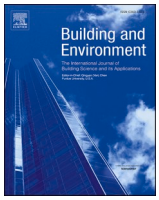
Building and Environment 271 (2025) 112593

Contents lists available at ScienceDirect

Building and Environment

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



Gap filling crowdsourced air temperature data in cities using data-driven approaches

Miao He [a], Zhiwen Luo [a,*], Xiaoxiong Xie [b], Peng Wang [c], Haichao Wang [c,d], Gabriela Zapata-Lancaster [a]

[a] Welsh School of Architecture, Cardiff University, United Kingdom
[b] School of Art, Design and Architecture, University of Plymouth, United Kingdom
[c] School of Infrastructure Engineering, Dalian University of Technology, China
[d] Department of Mathematics and Systems Analysis, School of Science, Aalto University, Finland

ARTICLE INFO

Keywords:
Machine learning model
Urban heat island
Urban weather observation
Crowdsourced weather data
Gap filling
Data imputation

ABSTRACT

Crowdsourced temperature data from citizen weather stations (CWS) in urban area offer valuable insights into intra-urban temperature distribution but are often challenged by a significant number of missing values. Existing gap-filling methods, typically effective for random gaps and low missing rates, are inadequate for the continuous gaps and high missing rates common in CWS recordings. This study introduces a novel data-driven approach to fill these gaps by modelling relationships between CWS data and official weather station (OWS) records during periods of data availability. We evaluate various feature sets and data-driven algorithms, including Multiple Linear Regression (MLR), Random Forest (RF), and Multilayer Perceptron (MLP), using a complete CWS temperature dataset from July 2018 in London. The MLP-based models, which include features such as preceding and subsequent air temperature along with past solar radiation, demonstrate superior performance across various missing data conditions. In the most challenging case, with a missing rate of 70–80% and continuous gaps, the MLP model achieves a Mean Absolute Error of 0.59 °C, a Root Mean Squared Error of 0.73 °C, and a coefficient of determination (R^2) of 0.94. The robustness of the MLP algorithm is further validated across multiple complete CWS datasets from different areas in London. This study offers effective strategies for handling common missing data conditions in CWS datasets and serves as a valuable reference for future machine learning applications in urban climatology.

1. Introduction

1.1. Background

With global warming, heatwaves have become more frequent, intense, and prolonged [28]. Extreme heat significantly impacts human health, wellbeing, and safety, making understanding heat stress a critical health concern, especially in urban areas where the urban heat island (UHI) effect exacerbates heat exposure compared to rural areas [17]. However, the diversity in urban land cover and urban morphology [41] leads to substantial temperature variations across urban area, with difference sometimes exceeding 5 °C [34]. To effectively understand and mitigate urban heat stress, it is crucial to study urban temperature distribution at a fine spatial resolution.

Urban temperature distribution can typically be obtained through three primary methods [30]: numerical modelling, remote sensing, and onsite measurements. Numerical models, while capable of fine spatial resolution, require significant computational resources, and their accuracy depends on the precision of inputs, some of which may be uncertain [32]. Remote sensing provides land surface temperature data, but these measurements do not fully represent air temperature [37]. Onsite measurements, on the other hand, offer the most accurate representation of air temperature. Onsite temperature data can be obtained from two key sources: official weather stations (OWS), typically established by the World Meteorological Organization, and citizen weather stations (CWS). The latter are installed by the public for personal or educational purposes and offer a crowdsourced alternative to traditional measurements. [26].

Traditionally, OWS have been used to record long-term time series of outdoor meteorological variables. However, their sparse

* Corresponding author.
E-mail address: LuoZ18@Cardiff.ac.uk (Z. Luo).

https://doi.org/10.1016/j.buildenv.2025.112593
Received 7 October 2024; Received in revised form 13 January 2025; Accepted 20 January 2025
Available online 20 January 2025

distribution—often located in rural areas or urban parks—limits their ability to accurately represent the diverse microclimates found in urban environments. In recent years, CWS have gained popularity due to their lower cost and ease of installation [33]. The dense presence of CWS in some cities has made their temperature data increasingly valuable for capturing urban air temperature distribution. For example, Fenner et al. [9] used both OWS and nearly 2000 CWS in Berlin to analyse how location choice and time of day influence the variability of urban heat island (UHI) intensity. Similarly, Benjamin et al. [4] used CWS temperature data to estimate the UHI effect and assess building heating/cooling loads in London. Brousse et al. [6] also employed CWS data to examine urban heat patterns and the impact of urban heat advection in southeast England and Greater London. These studies underscore the significant potential of CWS data for urban heat studies, offering a more granular view of temperature variability in cities.

