Empowering urban climate resilience and adaptation: Crowdsourcing weather citizen stations-enhanced temperature prediction

Daniel Castro Medina a,*, MCarmen Guerrero Delgado b, José Sánchez Ramos b, Teresa Palomo Amores b, Laura Romero Rodríguez a, Servando Álvarez Domínguez b

a Grupo Termotecnia, Escuela Superior de Ingeniería, University of Cadiz, Spain
b Grupo Termotecnia, Escuela Técnica Superior de Ingeniería, University of Seville, Spain

A R T I C L E   I N F O
Keywords:
Urban climate
Recovering street life
Adaptation climate change
Forecasting
Social tool

A B S T R A C T
The growing impact of climate change, including extreme weather events, represents a significant challenge for humanity. With most of the world’s population living in urban areas, the urban heat island effect and anthropogenic heat contribute to elevated city temperatures. This increase in urban warming threatens human health and demands a deeper understanding of thermal distribution in urban environments. Collecting accessible and widespread temperature data in urban areas is essential to address this challenge. This study aims to develop a methodology for anticipating temperature distribution in urban environments, leveraging Citizen Weather Stations (CWS) as valuable crowdsourcing data sources. The ultimate goal is to create a predictive model that estimates urban temperatures based on government meteorological station forecasts, improving urban planning, regulating temperature-based routes, preventing health issues in vulnerable populations, and enhancing urban livability. The methodology is divided into three fundamental stages: data acquisition through CWS with citizen collaboration, the development and evaluation of optimal forecast models based on government weather stations (SWS) data, and its exploitation in terms of utility and applicability. This methodology encompasses data collection and filtering to ensure its usefulness and implement reliable models. The resulting tool facilitates informed decision-making and precise seasonal event planning in urban environments, effectively addressing the challenges of climate extrapolation and contributing to more effective adaptation and mitigation strategies in climate change and heatwaves. The results obtained probe the feasibility of using CWS to predict temperatures in urban environments, which has been demonstrated accurately. This is a significant achievement, as CWS has proven to be a reliable source of climate data for this context. Also, the filtering process described and applied to the case study has proven effective, discarding approximately 34.87% of the data. This is achieved by detecting and eliminating anomalies, considering station availability, and adhering to specific quality criteria. Finally, the developed prediction model has demonstrated its ability to optimally estimate urban temperatures, utilizing climate prediction data provided by government weather stations (SWS). The model performance indicators support this claim. For the linear regression model, a Mean Squared Error (MSE) of 2.177 and an R-squared ($R^2$) of 0.960 are obtained, while for the neural network, an MSE of 1.284 and an $R^2$ of 0.976 are achieved.

1. Introduction

Climate change and extreme weather events challenge humanity (Santamouris, 2020). The consequences of climate change are becoming increasingly evident worldwide. Through social media, the World Meteorological Organization (WMO) warns of the health risks, as well as risks to energy, water supply, and agriculture, posed by the heatwave that will affect warm regions like Southern Europe. The concentration of people, assets, and economic activity makes cities particularly vulnerable and, therefore, a priority for assessing the impact of climate change.

According to projections by the United Nations (World Urbanization Prospects - Population Division - United Nations, 2023), it is estimated that 57% of the world’s population will live in urban areas by 2050. In Spain, depopulation of rural areas has reached such a level that 81% of the population resides in urban areas. This continuous growth of urban areas has various climatic effects (Wu et al., 2023a; Kalnay & Cai, 2003).

https://doi.org/10.1016/j.scs.2024.105208
Received 13 October 2023; Received in revised form 11 January 2024; Accepted 12 January 2024
Available online 14 January 2024

© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).
D. Castro Medina et al.

One of the most significant is the phenomenon of the urban heat island (UHI), which has been extensively discussed in the literature (Arnfeld, 2003; Oke, 1973; Yang et al., 2016). Urban heat islands are urban areas that experience higher temperatures than the surrounding areas due to buildings, roads, and other infrastructure absorbing and emitting more solar heat than natural landscapes like forests and bodies of water (Stewart & Mills, 2021). The temperature increase caused by the Urban Heat Island (UHI) phenomenon is influenced by factors such as building density, building height, sky view factor (Chen et al., 2023), and urban land cover (Kim and Brown, 2021a). These urban characteristics contribute to the heterogeneity of the urban environment, resulting in temperature variations within the same city (Song et al., 2020; Yang & Li, 2015; Guo et al., 2016; Wang et al., 2017). Zekar et al. explain that urban form accounts for approximately two-thirds of city temperature variations (Zekar et al., 2023). Xue et al. (2020) conclude that ventilation rate, building height, and the ratio of building-occupied area to total urban area are crucial in the daily urban air temperature cycle. Levermore et al. (Levermore et al., 2018) point out that increasing urbanization and the loss of green spaces are the leading causes of the intensification of the heat island effect. They predict an average increase of 2.4 °C in urban air temperature by the end of the century, in addition to the increase projected due to climate change.

Anthropogenic factors influence the phenomenon of the heat island. Anthropogenic heat results from human activities and energy consumption in urban areas (Wu et al., 2023b). One of the most prominent anthropogenic factors is vehicular traffic. The results obtained by Husni et al. (Husni et al., 2022) show that the average temperature in urban areas is higher than in suburban areas, and severe traffic jams lead to a significant temperature increase, up to 7 degrees Celsius on sunny days.

On the other hand, extreme weather events such as heatwaves are becoming more frequent and, in the future, are expected to be more intense and persistent with the increase in global temperatures (Perkins et al., 2012). Health Risks on the Rise as Heatwave Intensifies across Europe: WMO | UN News, 2023, Hu et al. (2023) state in their study that exposure to heat during a heatwave was approximately 5 °C higher in urban environments than in rural areas and that this effect is magnified in hot cities. Kiaris et al. (2023) note in their review that heatwaves caused by climate change have resulted in the deaths of thousands of people in recent years. According to a recent United Nations report, over 60,000 people died in Europe during the summer of 2022 due to extreme heat (Health Risks on the Rise as Heatwave Intensifies across Europe: WMO | UN News, 2023). As a result, authorities advise their citizens to stay indoors due to the uninhabitable conditions.

