Urban Air Temperature Modeling: Combining Physical Simulation and Data-Driven Fine-Tuning

Hiba Hamdi, Thomas Corpetti [1], Laure Roupioz [2], Xavier Briottet [3]

[1] LETG Rennes, *Université Rennes 2 - Haute Bretagne, CNRS*

[2] ONERA (DOTA), *ONERA (DOTA)*
[3] ONERA / DOTA, *Université de Toulouse*

Keywords: Air temperature, Deep learning, Urban Weather Generator (UWG), Urban Heat Islands (UHI), Neural UWG-City (NUWG-City).

Abstract

Accurate urban climate modeling is crucial for addressing the growing impacts of urban heat islands (UHI) and climate change. Physics-based tools such as the Urban Weather Generator (UWG) are widely used but often limited by high parameterization needs and a lack of specialized data. In this study, we develop a hybrid framework combining UWG simulations with deep learning, introducing two models: NUWG-Sim (Neural Urban Weather Generator on Simulations), trained solely on simulated data, and NUWG-city, which is fine-tuned with ground weather station data. To systematically evaluate model performance across heterogeneous urban contexts, we structure our experiments around Local Climate Zones (LCZs) in Toulouse, France. Our methodology involves generating over 3400 UWG initialization files, simulating urban air temperatures time series for diverse surface parameters, and training a neural model on these series. We then fine-tune the model with observed data from selected weather stations, analyzing how the number and diversity of stations environments impact performance on unseen stations from different LCZs. Results show that even limited fine-tuning significantly improves performance, particularly when training includes stations from LCZs similar to the test set. The approach highlights the potential of physics-informed neural models for city-specific urban climate monitoring.

1. Introduction

Earth's climate has naturally fluctuated over millions of years, with long periods of warming and cooling, shaping ecosystems and life on the planet (Zachos et al., 2001). In contrast, current climate change is unfolding at an unprecedented pace, primarily due to human activities (Trenberth, 2018). A key driver of this shift is global warming, a sustained rise in average global temperatures with far-reaching consequences.

Urban areas are particularly vulnerable to climate change impacts, especially through the Urban Heat Island (UHI) effect. The effects of heatwaves are especially severe in urbanised areas. These environments, predominantly composed of artificial surfaces, amplify rising temperatures and give rise to the UHI phenomenon characterized by higher air temperatures in cities compared to their rural surroundings (Sobrino et al., 2013). The intensity of the UHI varies within a city due to its heterogeneous urban form (Gao et al., 2022), highlighting the need to pinpoint neighbourhoods that are particularly vulnerable in order to implement targeted heat mitigation strategies. While air temperature is commonly measured in urban settings, fixed weather stations cover only limited areas and fall short in capturing the fine-scale spatial variability of UHI. Policymakers would greatly benefit from high-resolution air temperature maps that offer a detailed overview of thermal patterns across the city. As (Buttsta¨dt et al., 2011) notes, such spatially refined data are crucial for identifying hotspots and understanding the underlying drivers of urban thermal conditions.

Modeling urban climates with high spatial and temporal resolution is key for city planning and mitigation strategies. On one hand, tools like the Urban Weather Generator (UWG) offering a physics-based approach, are limited by complex parameterization and struggle to adapt to real-time urban heterogeneity due to their high computational cost. In contrast, physical models, grounded in atmospheric laws, offer greater reliability across diverse conditions (Ja¨nicke et al., 2021). On the other hand, data-driven models are efficient and scalable for estimating air temperatures, making them ideal for real-time and large-scale applications. However, they often lack physical interpretability and may struggle with rare weather events or unfamiliar environments. With the increasing availability of urban climate data and advances in machine learning, hybrid models can leverage both physics-based model and measured data.

This study proposes a novel hybrid modeling framework that couples UWG with deep learning. We develop two models: NUWG-Sim, trained purely on synthetic data from UWG, and NUWG-city, fine-tuned using observations from urban weather stations. To analyze the influence of urban heterogeneity, we structure the evaluation around Local Climate Zones (LCZs) (Stewart and Oke, 2012), a standardized classification of urban and natural landscapes based on both their physical characteristics and their thermal behavior. The main contributions of this study are:

- A neural network trained to simulate UWG outputs by varying key surface parameters : NUWG-Sim
- A neural network trained on UWG simulations, then fine-tuned on real observations: NUWG-City
- A systematic evaluation of fine-tuning strategies based on LCZ composition.

