of 3D building models". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 10 (2022), pp. 89–96.

[40] DT Mulder. "Automatic Repair of 3D City Building Model Using a Voxel–based Repair Method". PhD thesis. Master Thesis, Delft Univ. Technology, Netherlands, 2015.

[41] Pirouz Nourian et al. "Voxelization algorithms for geospatial applications: Computational methods for voxelating spatial datasets of 3D city models containing 3D surface, curve and point data models". In: *MethodsX* 3 (2016), pp. 69–86.

[42] BRUNO Willenborg, Maximilian Sindram, and THOMAS H Kolbe. "Semantic 3D city models serving as information hub for 3D field based simulations". In: *Lösungen für eine Welt im Wandel* (2016), pp. 54–65.

[43] Maximilian Sindram et al. "Voluminator 2.0–speeding up the approximation of the volume of defective 3D building models". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 3 (2016), pp. 29–36.

[44] Amol Konde and Sameer Saran. “Web enabled spatio-temporal semantic analysis of traffic noise using CityGML”. In: *J Geomatics* 11.2 (2017), pp. 248–59.

[45] Nurfairunnajiha Ridzuan, Uznir Ujang, and Suhaibah Azri. “3D vectorization and rasterization of CityGML standard in wind simulation”. In: *Earth Science Informatics* 16.3 (2023), pp. 2635–2647.

[46] M.W. Jahn and P.E. Bradley. “Computing watertight volumetric models from boundary representations to ensure consistent topological operations”. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* VIII-4/W2-2021 (2021), pp. 21–28.

[47] M.W. Jahn and P.E. Bradley. “A Robustness Study for the Extraction of Watertight Volumetric Models from Boundary Representation Data”. In: *ISPRS International Journal of Geo-Information* 11.4 (2022), p. 224.

[48] M. Jahn. “Distributed & Parallel Data Management to Support Geo-Scientific Simulation Implementations”. PhD thesis. Karlsruhe Institute of Technology, 2022.

[49] Klaus Pusacker et al. “A Concept for 3D Geological and Urban Subsurface Modeling with a Unified Voxel Model Examined by a Case Study for the City Center of Stuttgart (Baden-Württemberg), Germany”. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 10 (2024), pp. 193–200.

[50] Janina Konarska et al. “Influence of vegetation and building geometry on the spatial variations of air temperature and cooling rates in a high-latitude city.” In: *International Journal of Climatology* 36.5 (2016).

[51] Ho Jong Kim et al. “A study on the effectiveness of spatial filters on thermal image pre-processing and correlation technique for quantifying defect size”. In: *Sensors* 22.22 (2022), p. 8965.

[52] Marius Zumwald et al. “Mapping urban temperature using crowd-sensing data and machine learning”. In: *Urban Climate* 35 (2021), p. 100739.

[53] Open source CityGML data of Thuringia/Germany. https://geoportal.thueringen.de/gdi-th/download-offene-geodaten/download-3d-gebaeudedaten.

[54] S Krähenmann et al. “High-resolution grids of hourly meteorological variables for Germany”. In: *Theoretical and Applied Climatology* 131 (2018), pp. 899–926.

[55] DWD Climate Data Center (CDC): Annual mean of station observations of daily air temperature at 2 meter above ground in °C for Germany.
https://opendata.dwd.de/climate_environment/CDC/grids_germany/hourly/hostrada/air_temperature_mean/.

[56] Sandia National Labs, Kitware Inc, and Los Alamos National Labs. Paraview: Parallel visualization application.
https://www.paraview.org/.

[57] QGIS Development Team. QGIS Geographic Information System. QGIS Association. URL: https://www.qgis.org.

[58] Stelios Vitalis, Ken Arroyo Ohori, and Jantien Stoter. "CityJSON in QGIS: Development of an open-source plugin". In: *Transactions in GIS* 24.5 (2020), pp. 1147–1164.

[59] GDAL/OGR contributors. GDAL/OGR Geospatial Data Abstraction software Library. Open Source Geospatial Foundation, 2024. DOI: 10.5281/zenodo.5884351. URL: https://gdal.org.

[60] Hugo Ledoux et al. "CityJSON: A compact and easy-to-use encoding of the CityGML data model". In: *Open Geospatial Data, Software and Standards* 4.1 (2019), pp. 1–12.

[61] Ghazaleh Tanoori, Ali Soltani, and Atoosa Modiri. "Machine Learning for Urban Heat Island (UHI) Analysis: Predicting Land Surface Temperature (LST) in Urban Environments". In: *Urban Climate* 55 (2024), p. 101962.

[62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).