The rising temperatures due to the urban heat island effect, anthropogenic heat, and extreme weather events like heatwaves also present a significant public health challenge. The World Health Organization estimates that the warming and precipitation trends resulting from anthropogenic climate change over the past 30 years have already claimed more than 150,000 lives annually (Patz et al., 2005). Authors such as Ho et al. (2023) have investigated the relationship between mortality and temperature in various UHI scenarios, concluding that the risk of temperature-related mortality is most pronounced in individuals aged 75 or older, and with more intense heat island effects, mortality rates double. Cleland et al. (2023) emphasize that older adults and populations with chronic conditions are the most vulnerable to heat, with the UHI effects increasing the risk of cardiovascular diseases. Taylor et al. (2015) suggest that housing construction characteristics contribute more to variations in temperature exposure than the UHI itself, thereby increasing mortality risks.

Considering the perspective above, which underscores the increasing temperatures in urban environments due to various factors, and considering the rapid population growth, understanding thermal distribution in urban areas becomes paramount due to its relevance in health impacts associated with extreme temperatures.

Numerous studies have been conducted to investigate thermal distribution in urban environments. In their research, Ho et al. (2014) graph with the inherent difficulty of mapping temperatures in urban areas due to significant temperature gradients. To address this challenge, they propose creating regression models that employ Landsat and elevation data to predict daily maximum air temperatures. On the other hand, Tsin et al. (2016) confront the challenge of monitoring temperatures in the urban environment at a micro-scale using mobile devices to better understand temperature distribution in urban settings. (Bechtel et al., 2017) have proposed and evaluated a robust empirical modelling approach. This method utilizes Land Surface Temperature (LST) evolution at predefined time intervals to estimate air temperatures in rural and urban areas, relying on satellite data. Hrisko et al. (2020) present a model constructed by adjusting the discrepancy between actual ground-level air temperature data and satellite-based Land Surface Temperature, employing a Gaussian function. Lastly, Romero Rodríguez et al. (2023) demonstrate in their work the process of monitoring air temperature in the urban environment using a low-cost Arduino-based mobile device.

In summary, obtaining temperature data in urban environments presents challenges due to the need for more versatility and ease in acquiring this data. This is due to various reasons, including the limited availability of conventional weather stations in urban areas and the complexity of implementing traditional monitoring systems in densely populated environments. As a result, collecting temperature data in urban areas becomes a challenging and costly task. In this context, the present study proposes using crowdsourcing stations like Netatmo as an accessible data source for a broader audience (Meier et al., 2017a). This approach aims to overcome traditional barriers in temperature data acquisition by harnessing community collaboration, opening the door to more efficient and widespread data collection in urban environments. This can be of great value for creating a temperature prediction model in urban environments and enabling informed decision-making in urban planning.

The imperative to develop accurate temperature forecasting models in urban environments has led to the application of machine learning techniques (Wang et al., 2023), mainly using neural networks (Tran et al., 2021). For example, Li et al. (2019) employed a stacked Long Short-Term Memory (LSTM) neural network to process and predict time series temperature data. Jallal et al. (2019) based their research on an algorithmic approach that utilizes multi-layer perceptron (MLP) neural networks, including a lagged exogenous input sequence representing global solar radiation (GSR) data. Numerous studies have underscored the effectiveness of feed-forward artificial neural networks (ANN) ([PDF] Temperature Forecasting Using Artificial Neural Networks (ANN), 2023; Abhishek et al., 2012; Chattopadhyay et al., 2011) as valuable models for predicting temperature based on prior temperature data.

The current study proposes extrapolating data from state weather stations (SWS) for application to corresponding citizen weather stations (CWS). Implementing a Multi-Layer Perceptron (MLP) neural network is
suggested to achieve this objective. This network inputs the temperature data from the SWS station from the current time instant (t) to n previous time instants. As a result, the corresponding value at instant t for the CWS station in question is obtained. The process involves training the neural network using historical data and forecast data from the SWS to predict the corresponding values at the CWS.

The complexity of understanding the climate in urban environments arises from several challenges that impede precise data acquisition. Firstly, it is essential to note that there are two primary sources of reliable climate information in Spain: the Agrometeorological Information System for Irrigation (SIAR) (https://servicio.mapa.gob.es/websiar/, 2020) and the State Meteorological Agency (AEMET) (Agencia Estatal de Meteorología - AEMET, 2023).

SIAR is a government system that collects and provides meteorological information primarily intended for the agricultural sector. Its objective is to assist farmers and water managers in making informed decisions related to irrigation and other agricultural practices. SIAR has an extensive network of strategically located weather stations in agricultural areas, ensuring the availability of accurate data in rural zones. On the other hand, AEMET is the government agency responsible for meteorology in Spain. Its primary function is to provide accurate and timely meteorological information and forecasts at the national level. AEMET also operates a network of weather stations in various places, such as airports and mostly non-urban areas, to comprehensively capture and monitor weather conditions.

In conclusion, these government stations are often distant from the effects of the urban heat island. As a result, climatic measurements obtained at these government stations only sometimes accurately reflect local climate conditions in urban environments.

This research aims to develop a methodology to anticipate temperature distributions in urban environments using Citizen Weather Stations (CWS) (Meier et al., 2017b). CWS, which are collaborative weather stations located in urban areas and exposed to various factors such as urban heat islands, heatwaves, and anthropogenic heat, serve as valuable sources of specific climatic data for these settings. The main challenge of this research lies in the extrapolation of temperature data collected by government weather stations, typically situated in rural areas, to the urban context. The research gap lies in the absence of a practical approach to extrapolate climate data to urban areas reliably and subsequently apply it as input parameters in thermal comfort models for establishing temperature-regulated routes, preventing health issues in vulnerable individuals, and enhancing the utility and livability of urban spaces. The goal is to address this gap by creating a predictive model that estimates urban temperatures based on predictions made by government stations.