2. Methodology

2.1 Simulated Data Generation with UWG

The generation of simulated data is crucial for training the neural model, as it provides a foundational dataset with controlled conditions. To ensure that the model is trained effectively and can generalize across a range of urban environments, we adopted a detailed experimental framework using the UWG, a widely recognized tool for simulating urban microclimates. It is particularly well-suited for simulating urban air temperatures based on urban surface characteristics. Using physically based equations, it extrapolates air temperature values in urban areas from a well-chosen rural reference station (Bueno et al., 2013). This subsection delves into the step-by-step process of how we generated the simulated data and highlights the challenges and solutions encountered throughout the experiment.

The first step in generating the simulated data involves defining the urban surface characteristics that influence air temperature. The key parameters selected for this experiment were building density, building height, vertical-to-horizontal ratio (VerToHor), grass cover, and tree cover. These parameters were chosen based on their significant impact on the urban heat island effect, which is a primary focus of the study. A comprehensive sensitivity analysis was conducted on these parameters to understand their individual contributions to the simulated urban air temperatures (Hamdi et al., 2024). The next phase involved varying these parameters within the typical ranges observed in urban environments. These variations were informed by real-world data and represented a wide spectrum of urban configurations. To simulate diverse urban environments accurately, a total of 3450 initialization files were generated, each corresponding to a unique combination of these surface and morphological parameters. Each simulation covered a 3-day period between May and September 2020, a timeframe selected to capture representative summer conditions that strongly influence urban climate dynamics.

A key challenge in this process was managing the interdependencies between certain parameters. For example, the vertical-to-horizontal (verToHor) ratio cannot be sampled independently over a fixed range, as it depends on building height and fac¸ade length. Sampling it randomly would lead to many unrealistic and physically inconsistent urban configurations. To address this, we adopted a more controlled strategy by training a linear regression model to predict VerToHor based on other parameters, including building height, building density, grass cover, and tree cover. This method ensured that all generated values remained coherent and representative of actual urban environments.

Additionally, ensuring the physical validity of each parameter combination was essential during the simulation process. Given the wide range of configurations explored, some led to unrealistic outputs, such as abnormally high or low temperatures. To address this, we implemented a set of pre-simulation checks to filter out combinations that did not align with physically plausible urban conditions. This allowed us to retain only consistent and reliable simulations for model training.

The resulting simulated dataset formed the backbone of the training data for the neural model. By systematically varying the key urban surface parameters, we ensured that the model was exposed to a wide range of urban conditions, increasing its ability to generalize to unseen environments. However, the simulated data alone was not sufficient. It needed to be combined with real-world data to ensure that the final model could accurately reflect actual urban climates, which we address in subsequent sections.

In conclusion, the simulated data generation process using UWG was an essential step in the development of the final neural model. It not only provided a large and diverse dataset for model training but also highlighted the challenges involved in simulating urban climates, such as managing parameter dependencies and computational anomalies. The insights gained from this simulation process have been invaluable in refining the experimental design and preparing the dataset for subsequent integration with real-world data.

2.2 Data Preprocessing

Data preprocessing plays a critical role in preparing the dataset for training machine learning models. This step is essential for ensuring that the data is clean, complete, and in a suitable format for model ingestion.
Real-world data were preprocessed with careful attention to handling missing values, outliers, and temporal consistency, as well as aligning data from various sources. The primary goal of the preprocessing stage was to make sure that the dataset used to train and test the model would accurately represent the underlying urban weather conditions while minimizing the risk of introducing biases that could negatively affect model performance.

Handling Missing Data and Outliers: One of the primary challenges in preprocessing was dealing with missing data, which is a common occurrence in real-world weather datasets. Missing values can arise from sensor malfunctions, data transmission errors, or simply due to environmental factors that prevent measurements from being recorded. In the case of both rural and urban weather stations, missing data points were identified and handled based on the specific characteristics of the data and the station.

For rural stations, data gaps were filled using linear interpolation, provided that there were no more than five consecutive missing data points within a given 15-minute interval. This interpolation method ensured that the temporal continuity of the data was preserved while minimizing the introduction of artificial trends. However, in cases where the missing data exceeded five consecutive points, no interpolation was applied, and the missing values were retained as 'NaN' (Not a Number). This cautious approach prevented the imposition of unrealistic data points and maintained the integrity of the dataset.