The main challenge with CWS is the relatively low data quality compared to OWS data. CWS time series often contain gaps due to sensor failures, connection issues, or misuse of sensors [3,24]. To improve the reliability of CWS data, a statistically based quality control method has been developed by Fenner et al. [8]. Through statistical analysis, unreliable data can be removed, and gaps with a single missing value can be filled using linear interpolation. However, for time series with higher missing rates (e.g. more than 20% of the data), which typically indicates large gaps, linear interpolation is not valid for gap filling, and all values during that period are deleted in the quality control process [8,36]. This leads to a significant loss of CWS data and reduces the amount of available data. Furthermore, biased results are more likely to be inferred from such smaller samples, compromising the robustness of statistical analysis [19]. Additionally, major existing temperature forecasting models cannot proceed with gaps in training data [29]. Thus, the application of CWS data, such as forecasting local climate conditions, is further limited by these gaps. To improve availability and application of CWS data, effective data filling methods, especially for CWS data with high missing rates, are required.

1.2. Related work

Currently, no established method exists for filling missing data in CWS air temperature datasets with high missing rates. To address this, we review existing data-filling techniques to identify the most suitable method for CWS data imputation. Due to the limited research on outdoor air temperature imputation, we also consider gap-filling methods from other time series data, such as energy use, for a broader comparison.

Missing data conditions in previous research vary in both the missing rate and gap length. Both are primary factors influencing the effectiveness of gap-filling techniques. The missing rate refers to the percentage of data points missing from the time series. Data gaps are typically categorised by length into random (short-term) and continuous (long-term) gaps. Random gaps involve short periods without data, such as hourly or daily intervals, whereas continuous gaps refer to extended periods, such as weekly or monthly intervals in the time series [10].

1.2.1. Filling methods

Simple statistical models, such as interpolation, moving averages, local linear regressions, and K-nearest neighbours, are commonly used to fill random gaps by leveraging nearby points [7,16]. However, these models are less effective when significant changes occur within the gaps, and their performance declines as gap length increases due to insufficient data points near the gap's centre [36].

Models that capture temporal dependencies in time series can overcome some of the limitations of simple statistical models. By using all available observations, rather than just nearby data, these models are more effective at managing sudden changes within data gaps and perform better as gap lengths increase. For example, Sarafanov et al. [29] trained a temporal model using evolutionary algorithms, which outperformed various interpolation methods when gap lengths exceeded 30% of the sea surface height time series. Autoregression and recurrent neural networks (RNNs) are commonly used to capture time dependencies. Afrifa-Yamoah et al. [1] used an autoregressive model to fill 10% of missing data in a 12-month hourly outdoor air temperature series, effectively handling gaps of up to 30 consecutive hours with root mean squared error (RMSE) values between 1.03 °C and 1.29 °C. However, autoregressive models are limited in capturing complex nonlinear relationships. In contrast, RNNs such as Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM), can model more intricate temporal dependencies. Han et al. [13] used GRU and LSTM to forecast local air temperature 24 h ahead, achieving RMSE of 2.96 °C for GRU and 4.72 °C for LSTM, though these models were employed for forecasting rather than imputation. For gap filling, Ma et al. [23] proposed a bi-directional LSTM model for building energy data imputation, which outperformed the standard LSTM by incorporating both pre-history and post-history of the gaps. However, despite their strengths, these methods face limitations, particularly the propagation and accumulation of filling errors, which can reduce their effectiveness in addressing continuous gaps.

Some researchers have leveraged data segment similarity to fill gaps, thereby avoiding error propagation. For example, Lucbert et al. [22] utilised the hot deck method, which replaces gaps with observed values from similar cases identified based on relevant characteristics or variables. Wang et al. [38] employed Generative Adversarial Networks to fill data gaps by training a generator to capture the patterns and structures of the original dataset, producing synthetic data as gap estimates.

With advancements in deep learning, researchers have begun transforming time series data into a two-dimensional matrix to apply image processing techniques such as graph neural networks, Partial Convolution (Pconv), Convolutional Neural Networks for data imputation and forecasting [5,10,20]. While these techniques can capture both patterns and temporal relationships, their effectiveness in filling continuous gaps remain limited. For example, Fu et al. [10] used PConv to impute missing values with a large training dataset, achieving a coefficient of determination (R^2) of 0.6 for time series data with a 50% missing rate and gaps of up to two weeks. This result represents a meaningful advancement, while further improvement in accuracy may be necessary to meet practical application requirements. Moreover, these methods require substantial datasets and significant computing resources.