The developed methodology focuses on data collection and filtering from Citizen Weather Stations (CWS) using a crowdsourcing approach and creating an optimal fitting model that allows for obtaining CWS temperature data based on the data provided by government weather stations (SWS). Two types of models are generated, one based on multiple linear regression and another based on neural networks. The model that exhibits better performance will be selected for future applications.

The expected outcomes of this research encompass several aspects:

1. Demonstrate the feasibility of using CWS to predict urban temperatures accurately.
2. Anticipate that the developed prediction model will be capable of estimating temperatures in urban areas optimally using climate prediction data provided by SWS.
3. The proposed methodology can be applied in other cities and urban environments.
4. These findings will underpin the utility of CWS as a resource to address challenges associated with extrapolating climate data in urban environments. Moreover, they will drive the development of more effective adaptation and mitigation strategies in urban areas, addressing issues related to urban design and planning, social aspects, impacts, and resilience of cities, as well as health monitoring and improvement.

2. Methodology

The design methodology developed (see Fig. 1) is applicable to any region with reliable climate data and the utilization of local data sources, such as Netatmo stations distributed within the urban context.

The methodological approach is broken down into three fundamental blocks: data acquisition, the development of the optimal forecasting model, and, most importantly, its application in terms of utility and practicality. The strength of this methodology lies in its ability to meticulously address every detail and consideration, leaving nothing to chance. Each critical aspect has been carefully implemented in an open-source tool and has undergone rigorous testing and validation across many case studies. This ensures not only a robust and reliable method but also a valuable tool for authorities and society at large. This tool enables informed decision-making and effective planning of seasonal events while addressing the challenges of extrapolating climate data in urban contexts. Furthermore, it contributes to developing more effective adaptation and mitigation strategies within the climate change framework, thus addressing an urgent issue today. The following sections delve into each aspect of the developed methodology in detail.

2.1. Data collection: City Weather Station data and quality control of temperature data

Citizen Weather Stations (CWS) play a fundamental role as they enable the collection of real-time and historical data from multiple points within the city (Muller et al., 2015). These stations provide more detailed information about temperature variations in different locations, facilitating an understanding of the urban environment’s effects on temperature patterns. This information makes it possible to design appropriate mitigation strategies to address adverse thermal impacts on public health.

Various sources of climatic information are available from personal weather stations like Netatmo (2023). These internet-connected stations are distributed throughout urban environments and gather real-time climatic data. For the current study, the use of Netatmo is chosen for the following reasons:

- The availability of an API that simplifies the acquisition of climatic data collected by these stations.
- It offers extensive coverage, making it possible to find numerous stations within the urban environment of a specific city.
- The sensors offering high precision (±0.3 °C).

To acquire temperature data from the CWS, a Python script has been developed utilizing the Netatmo API. When running the script, users are prompted to input the coordinates of a rectangle that defines the area of interest. This enables the retrieval of temperature data from all weather stations in that region. The script accesses the Netatmo API and collects temperature data from the stations within the selected interval. These data are stored in individual CSV files for each station, facilitating their handling and subsequent analysis. Furthermore, a master CSV file contains general information about the weather stations, such as their unique identifiers, locations, and other relevant details. This file provides an overview of the stations in the area of interest (Fig. 2).

Due to the variability in the status and location of Netatmo stations, additional validations are necessary to ensure data quality. Therefore, it is essential to implement a quality control process in any crowdsourcing data-based approach to prevent incorrect measurements. After a comprehensive analysis of the available literature, the following three quality controls focused on data obtained from weather stations are presented (Table 1):

Below, a data filtering/quality control methodology is presented at
various levels based on what was previously discussed to obtain accurate and consistent climate data for subsequent analysis:

- Level 0 data refers to the raw data set expected to be collected. This data set is determined by the number of stations used and the data sampling frequency and study interval.
- At Level 1, a filter for matching station locations by latitude and longitude is implemented to eliminate duplicate data.
- At Level 2, actual data from the selected climate stations is obtained. However, it is common for data availability issues to arise due to Wi-Fi connection errors, battery depletion, or user device shutdown. These problems can result in gaps in the collected information.
Table 1
Literature review of data filtering.

<table>
<thead>
<tr>
<th>Author</th>
<th>Levels of air temperature data quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Beele et al., 2022)</td>
<td></td>
</tr>
<tr>
<td>Naples et al. (Napoly et al., 2018)</td>
<td>M1: Removal of Misconfigured Stations Using Metadata. M2: Data Adjustment Based on Elevation and Detection and Concealment of Suspicious Data. M3: Removal of All Data from a Station for a Month if Over 20 % of the Station’s Data is Marked as Suspicious. M4: Comparison of Data from Each Station with the Mean Value of All Stations and Removal of Data from Stations with Low Correlation. O1: Interpolation of Missing Values for a Single Time Step Using the Mean of the Two Nearest Values from the Same Station. O2: Removal of Values if They Belong to a Station with Less than 80 % of Data Available Per Day. O3: Removal of Values if They Belong to a Station with Less than 80 % of Data Available Per Month.</td>
</tr>
</tbody>
</table>

2.2. Forecasting models

This section presents the proposed prediction models, namely the multiple linear regression and neural network models. For each Citizen Weather Station (CWS), creating at least one or both models is required to select the most suitable one. In the context of this study, the decision has been made to develop both models to facilitate a comprehensive comparison of the results. Data collection spanned a complete year (year 2022) for each meteorological station, with this period divided into two halves, allocating one for model training and the other for validation. This process is conducted after the corresponding data filtering, ensuring that training and execution occur with reliable data. When a CWS exhibits availability below 80 %, a more extensive data period is extracted to ensure appropriate model training and validation.