For urban stations, missing values did not pose a significant challenge, as the absence of certain data points merely resulted in less available data for training, validation, and testing. No interpolation was performed for urban stations; instead, the missing values were simply ignored during model training, and the model was trained with the available data. This approach was consistent with the general principle that real-world urban weather data often have gaps, and machine learning models should be capable of handling such scenarios without overfitting or biasing the results.

In addition to managing missing values, we addressed outliers in both observed and simulated temperature data. For rural and urban stations, outliers were identified using fixed thresholds, set to missing, and interpolated. Simulated data from UWG often exhibit early morning anomalies caused by instabilities during the night-to-day transition. These were corrected by detecting abrupt slope changes using first and second derivatives; affected values and their neighbors were masked and interpolated with a second-degree polynomial. This preprocessing step improves data continuity and enhances the quality of inputs for the neural model training.

Temporal Aggregation: Weather station data, originally recorded at 15-minute intervals, were aggregated to hourly intervals to match the temporal resolution used by UWG. This aggregation ensured consistency across datasets. Variable-specific methods were applied:

For variables like temperature, humidity, and pressure, which are continuous and typically show gradual changes, the data were aggregated by computing the average value over the 15-minute intervals. This method ensured that the temporal resolution of the data was reduced without losing important trends. For wind speed and direction, however, the last recorded value within the 15-minute window was used. Rainfall, being an accumulative variable, was summed over the 15-minute periods to obtain a total value for each hour. This aggregation strategy ensured that each variable was treated in a way that preserved its underlying characteristics, making the data suitable for model training.

Merging Data from Multiple Sources: One of the unique aspects of this study was the integration of data from multiple sources. We merged data from the Meteopole station with each rural station studied to create complete input datasets. From the Meteopole, we extracted additional variables such as solar radiation and ground temperature, which were not available at the other rural stations. Conversely, the other rural stations provided key meteorological variables including air temperature, wind speed, humidity, and pressure. This merging was justified by the fact that the study focused on the summer period, during which sky conditions were mostly clear and cloud cover minimal. As a result, it was reasonable to assume that radiation data did not vary significantly across different locations within the Toulouse area. The merging process involved aligning datasets based on timestamps and ensuring consistency in variable formats and units, ultimately allowing us to construct unified and coherent datasets for training and evaluating our models.

Handling Temporal Overlaps: A critical aspect of data preprocessing was ensuring that no time-series overlaps existed between the training, validation, and test sets. This data split was essential for maintaining the integrity of the data split and ensuring that the model was evaluated on truly unseen data. Any instances where the same data point appeared in both the training and test sets were removed. This process helped avoid data leakage and ensured that the model's performance metrics were accurate and reflective of its ability to generalize to new, unseen time periods.

2.3 Neural Network Model

After simulating a large synthetic dataset with UWG and cleaning all datasets (both simulated and observed), a neural network model is trained to learn the relationship between rural weather conditions, urban morphology, and simulated urban air temperature. This initial model, NUWG-Sim, leverages the physical consistency and broad coverage of UWG outputs. In a second step, NUWG-Sim is fine-tuned using real measurements from urban weather stations in the target city (Toulouse), yielding NUWG-City, which integrates local climate characteristics and better captures site-specific microclimate effects. Throughout this process, inputs include real-time rural station observations and urban surface parameters (e.g., building density, vegetation cover). Artificial neural networks (ANN) are particularly well suited for simulating air temperatures. For example, (Snell et al., 2000) employed a multilayer perceptron to spatially interpolate daily maximum air temperatures and found that the ANN outperformed conventional methods (spatial average, nearest neighbor, and inverse distance weighting) in 94% of cases.

The neural model is built around a dual-branch architecture that separates the processing of temporal meteorological signals from static urban surface and morphological descriptors, then merges them to produce a 36-hour urban air temperature simulation. In the first branch, 72 hours of 16 rural weather variables (e.g., air temperature, humidity, pressure, wind speed) are fed into a stack of 1D convolutional layers. Each convolution is followed by non-linear activations (LeakyReLU), pooling for dimensionality reduction, and later up-sampling to recover temporal resolution, allowing the network to learn rich, local temporal patterns without relying on longer-term memory mechanisms.