Both models demonstrate the ability to forecast temperature; their choice will depend on various factors. Linear regression and neural networks represent contrasting approaches in mathematical modelling for variable prediction. Linear regression, by adopting a linear equation, stands out for its simplicity and robustness, making it a less complex and more computationally efficient option. The direct interpretation of linear regression is facilitated by coefficients, which indicate the relative contribution of each variable to the model. In contrast, by incorporating a more complex model with layers and activation functions, neural networks can learn non-linear patterns and adapt to more intricate relationships in the data. While this complexity gives them the flexibility to handle situations with more significant uncertainty, resulting in a better fit, it also entails higher computational requirements and additional complexity in understanding the contributions of variables to the model. In conclusion, both alternatives, which are the most recurrent in the literature, are presented and compared to ensure the replicability of the methodology.

2.2.1. Multiple linear regression model

The proposed multiple linear regression model (see Fig. 3) is a way to model urban temperature corresponding to the CWS (City Weather Station) as a function of non-urban temperature corresponding to the State Weather Station over a specific period and its past values.

In this formulation, it is considered that the current urban temperature at time \( t (T_{CWS}(t)) \) is equal to the sum of the temperature corresponding to the State Weather Station at time \( t (T_{SWS}(t)) \) and the terms corresponding to the past values of this temperature. In other words, it has been considered that the variable can be characterized as a time series (Box et al., 2013). Eq. (2) shows the proposed formulation:

\[
T_{CWS}(t) = T_{SWS}(t) + \beta_1T_{SWS}(t-1) + \beta_2T_{SWS}(t-2) + \ldots + \beta_nT_{SWS}(t-n)
\]

Where:

- \( T_{CWS}(t) \) is the temperature of the City Weather Station at time \( t \).
- \( T_{SWS}(t) \) is the temperature of the State Weather Station at time \( t \).
- \( \beta_1, \beta_2, \ldots, \beta_n \) are the coefficients representing the weight or contribution of past values of the State Weather Station temperature in predicting the City Weather Station temperature.
- \( t - 1, t - 2, \ldots, t - n \) correspond to past time periods used in the model.

The value of \( n \) (previous time periods) in the proposed model will depend on the quality of the fit. A condition will be established to determine how many previous time periods should be considered in the model to achieve an accurate prediction of urban temperature based on non-urban temperature. The entire procedure for choosing the number of coefficients, as well as identifying them, has been automated and is part of the resulting open-source code.

The condition used to determine the number of previous time periods will be based on the Mean Squared Error (MSE) between the model with
Suppose the difference in MSE between these two models is less than a previously established threshold, typically 1%. In that case, the execution will stop, and it will be considered that the best configuration for the number of previous time periods has been achieved. Fig. 4.

This criterion is based on the idea that adding more previous time periods to the model can increase its complexity and fitting capacity. However, there is a point where adding more previous time periods does not provide a significant improvement in prediction accuracy. Setting a threshold on the MSE difference aims to avoid overfitting and find the appropriate balance between accuracy and model complexity.

### 2.2.2. Neural network model

The proposed model adopts a Multilayer Perceptron (MLP) neural network configuration with multiple dense layers to predict the City Weather Station (CWS) temperature based on the State Weather Station temperature. The architecture consists of three main layers: an input layer, two hidden layers, and an output layer. Fig. 5 illustrates a flow diagram of this model.

In the input phase, the neural network receives a sequence of temperature values from a government weather station, representing the non-urban environment. This sequence spans from the current time (t) to a previous time (t-n), where ‘n’ denotes the length of the input sequence used for making predictions. The resulting value is derived from the multiple linear regression model. The neural network model configuration involves a series of iterations in the training process of the MLP model. The model is trained multiple times for each urban climate station (CWS), incorporating variations in the neural network architecture. Different combinations of hidden layers and varying numbers of neurons in these layers and activation functions are explored. Each model is trained using training datasets representing 50% of the available data. During the training process, internal model parameters are adjusted over multiple epochs and batches of samples. Evaluation metrics include Mean Squared Error and the coefficient of determination ($R^2$). Depending on whether the metrics reach satisfactory levels, the model is deemed suitable for application, or optimization strategies can be identified for fine-tuning and improvement. The procedure applied to all urban climate stations (CWS) in the study reveals that the optimal configuration consists of two hidden layers using the Rectified Linear Unit (ReLU) activation function, with the configurable parameter for each CWS being the number of neurons in each hidden layer.

### 2.3. Validation and exploitation: case of study

In this study, the proposed methodology from Section 2 is implemented in the urban center of the city of Sevilla. Temperature data for the entire year 2022 is collected at hourly intervals. The data collection process utilizes the Python-developed Netatmo API call script (see Section 2.1). The specific data configuration used is shown in Fig. 6 (left), which corresponds to an area of approximately 35 km$^2$ in the urban environment of Sevilla (Fig. 6 (right)).

Seville was chosen for this study in collaboration with the city’s urban authority due to the issues of urban space usage during the
summer. The State Meteorological Agency (AEMET), based on data from the Sevilla airport station (non-urban environment), reports that during July and August in 2023, there have been more than 184 h with temperatures exceeding 35 degrees Celsius, with 60 of them surpassing 40 °C. However, the effects described in Section 1 significantly exceed this number in the urban environment. This situation emphasizes the urgent need for a comprehensive approach to address the challenges posed by extreme summer temperatures. It is imperative to have a methodology that provides urban authorities with advanced climate forecasts and information on the suitability of specific hours of the day for outdoor activities. The urban community, which includes children and the elderly, should be able to safely enjoy the city’s streets and public spaces during the summer season. Therefore, a robust data-driven methodology has been developed to objectively guide urban authorities in planning and organizing activities that help revitalize urban life during these challenging weather conditions. This approach goes beyond traditional weather alerts. It aims to empower the urban population with accurate and detailed information to make informed decisions and ensure their well-being during extreme heat.

Data has been collected from 36 Citizen Weather Stations (CWS) for the case study. These stations are distributed over an altitude range, ranging from a minimum of 8 meters to a maximum of 24 meters, with an average altitude of 15 meters. Thanks to the use of the folium(Folium — Folium 0.1.Dev1+g57e8eae Documentation, 2023) library in Python and the precise information provided by these stations, it is possible to create a geographical map of their locations. Fig. 7 displays the result of this mapping, where the exact geographic location of each station is visualized.