Parallel to this, the second branch handles surface and morphological parameters, such as building density, tree cover, and vertical-to-horizontal ratios, as a 1D vector input into a multilayer perceptron (MLP). Through successive dense layers and activations, the MLP distills spatial features reflective of urban morphology. At multiple fusion points, the feature maps from both branches are concatenated, then further processed via convolutional and up-sampling blocks, enabling the model to jointly leverage temporal trends and spatial heterogeneity when generating hourly temperature estimates. Convolutional neural networks (CNNs) were selected over alternatives such as LSTMs or GRUs due to their versatility and strong track record in the literature. CNNs are particularly well-suited for processing both time series data (Pelletier et al., 2019) and imagery (Maggiori et al., 2016), making them an ideal choice for this study and for future extensions. Their flexible architecture facilitates the integration of remote sensing imagery or digital terrain models, ensuring that the core model can scale to accommodate richer geospatial inputs over time.

Custom Loss Function: Our neural network is trained to simulate hourly air temperature from the surface parameters and time. We present in this section the custom loss function developed to train the neural network for simulating urban air temperature. The model is trained by minimising a loss that combines three complementary components to improve prediction performance.
First, the Mean Absolute Error (MAE) is used to quantify the average absolute difference between predicted and observed temperatures. As a standard choice in regression tasks, MAE ensures overall accuracy by encouraging the model to minimize deviations from the actual values.
However, for meteorological time series such as urban air temperature, it is also essential to capture the temporal dynamics and not just pointwise accuracy.This motivates the inclusion of a second component: the MAE on the gradient, which compares the rate of change (i.e., the slope) of the predicted and true temperature series. This term enhances the model's sensitivity to sudden variations, such as those caused by atmospheric processes like changes in wind or cloud cover, and helps the network learn more realistic temperature evolution patterns over time.
The third component, the cosine loss, addresses a different but equally important aspect. It measures the directional similarity between predicted and observed temperature vectors and is particularly useful in detecting whether the model correctly captures the general shape of temperature trends, even in the presence of time shifts. As shown in previous work on hyperspectral data analysis (Zhang et al., 2012), cosine similarity is effective

when the direction of variation is more relevant than the magnitude. In this context, it helps the model better align its outputs with the overall trend of the observed time series.

The final loss function is defined as a weighted sum of these three components, with coefficients controlling their respective contributions. This combined approach ensures that the model not only achieves good numerical accuracy but also learns to reproduce realistic temporal behaviors and directional patterns, features that are particularly important when modeling air temperature dynamics in urban environments.

In conclusion, the custom loss function designed is combining three components:

- Mean Absolute Error (MAE)
- Gradient MAE to emphasize changes
- Cosine loss to encourage correct temporal patterns

The final loss is a weighted sum:

$$L = \alpha \cdot MAE + \beta \cdot MAE_{\nabla} + \gamma \cdot Cosine \tag{1}$$

where $\alpha = 0.9$
$\beta = 0.1$
$\gamma = 0.5$

2.4 Model Training and Evaluation

The model fine-tuning and evaluation process is crucial to adapt the neural network model to the specific urban conditions and ensure its ability to generalize to unseen environments. The primary objective of fine-tuning is to optimize the model's performance on real-world weather data while maintaining the physical consistency introduced by the simulated UWG data. In this section, we describe the steps taken to fine-tune the NUWG-Sim model using observed weather station data and how the model's performance was evaluated through multiple experiments.

Fine-Tuning NUWG-Sim Model: The NUWG-Sim model was first trained on simulated data generated by the UWG model. This step was essential for providing the model with a general understanding of urban climate dynamics and ensuring that it could simulate temperature variations based on key urban surface parameters. However, since simulated data alone may not fully capture the complexities of real-world urban environments, fine-tuning the pre-trained model on actual weather station data is necessary.

Fine-tuning involves adjusting the weights of the pre-trained model by training it on a smaller, real-world dataset. In our case, we used observed air temperature data from various weather stations in Toulouse. Fine-tuning allows the model to adapt its internal representations to the specific conditions of the city, improving its ability to simulate local weather patterns more accurately.