In this study, two important validation stages have been conducted to assess the accuracy of the Citizen Weather Stations (CWS) measurements and the utility of the proposed data filtering stages in Section 2.1. The first of these validations occurred in Granada, Spain, in an area characterized by a high density of CWS. In this context, the research team installed their weather station to verify the similarity of the measurements obtained with those from other stations in the same area. The goal was to confirm that stations nearby exhibit temperatures and other climatic parameters comparable to those recorded by the newly installed station. This validation exercise laid the groundwork for applying the
data filtering methodology developed in the study and examining its effectiveness in improving measurement quality.

The second validation stage was conducted in the city of Sevilla, where a weather station was placed on the roof of a school. In this case, the focus was on spatial interpolation of climate data. Data from Netatmo stations located near the school station were collected for this purpose. Using spatial interpolation techniques, climate values were estimated at various locations in the area. The primary goal of this validation was to compare these interpolated values with the actual measurements obtained by the weather station installed at the school.
This comparison sheds light on the ability of interpolation to provide accurate and reliable data in areas where direct weather stations are not present.

Both validation stages are essential to ensure the quality and accuracy of the data collected by the weather stations and to confirm the effectiveness of the methodologies used in the study. Validation not only supports the reliability of climate data but also establishes a solid foundation for future analyses and forecasts based on this data. These validations are a fundamental step in ensuring that the measurements taken by CWS are reliable and helpful in understanding and predicting weather conditions in the study region.

2.3.1. First check

The first validation stage takes place in February 2023 in Granada, Spain. In this geographical context, 61 stations belonging to the Netatmo network are identified. Notably, six stations are located near the personal weather station, as illustrated in Fig. 8.

Applying the filtering process to the neighboring stations deletes 1/3 of the stations due to low data availability and consistency. After completing the filtering process, four stations that remain and exhibit suitable levels of similarity and consistency in their records are selected. Subsequently, a comparison is made between the data collected by these chosen stations and the data from the personal weather station (See Fig. 9).

Based on Fig. 9, it can be confirmed that the filtering process is highly beneficial, as it eliminates all those CWS with out-of-range values, low availability, or those located indoors. It is also possible to assert that the data is of good quality because it closely matches the values obtained by the installed personal weather station.

2.3.2. Second check

In the second validation phase, measurements were taken using a personal weather station located at the Arias Montano School in Sevilla, as indicated in Fig. 10, highlighted with a red circle. In the figure, it can be observed that there are no nearby CWS (blue icons) near the personal weather station’s location. Therefore, if it is necessary to know the temperatures at this specific location, spatial interpolation is required (Fig. 11).

Data from multiple nearby stations in the Netatmo station network are collected to carry out this spatial interpolation process. These data are used to estimate and generate a complete set of temperature data at the location of the personal weather station. Subsequently, a detailed comparison was conducted between these interpolated data and the data recorded by the personal weather station (See Fig. 10).

Fig. 10 presents the data acquired by the personal weather station and the temperature values at that location obtained through spatial interpolation (CWS interpolate). It is evident that thanks to the data filtering process, the information collected from Netatmo CWS stations is valuable, as the spatial extrapolation of the data shows a high level of accuracy compared to the data measured at that specific location. This validation reinforces the reliability of the measurements taken by CWS stations as long as the data filtering process is carried out effectively, as it has been done in this particular case. The ability to use data from nearby stations to estimate values at specific locations underscores the usefulness of the Netatmo CWS network.

3. Results

3.1. Application and validation of the data filtering methodology

As mentioned in Section 2.1, performing a data filtering process for the data obtained from the various CWSs is necessary to obtain precise and consistent data for subsequent use in the prediction model training. Based on 6 levels, the filtering methodology described is applied to the case study.

At level 0, 315,360 temperature measurements were expected for these 36 CWSs, considering the temporal scale and time interval used.

At level 1, the location of these stations is verified. If some CWSs have coincident latitude and longitude, one is discarded to avoid duplicates. In the study, only two stations were found with the exact location and altitude, so one was removed. Consequently, 35 CWSs were obtained, representing 306,600 raw temperature data.

At level 2, access is made to the data from each station. Out of the 35 CWSs, only 231,475 had data available, which represents a reduction of...
24.5% compared to the expected data. This loss of information is due to Wi-Fi connection problems or lack of battery in some of the stations. Fig. 12 (a) shows the distribution of the data obtained from each station throughout 2022, highlighting the presence of anomalies due to the wide dispersion of some CWSs.

Level 3 of the filtering process is applied to detect and remove these anomalies. At this level, the mean and standard deviation of the data for each hour are calculated, and a specific formula is applied to check if a data point is abnormal (greater or less than 3 standard deviations). Thanks to this filtering, a total of 4308 anomalous data points were eliminated. Fig. 12 (b) shows the distribution of the remaining data after applying level 3, demonstrating a reduction in the temperature range and removing previously detected anomalous values in Fig. 12 (a).

Levels 4 and 5 are related to data elimination due to the lack of availability of the CWSs, indicating abnormal or discontinuous operation. At level 4, days with CWS availability below 80% are removed. Applying this level, 5446 temperature data points were discarded, leaving 221,721. Fig. 12 (c) displays the data distribution after applying
this level, showing a reduction in noise in the first half of the year, corresponding to minimum temperatures.

Finally, level 5 is even more restrictive than level 4, requiring at least 80% of monthly data to be retained. Implementing this level results in a significant reduction in the volume of data, with a total of 205,401 records remaining. Fig. 12 (d) presents the distribution of all the data used for subsequent analysis. Once again, there is substantial data removal for minimum temperatures during the year’s first half due to non-compliance with established availability requirements.

In Fig. 13, a bar graph illustrates the progressive reduction of data
across the different filtering levels. Summarizing the process, starting from 100 % of the expected data at level 0, it can be observed that, following the filtering process, only 65.13 % of the original data is retained. This implies that approximately 34.87 % of the data has been discarded due to the detection and elimination of anomalies, lack of station availability, or non-compliance with specific quality criteria.