To fine-tune the model, we used a strategy where the learning rate was adjusted to prevent the model from overfitting to the real station data. A lower learning rate was used to preserve the knowledge learned during pre-training, while allowing the model to gradually adjust its parameters to fit the real data.

Evaluation of Model Performance: The evaluation process is designed to test how well the fine-tuned model generalizes to new, unseen data. Evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Standard Deviation (STD) were used to assess the model's accuracy and predictive power. These metrics provide a comprehensive understanding of how close the model's predictions are to the true observed values, with lower MAE and RMSE values indicating better model performance. These evaluation metrics were calculated for each station, with particular emphasis placed on stations that had more challenging conditions or higher variability, such as urban parks with dense vegetation (e.g., Compans-Cafarelli) or areas with limited station data.

The temporal coherence of the data splits was ensured by aligning the data from both simulated and real weather stations, applying the same time period split across both types of data. This process ensured that the model was evaluated on novel data that had never been seen during training or validation. This strategy was critical for testing the generalization ability of the model and ensuring that it could effectively predict temperature values for future, unseen time periods.

Cross-Station Evaluation and Generalization: To assess the generalization ability of the fine-tuned model, it was evaluated on a set of stations that were not included in the training process. This cross-station evaluation allowed us to determine whether the model was able to adapt to different LCZs (Local Climate Zones) and provide reliable predictions across diverse urban configurations. For instance, urban parks with high vegetation density, like Compans-Cafarelli, were compared against more built-up areas, such as Avenue de Grande Bretagne, to evaluate how well the model performed across different LCZs.

Model Validation Across LCZs: One of the key objectives of this study was to investigate how well the model generalizes across different LCZs. As part of the fine-tuning process, we performed experiments to evaluate the impact of training on stations with different LCZs. We trained the model on stations located within similar LCZs and tested it on stations located in different LCZs to examine the model's ability to generalize across diverse urban environments. The results were compared to determine the optimal number of stations needed for robust model performance and whether training on multiple stations with different LCZs improved the model's ability to adapt to new environments.

Fine-Tuning Strategies: Different fine-tuning strategies were explored to determine the optimal configuration for the model. These strategies included:

- **Single-Station Fine-Tuning:** In this setup, the model was fine-tuned using data from one station at a time. This approach was useful for investigating how well the model could adapt to the characteristics of individual stations and how much information could be extracted from a single source.
- **Pairwise Fine-Tuning (Extreme LCZs):** In this case, the model was fine-tuned using data from two stations located in ”extreme” LCZs, such as areas with high building density or dense vegetation. This setup aimed to assess whether training on contrasting LCZs improved the model's robustness to different urban conditions.

- **Multi-Station Fine-Tuning (Extreme + Average LCZs):** This strategy involved fine-tuning the model on a combination of extreme LCZs and a middle-range LCZ, such as stations located in moderately built-up areas. This setup aimed to strike a balance between the extremes of urban environments and typical city conditions.

- **All-Station Fine-Tuning:** This final strategy involved training the model on data from all available stations, regardless of their LCZ classification. This approach tested the model's ability to generalize across all types of urban environments and maximize the diversity of training data. It also aimed to evaluate the overall contribution of combining all stations in improving predictive performance.

3. Results

3.1 Evaluation of NUWG-Sim

The baseline NUWG-Sim model, trained solely on UWG-generated data, was evaluated on 6 held-out urban stations in Toulouse. Overall, its accuracy varies with local morphology: in densely built areas (LCZ 2 and 8) it achieves acceptable errors, but performance degrades markedly in heavily vegetated, low-density zones. For instance, at Compans-Cafarelli (6% tree cover) NUWG-Sim yields an RMSE of 2.19 °C, while at Cote Pavée (13% building density, 29% tree cover) it achieves 1.85°C versus 2.22°C at Parc Maourine (5% building density, 21% tree cover) (See Table 1). Compared to raw UWG simulations, NUWG-Sim shows a modest RMSE increase of about 0.2 °C (mean RMSE: 1.81°C for UWG vs. 2.02°C for NUWG-Sim), but delivers a 33% speedup, simulating a three-day series in 0.7s instead of 1.06s. These results establish a physically consistent yet computationally efficient baseline, highlighting the need to incorporate real-world data, particularly from LCZs, where UWG struggles to simulate accurate air temepetaures, to improve generalization across heterogeneous urban microclimates.

Trained on (data points) / **Tested on**	**Metrics** (STD) in ◦C	**Simulated data from UWG** (101520)
Carmes	MAE	**1.46** (0.03)
	RMSE	**1.88** (0.07)
	STD	**1.84** (0.09)
Thibaud	MAE	**1.53** (0.03)
	RMSE	**2.00** (0.04)
	STD	**1.97** (0.05)
Sabatier	MAE	**1.48** (0.09)
	RMSE	**1.93** (0.13)
	STD	**1.92** (0.14)
St-Exupery	MAE	**1.58** (0.05)
	RMSE	**2.08** (0.09)
	STD	**2.07** (0.10)
Busca	MAE	**1.67** (0.02)
	RMSE	**2.05** (0.02)
	STD	**1.89** (0.07)
Compans-Cafarelli	MAE	**1.70** (0.03)
	RMSE	**2.19** (0.07)
	STD	**2.19** (0.06)
Mean ± std	MAE	**1.57** ± 0.10
	RMSE	**2.02** ± 0.11
	STD	**1.98** ± 0.13

Table 1. Comparison of results of models trained on UWG simulated data solely, tested on test set stations in Toulouse.

3.2 NUWG-City Evaluation : Fine-Tuning Experiments with observed data

We fine-tune NUWG-Sim on observed data collected from urban weather stations in Toulouse. These stations are grouped into two sets covering diverse LCZs (a training/validation set and a test set) and several strategies are tested:

- Fine-tuning on a single station
- Using two stations with extreme LCZs
- Adding an average LCZ to the extreme pair
- Using all stations from the training group

Each fine-tuned model is evaluated on the evaluate group to test generalizability across LCZs.

Model Performance and Analysis: After fine-tuning and evaluation, the results were analyzed to determine the impact of different fine-tuning strategies on model performance. The results indicated that the best performance was achieved by training the model on a combination of stations from various LCZs, as this approach allowed the model to learn from a more diverse range of urban conditions. The performance metrics showed that the model was able to generalize well to stations located in different LCZs, with a significant reduction in error compared to the baseline model trained solely on simulated data. The results confirmed that including diverse LCZs in the training set notably improved model performance. The best results were obtained when training included stations from both extreme LCZs (Valade and Parc Maourine) along with an intermediate one (Avenue de Grande Bretagne). This combination achieved the lowest average MAE of 0.87°C and RMSE of 1.25°C across test stations. For instance, at Sabatier, this strategy reduced the MAE from 0.94°C (using only extreme LCZs) to 0.82°C, and at Busca from 1.10°C to 0.89°C (see Table 2).

In summary, the fine-tuning and evaluation process was essential for adapting the NUWG model to real-world urban conditions. By adjusting the model's weights using real weather station data and evaluating it across various LCZs, we ensured that the model was able to generalize effectively to unseen environments. The combination of simulated data pre-training and real-world data fine-tuning provided a powerful approach to simulating urban weather conditions, and the evaluation results confirmed that the model can successfully predict air temperatures in diverse urban settings. While training on all available stations yielded comparable performance (average MAE of 0.89°C), it did not consistently outperform the more strategic LCZ selection. This suggests that LCZ diversity is more important than simply increasing the volume of training data. Overall, these results underscore the importance of thoughtful fine-tuning. Incorporating stations with contrasting LCZs enhances the model's ability to generalize across different urban environments, while training solely on green zones or parks (e.g., Compans-Cafarelli or Parc Maourine) limits generalization to denser urban fabrics, and vice versa.

Our results show that NUWG-city significantly outperforms NUWG-Sim on test stations, especially when training includes LCZs similar to the test site. Training on extreme LCZs plus one middle LCZ often yields the best compromise. Models trained on very green urban parks (e.g., with high tree cover) generalize poorly to denser zones and vice versa, confirming the need to consider LCZ diversity in fine-tuning.