From the perspective of the data collected from each CWS, it is observed that out of a total of 35 initial CWSs, only 15 have complete data in at least 95 % of the cases. On the other hand, 7 out of the 35 CWSs have availability lower than 20 %. These CWSs have significant missing or incomplete data, which may limit their utility for further analysis. The remaining 13 CWSs, representing over a third of the total, show availability between 20 % and 95 %. These CWSs have an average availability of 64 %, indicating that they have some gaps or missing data but are mostly available.

3.2. Comparative analysis and validation of neural network and regression models

For this analysis, one of the urban temperature monitoring stations (CWS) with significant data availability (>98 %) has been selected. The reference temperature used is derived from data provided by the State weather station “SIAR” and corresponds to a location in a non-urban environment (La Rinconada, in Sevilla). Data collected during the first half of 2022 have been used to adjust the multiple linear regression and neural network-based models. Performance evaluation is carried out using data from the second half of 2022. This evaluation involves determining the Mean Squared Error (MSE) and the coefficient of determination ($R^2$).

Fig. 14 displays the temperatures of the non-urban SWS (State Weather Station) and the urban temperature (CWS). The average temperature values for the reference and urban temperatures were 19.30 °C and 21.46 °C, respectively. These results demonstrate the phenomena described in the introduction of this study.
Applying what was described in Section 2.2.1, the models consist of 29 previous time steps (n = 29), and the hidden layers contain 64 and 32 dense units, respectively, with the ReLU (Rectified Linear Unit) activation function. The table below presents the results of the neural network model and the regression model in terms of Mean Squared Error and the Coefficient of Determination (Table 2, Fig. 15):

Based on the results obtained, it can be observed that the Neural Network model has a lower Mean Squared Error (MSE) of 1.284, indicating a better fit to the data compared to the Linear Regression model, which had an MSE of 2.177. The Neural Network model also exhibits a higher Coefficient of Determination ($R^2$) of 0.976, implying that it can explain approximately 97.6 % of the data’s variability. On the other hand, the Linear Regression model achieved an $R^2$ of 0.960, indicating slightly lower but still high explanatory capability of 96 %. These results demonstrate the effectiveness and accuracy of the Neural Network model compared to the Linear Regression model in the prediction task. The lower MSE and higher $R^2$ values suggest that the Neural Network model may be more suitable for making accurate predictions on the interest data.

To visualize the improvement between the Linear Regression and Neural Network models in terms of urban temperature prediction, the forecasted values of urban temperature, as well as the measured values (CWS) and the reference temperature (Reference), are plotted for the week of August 21–28, 2022. Fig. 16 shows this for the Linear Regression model, while Fig. 17 presents it for the Neural Network model.

The Linear Regression model achieved a Mean Absolute Error (MAE) of 1.61 and a Mean Squared Error (MSE) of 2.99 for the specified week. In contrast, the Neural Network model significantly improves prediction accuracy. It obtained a MAE of 0.79, and the MSE was 0.88. The lower values of MAE and MSE indicate that the Neural Network model can make predictions closer to the actual measured values, making it a more reliable choice for predicting urban temperature.

4. Applicability

The applicability of what has been developed in this research extends to various contexts where urban climate management and outdoor activity planning are paramount. Below, it is detailed how the results and proposed methodologies can be effectively used in various situations:

1. Selection of Outdoor Event Locations: The findings of this study have particular relevance in planning sports and recreational events in urban settings during the summer months. Municipal authorities, event coordinators, and urban planners can leverage the climate prediction models derived from the methodology outlined to anticipate thermal conditions in the urban environment. Using forecasts provided by state weather stations (SWS), these stakeholders can differentiate between various locations and opt for those offering a more favourable climate for participants. As a result, a more comfortable and safe experience would be ensured for all attendees.

2. Activity Scheduling: The ability to predict specific weather conditions allows for precise scheduling of outdoor activities. Event organizing bodies can use temperature predictions to determine optimal activity timings, avoiding the day’s hottest hours and minimizing exposure to extreme heat. This contributes to the well-being of participants and reduces risks associated with high temperatures.

3. Urban Infrastructure Design: The results of this research can also be valuable for the planning and design of urban spaces. Architects and urban planners can use climate data and temperature predictions to inform the design of public areas, such as parks and squares, incorporating shading and vegetation strategies to mitigate the urban heat island effect and enhance the comfort of citizens.

4. Climate Change Adaptation Strategies: In climate change, where heatwaves are becoming more frequent and intense, the proposed methodologies can be essential for developing adaptation strategies. Municipal authorities can use temperature predictions to identify areas at higher risk of extreme thermal conditions and design specific interventions to protect vulnerable groups of people.

4.1. Example of application

The city council plans to organize an outdoor soccer marathon in the urban context of Seville, scheduled from August 21st to August 28th. In this situation, the choice of an optimal location plays a crucial role, given the urgent need to ensure the safety of participants, particularly in light of the high temperatures inherent to the summer season. Following this premise, the methodology developed during this research stands out as a valuable resource.

The city council is considering two possible sites to host the event: Zone A and Zone B (see Fig. 18). Both locations have an indoor futsal court, which protects them from direct solar radiation, with ambient temperature being the critical factor. This temperature must not exceed 32 °C to ensure the viability of the matches (Extreme Heat: When Outdoor Sports Become Risky Climate Central, 2023).

The marathon takes place over 8 days, during which a maximum of 120 h is allocated for the competition, with a schedule ranging from 9:00 AM to midnight. The competition consists of 6 groups, each with 4 teams competing against each other. The two best-placed teams from each group and the two third-best-placed teams advance to the final stages, including the round of 16, quarter-finals, semi-finals, and the grand championship final. In total, these stages encompass 51 matches, with each match lasting 1 hour.

Applying the described methodology allows for obtaining climate forecasts for each of the locations (CWS) within the urban environment of Seville, using the data provided by the climate forecasting station (SWS). In the presented case, Zone A and Zone B are close to existing CWS stations. When this is not the case, it is possible to interpolate the data geographically due to the density of CWS stations. Fig. 19 depicts the temperature forecast for Zone A and Zone B from August 21st to August 28th. The temperature threshold is shown in red and as discontinuous, indicating that all matches must be conducted below this threshold.

It can be observed that during the night, both in Zone A and Zone B, the air temperature is similar, while during the midday hours, significant differences exist (ΔT, Fig. 19 (green line)), with the temperature being higher in Zone B. Table 3 displays the maximum temperature difference per day and the difference between the daily average temperature per day.

In light of the collected data, if there were no temperature restrictions below 32 °C for the tournament’s viability, Zone A would be the chosen option to host the championship. However, due to the temperature limitation, it is imperative to determine if the tournament can be held within the stipulated 8-day period during which a total of 51 matches must be played. Given this situation, an analysis of the availability of feasible daily hours in which the championship can be executed within the established schedule is proposed.

In Zone A, the percentage of hours below the threshold is 52.08 %, in contrast to the lower 31.77 % in the case of Zone B. Extrapolating these data to the temporal context, it can be inferred that in Zone A, it is feasible to use the pitch to carry out the tournament for a total of 62.5 h. In contrast, in Zone B, this time is reduced to 38 h out of a minimum of 51 h required to ensure the marathon’s viability. Fig. 20 illustrates the daily availability for both Zone A and Zone B.

Table 2

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Squared Error (MSE)</th>
<th>R-squared ($R^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Linear Regression</td>
<td>2.177</td>
<td>0.960</td>
</tr>
<tr>
<td>Neural Network</td>
<td>1.284</td>
<td>0.976</td>
</tr>
</tbody>
</table>
Fig. 15. Linear regression (Left) neural network (Right): model accuracy.

Fig. 16. Urban temperature prediction - linear regression.

Fig. 17. Urban temperature prediction - neural networks.
The results of this study have completely transformed the event planning. Initially, a schedule was followed based on state recommendations regarding heatwave alerts and traditional climate parameters. However, the developed methodology reduces the number of hours with critical temperatures, ensuring optimal conditions for all activities. In the original planning, over 60% of the activities were scheduled with temperatures above the threshold. Thanks to our methodology, we have eliminated these adverse conditions, ensuring a safe and comfortable
environment for all participants.

5. Discussion

This study focuses on developing and applying a methodology to estimate air temperature distribution in urban environments, exhibiting significant differences from non-urban settings. It provides a solution to a social problem in the literature: the need to generate developments that allow the recovery of street life and the adaptation of life in cities under climate change conditions and urban heat island issues. Based on crowdsourcing, this methodology allows the results to be utilized in formulating specific mitigation and adaptation policies for the urban environment. These policies address aspects such as planning activities in the urban environment and managing exposure for individuals most susceptible to health risks.

To date, the limited availability of air temperature measurements in urban environments has hindered the ability to predict thermal distribution in such contexts [(Chapman et al., 2023)], posing a challenge in assessing the magnitude of the urban heat island (Romero Rodríguez et al., 2023b) and the associated risk for the population (Kim et al., 2022). In this research, using Citizen Weather Stations (CWS) and State Weather Stations (SWS) enables the acquisition of information regarding temperature distribution in the urban environment through the developed methodology. It facilitates prediction, addressing the inherent complexity of this situation. In contrast to other methodologies described in the literature (Bechtel et al., 2017; Romero Rodríguez et al., 2020; Yu et al., 2020), the present one adheres to a crowdsourcing and open-source approach, providing the opportunity for any individual or governmental entity to apply it according to their needs.

Instead of employing complex prediction models, the methodology relies on state-of-the-art temperature prediction. The implementation of a fast-processing predictive model, such as linear regressions, and a computationally high-capacity model, such as neural networks, are two alternatives that, in conjunction with temperature data provided by CWS, enable the extrapolation of reliable data from state stations to points within the urban environment. This facilitates understanding temperature distribution and prediction, allowing the anticipation of scenarios in the urban environment and aiding in planning against undesirable phenomena such as the urban heat island and heatwaves. Both models, developed from a dataset collected for one year, yield satisfactory results. The linear regression model exhibits a mean squared error in the analyzed case of 2.17 °C, while the high-capacity computational model based on neural networks shows a result of 1.28 °C. Additionally, the coefficient of determination presents a value of 0.96 for the LR-based model and 0.976 for the neural network-based model.

From the perspective of replicability, the methodology developed in this study to predict the temperature distribution in urban environments is readily applicable to any location on Earth, provided it has a substantial number of Citizen Weather Stations (CWS). Following the filtering process, many stations must persist and have a considerable data history to train the proposed prediction models. In their research, Venter et al. (Venter et al., 2020) address the necessary quantity of CWS to ensure adequate accuracy in temperature distribution in the urban environment, placing this value at 250 CWS or one station per km². This criterion is met in most European cities, as documented in numerous studies employing CWS (Chapman et al., 2017; Feichtinger et al., 2020; Napoly et al., 2018; Uteuov et al., 2019; Venter et al., 2020; Zumwald et al., 2021). On the other hand, in Section 4, we have discussed and demonstrated the applicability of this methodology, which proves to be

![Fig. 20.](image-url)

Table 3
Maximum temperature difference per day and average absolute temperature difference per day.

<table>
<thead>
<tr>
<th>Date</th>
<th>Maximum Temperature Difference per Day [°C]</th>
<th>Average Absolute Temperature Difference per Day [°C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2022-08-21</td>
<td>3.80</td>
<td>2.30</td>
</tr>
<tr>
<td>2022-08-22</td>
<td>2.74</td>
<td>1.64</td>
</tr>
<tr>
<td>2022-08-23</td>
<td>2.68</td>
<td>1.22</td>
</tr>
<tr>
<td>2022-08-24</td>
<td>2.80</td>
<td>1.21</td>
</tr>
<tr>
<td>2022-08-25</td>
<td>2.83</td>
<td>1.35</td>
</tr>
<tr>
<td>2022-08-26</td>
<td>3.08</td>
<td>1.82</td>
</tr>
<tr>
<td>2022-08-27</td>
<td>2.61</td>
<td>1.70</td>
</tr>
<tr>
<td>2022-08-28</td>
<td>3.98</td>
<td>1.93</td>
</tr>
</tbody>
</table>
highly valuable for local authorities in organizing events during the summer season.