Trained o (data points **Tested on**	**Metrics (°C)**	**Valade & Parc Maourine**	**Valade & P. Maourine & Av. Gd Bretagne**	**All stations**
Carmes	MAE	0.83 (0.02)	0.80 (0.04)	0.78 (0.02)
	RMSE	1.21 (0.01)	1.17 (0.04)	1.16 (0.02)
	STD	1.20 (0.01)	1.16 (0.02)	1.16 (0.02)
Thibaud	MAE	0.93 (0.02)	0.85 (0.02)	0.86 (0.03)
	RMSE	1.35 (0.02)	1.22 (0.03)	1.23 (0.06)
	STD	1.35 (0.02)	1.22 (0.02)	1.20 (0.01)
Sabatier	MAE	0.94 (0.03)	0.82 (0.02)	1.04 (0.18)
	RMSE	1.35 (0.04)	1.22 (0.02)	1.43 (0.17)
	STD	1.31 (0.02)	1.22 (0.02)	1.37 (0.14)
St-Exupery	MAE	0.97 (0.02)	1.00 (0.02)	0.89 (0.03)
	RMSE	1.38 (0.03)	1.41 (0.03)	1.29 (0.03)
	STD	1.38 (0.03)	1.37 (0.05)	1.29 (0.03)
Busca	MAE	1.10 (0.06)	0.89 (0.08)	0.93 (0.08)
	RMSE	1.47 (0.05)	1.27 (0.07)	1.29 (0.07)
	STD	1.26 (0.01)	1.20 (0.03)	1.17 (0.02)
Compans-Cafarelli	MAE	1.02 (0.02)	0.99 (0.02)	0.99 (0.03)
	RMSE	1.42 (0.02)	1.38 (0.03)	1.37 (0.04)
	STD	1.41 (0.02)	1.35 (0.04)	1.37 (0.04)
Mean	MAE	0.95 ± 0.09	0.87 ± 0.07	0.89 ± 0.09
	RMSE	1.35 ± 0.08	1.25 ± 0.08	1.27 ± 0.09
	STD	1.31 ± 0.07	1.23 ± 0.07	1.23 ± 0.08

Table 2. Results of models trained on UWG simulations and **fine-tuned** with combinations of training set stations data and tested on test set stations. **MAE** (Mean Absolute Error), **RMSE** (Root Mean Square Error), **STD** (Standard Deviation) of errors are performance metrics of a certain trained model. The **STD** values **in brackets** represent the standard deviations across model training runs, while the **± std** values indicate the standard deviations across stations.

4. Discussion and Conclusion

To date, techniques for mapping urban air temperature at fine spatial resolutions generally fall into two main categories: data-driven approaches, which rely on interpolation and statistical inference, and physics-based approaches, which simulate temperature using physical principles (Taheri-Shahraiyni and Sodoudi, 2017).

We propose a robust framework combining physics-based and data-driven models for urban climate simulation. By leveraging simulated data for pre-training and real data for fine-tuning, our approach adapts well to heterogeneous urban environments. LCZ-based strategies offer a promising path for guiding data collection and optimizing neural model performance in data-sparse cities.

This reasearch demonstrates that combining physics-based microclimate simulations with deep learning yields robust, accurate urban temperature models. Key findings include a 30–35% reduction in RMSE compared to UWG and NUWG-Sim baselines, improved stability through multi-station fine-tuning, and a 33% speedup in simulation runtime. These results highlight the value of judiciously integrating simulated and observed data to capture the spectrum of urban microclimates.

The evaluation of NUWG-City across Toulouse weather stations reveals that pretraining on UWG simulations followed by targeted fine-tuning with real data delivers the most reliable air temperature estimates. Models trained exclusively on UWG outputs (NUWG-Sim) exhibited the largest errors, especially in heterogeneous microclimates such as dense, vegetated areas, underscoring UWG's limited sensitivity to fine-scale urban features when used without calibration. Fine-tuning NUWG-Sim on data from a single, well-chosen station significantly reduced error—achieving RMSEs near 1.3°C, but performance varied widely depending on the station's representativeness. By contrast, multi-station fine-tuning (combining two extreme LCZs and one average LCZ) not only matched or exceeded the best single-station results but also halved the variability in error across runs and sites. In quantitative terms, the best NUWG-City configuration outperformed NUWG-Sim by approximately 35% and raw UWG by 30%, while maintaining a computational speed roughly 33% faster than UWG itself.

Despite these advances, several limitations temper the approach's broader applicability. First, UWG was used "out of the box," without tuning its physical parameters or code, which constrained the fidelity of its synthetic data. Second, our fine-tuning leveraged only 12 of the 39 available stations in Toulouse metropolis, limiting the diversity of urban contexts represented in training. Third, NUWG-City's reliance on both rural meteorological data and radiative flux measurements restricts its use in locations lacking such inputs, and the model omits the influence of water bodies on local temperature dynamics. Finally, although pretraining on UWG embeds an implicit physical framework, NUWG does not enforce explicit physical laws—making expert post-hoc analysis essential to validate its outputs. Looking forward, several avenues can further strengthen NUWG-City. Enhancing the training database with more advanced or calibrated physical models could improve baseline fidelity. Expanding validation to longer time periods and additional years would test temporal robustness. Transferring the model to other cities, via direct retraining, domain adaptation, or knowledge distillation, could establish its generality across climatic and urban forms. Finally, integrating remote-sensing imagery (from surface parameter maps to high-resolution Sentinel-2 and future thermal missions) promises to automate feature extraction, enrich spatial context, and potentially enable end-to-end temperature mapping directly from satellite data. Such developments would broaden NUWG-City's applicability as a decision-support tool for urban planners confronting the challenges of heat mitigation and sustainable city design.

Acknowledgments

This work contributes to a broader doctoral research on urban climate modeling in Toulouse. We gratefully acknowledge Toulouse Me´tropole for providing access to weather station data, and thank Kermap and the ANRT for their generous financial support of this work.

References

Bueno, B., Norford, L., Hidalgo, J., Pigeon, G., 2013. The urban weather generator. *J. Build. Perform. Simul.*, 6, 269–281.

Buttsta¨dt, M., Sachsen, T., Ketzler, G., Merbitz, H., Schneider, C., 2011. A new approach for highly resolved air temperature measurements in urban areas. *Atmospheric Measurement Techniques Discussions*, 4(1), 1001–1019.

Gao, Y., Zhao, J., Han, L., 2022. Exploring the spatial heterogeneity of urban heat island effect and its relationship to block morphology with the geographically weighted regression model. *Sustainable Cities and Society*, 76, 103431.

Hamdi, H., Roupioz, L., Corpetti, T., Briottet, X., 2024. Evaluation of the Urban Weather Generator on the City of Toulouse (France). *Applied Sciences*, 14(1).

Ja¨nicke, B., Milosˇevic´, D., Manavvi, S., 2021. Review of User-Friendly Models to Improve the Urban Micro-Climate. *Atmosphere*, 12(10).

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2016. Fully convolutional neural networks for remote sensing image classification. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 5071-5074.

Pelletier, C., Webb, G. I., Petitjean, F., 2019. Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sensing*, 11(5).

Snell, S. E., Gopal, S., Kaufmann, R. K., 2000. Spatial Interpolation of Surface Air Temperatures Using Artificial Neural Networks: Evaluating Their Use for Downscaling GCMs. *Journal of Climate*, 13(5), 886 - 895.

Sobrino, J. A., Oltra-Carrio´, R., So`ria, G., Jime´nez-Mun˜oz, J. C., Franch, B., Hidalgo, V., Mattar, C., Julien, Y., Cuenca, J., Romaguera, M., Go´mez, J. A., de Miguel, E., Bianchi, R., Paganini, M., 2013. Evaluation of the surface urban heat island effect in the city of Madrid by thermal remote sensing. *International Journal of Remote Sensing*, 34(9-10), 3177–3192.

Stewart, I., Oke, T., 2012. Local Climate Zones for Urban Temperature Studies. *Bull. Am. Meteorol. Soc.*, 93, 1879–1900.

Taheri-Shahraiyni, H., Sodoudi, S., 2017. High-resolution air temperature mapping in urban areas: A review on different modelling techniques. *Thermal Science*, 21(6 Part A), 2267–2286.

Trenberth, K. E., 2018. Climate change caused by human activities is happening and it already has major consequences. *Journal of Energy & Natural Resources Law*, 36(4), 463–481.

Zachos, J., MO, P., Sloan, L., Thomas, E., Billups, K., 2001. Trends, Rhythms, and Aberrations in Global Climate 65 Ma to Present. *Science (New York, N.Y.)*, 292, 686-93.

Zhang, J., Zhu, W., Wang, L., Jiang, N., 2012. Evaluation of similarity measure methods for hyperspectral remote sensing data. *2012 IEEE International Geoscience and Remote Sensing Symposium*, 4138–4141.