This methodology has its limitations. In the context of the research conducted in this study, Sevilla demonstrates an approximate density of one Citizen Weather Station (CWS) per square kilometre. However, the previously unaddressed distribution of these stations constitutes a critical aspect to consider in determining the feasibility of ensuring a certain level of accuracy when estimating temperature distribution across the entire urban grid. Various parameters contribute to the variability in urban air temperature (Kim & Brown, 2021). Urban morphology and proximity to heat sinks, such as significant bodies of water or green spaces, emerge as particularly relevant factors (Kianmehr et al., 2023; Pena Acosta et al., 2023; Zhao et al., 2023). Therefore, for a thorough assessment of urban temperature distribution, it is necessary that urban morphology does not exhibit significant differences in form and use and, of course, that there are no significant bodies of water or green spaces. To evaluate the accuracy of temperature distribution within the urban environment using CWS and considering the factors mentioned, a geographical map of Sevilla has been created, applying a grid. Each grid cell represents one square kilometre of surface area and has been assigned a specific colour based on accuracy (Fig. 21):

- **Adequate Accuracy (Green):** Adequate precision is achieved through the presence of one or more stations within one km² area, with no interruptions in the urban grid. This enables precise mapping of temperature in the cell.
- **Moderate Accuracy (Yellow):** Characterized by uniformity in the urban grid, without discontinuities due to large bodies of water or green areas. However, it is situated in an area with a low density of stations, and the distance to the nearest station exceeds 1 km.
- **Inaccuracy (Red):** In this region, any potential data obtained through interpolation is unreliable due to interruptions in the urban grid, such as the presence of the Guadalquivir River crossing Sevilla and the low or nonexistent density of Citizen Weather Stations (CWS).

These limitations could be overcome through policies that support the implementation of Automatic Weather Stations (AWS) with the allocation of public funds. These stations, strategically distributed by local authorities in the urban environment and adhering to high-quality standards, enable a precise understanding of temperature distribution in the urban area, facilitating and ensuring proper applicability. The use of AWS also guarantees the intrinsic reliability of the data, obviating the need for the proposed filtering in CWS.

### 6. Conclusions

The research aims to demonstrate the feasibility of utilizing Citizen Weather Stations (CWS) to predict urban temperatures accurately. Additionally, it foresees that the developed prediction model will be capable of optimally estimating temperatures in urban areas using climate forecast data provided by governmental weather stations (SWS). The proposed methodology holds potential for application in other cities. Moreover, it enables the development of more effective adaptation and mitigation strategies in urban areas, addressing aspects related to urban design and planning, social factors, impacts, and resilience of cities, as well as health monitoring and improvement.

The results obtained have met the expectations in several aspects:

- The feasibility of using Citizen Weather Stations (CWS) to predict temperatures in urban environments has been demonstrated accurately. This represents a significant achievement, as CWSs have proven to be a reliable source of climate data for this context.
- The filtering process described and applied to the case study has proven effective by discarding approximately 34.87 % of the data. This is achieved by detecting and eliminating anomalies, considering station availability and meeting specific quality criteria. This step is essential to ensure that only reliable data are used in the interpolation and forecasting.
- The developed forecast model has demonstrated its ability to optimally estimate urban temperatures by taking advantage of climate forecast data provided by government weather stations (SWS). Model performance indicators support this claim. For the linear regression model, a Mean Squared Error (MSE) of 2.177 and an R-squared (R²) of 0.960 are obtained, while, for the neural network, an MSE of 1.284 and an R² of 0.976 are achieved.

This methodology has considerable applicability and replicability in planning outdoor events, the precise scheduling of activities, the design of urban infrastructures and adaptation to climate change in urban environments. In addition, it facilitates the scheduling of activities to avoid the hottest hours, optimizes the design of public areas and supports the creation of climate change adaptation strategies, all to improve the quality of life in urban environments. The example presented in the city of Sevilla illustrates how the methodology helps to choose optimal locations and plan outdoor events efficiently.

Future lines of research are oriented towards integrating and developing a practical application for citizens and authorities. This application, equipped with the developed urban weather prediction methodology, helps provide optimal routes, highlight shaded options, and provide information about outdoor sports and weather conditions. This approach can significantly improve citizens’ quality of life by facilitating the planning of outdoor activities and promoting safety and comfort at events. The application would undoubtedly become a valuable resource for addressing the challenges of climate change in urban environments, and its promising development points towards future research and highly relevant practical applications. For this reason, all developments made in this study have been published and are available in open source on GitHub, making them accessible to anyone. It allows these resources to be used and integrated into various platforms, websites, applications and tools designed for the benefit of citizens.

**CRediT authorship contribution statement**

**Daniel Castro Medina:** Writing – original draft, Investigation, Data curation. **MCarmen Guerrero Delgado:** Writing – review & editing, Supervision, Methodology, Investigation. **José Sánchez Ramos:**…

![Fig. 21. Accuracy in estimating the distribution of temperatures throughout the urban fabric.](image-url)
Validation, Methodology, Investigation, Formal analysis. Teresa Palomos Amores: Investigation, Formal analysis, Data curation. Laura Romero Rodríguez: Investigation. Servando Álvarez Domínguez: Validation, Methodology, Investigation, Conceptualization.

Declaration of competing interest
The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Servando Álvarez Domínguez reports financial support was provided by Spain Ministry of Science and Innovation.

Data availability
Data will be made available on request.

Acknowledgements
This study has been funded by the projects "CONSTANCY - Resilient urbanisation methodologies and natural conditioning using imaginative nature-based solutions and cultural heritage to recover the street life" (Grant Agreement PID2020-118972RB-100) and the project "NATUR-BEAM - Lighting the way to a greener future to restore urban habitability through nature-based solutions" (Grant Agreement TED2021-130416B-100) by the Spanish Ministry of Science and Innovation.

References