Some studies have addressed data gaps by using variables that are highly correlated with the time series. This approach builds relationships between the available data in the time series and corresponding highly correlated data during the same period, which are then applied to the missing periods. This method has proven effective for random gaps [1,16], though its effectiveness for continuous gaps remains unverified. Moreover, the approach is limited by the availability of highly correlated data, which may be absent for certain missing periods or entirely unavailable [16].

1.2.2. Filling performance

We further summarise the performance of different methods (Section 1.2.1) under varying missing data conditions, as shown in Table 1. The best-performing model in each study is highlighted in bold in the 'Filling model' column. To emphasise their maximum potential, the "Best filling performance" column includes only the results for the most challenging conditions (i.e., largest gaps and highest missing rates). Under these conditions, complex algorithms like LSTM and image-based techniques generally outperform others due to their ability to capture intricate temporal dependencies and data patterns.

However, regardless of the method used, filling performance generally decreases as gap length increases. Several studies [10,21–23] have highlighted the particular challenges associated with filling continuous data gaps. Regarding the missing rate, while a high rate of missing data poses less of a challenge when gaps are short [16], longer gaps combined with a high missing rate generally lead to poorer

Table 1
Filling performance for various filling methods in previous research. The best-performing model in each study is highlighted in **bold**. NRMSE: Normalised Root Mean Square Error. See footnote for details.

Data type	Worst missing condition		Length of training data	Filling methods [c]	Predictors [d]	Best filling performance	References
	Missing rate	Largest gap length					
Outdoor air temperature	0.1	30	5242 × (1 - missing rate)	Kalman Smoothed Time Series Model, Kalman Smoothed ARIMA, **MLR**	MLR: other weather parameters; Others: outdoor air temperature	RMSE = 0.9448 °C	[1]
Zone temperature of interior zone	0.4	312	1680 × (1 - missing rate)	LIN, SPL, KNN, **MICE, MF**, SVD-EM, SGD	Zone temperature of interior zone	NRMSE = 0.834	[7]
Outdoor air temperature	0.5	1	744 × (1 - missing rate)	Mean imputation, **LIN**, MLR, MLP, SVM, RF	MLR: outdoor relative humidity; Others: outdoor air temperature	NRMSE = 0.047	[16]
Outdoor air temperature	0.21	168	17,545 × (1 - missing rate)	KNN, RNN, **HD**, LOCF	Outdoor air temperature	RMSE = 14.67 °C	[22]
Energy use (Electricity consumption)	0.9	70,176 × 0.9 [a]	Length of other time series + 70,176 × (1 -missing rate) [b]	Mean imputation, LIN, KNN, SVM, RF, FCNN, RNN, LSTM, **LSTM-BIT**	Electricity consumption	R^2 = 0.3876	[23]
Indoor air temperature	0.9	22	4 years × 0.3	CONV, **FEED**, LSTM	Indoor air temperature	RMSE = 0.49 °C	[21]
Energy use (Energy consumption)	0.5	Two-week absence	1 year (8736)	Image techniques (1D-CNN, 2D-CNN, **PConv**)	Energy consumption	R^2 = 0.6	[10]

[a] 70,176 is the time series length, and 0.9 represents the missing rate, assuming the missing data forms a continuous gap.
[b] Length of other time series refers to the length of the other time series used to train the basic LSTM model, while 70,176 × (1 - missing rate) represents the data used for transfer learning.
[c] ARIMA: autoregressive integrated moving average; MLR: multiple linear regression; LIN: linear interpolation; SPL: spline interpolation; KNN: k-nearest neighbour; MICE: multiple imputations through chained equations; MLP: multilayer perception; SVM: support vector machines; RF: random forest; RNN: recurrent neural networks; HD: hot deck; LOCF: last observation carried forward; LSTM: long short-term memory; LSTM-BIT: LSTM with bi-directional imputation and transfer learning; FCNN: fully connected neural network; CONV: convolutional; FEED: feed-forward; 1D-CNN: one-dimensional convolutional neural networks; PConv: Partial Convolution.
[d] Others: other filling methods; other weather parameters: such as precipitation, humidity, wind speed, wind direction (sine and cosine transformed), sea level pressure.

performance in data gap filling [10,23].

In addition, most studies rely solely on time series themselves to fill gaps, possibly due to the lack of strong correlations with external variables or the focus on testing specific methods. However, in the case of CWS air temperature data, a strong correlation exists with meteorological variables consistently available from OWS. This correlation suggests that leveraging these relationships could serve as an effective approach for CWS data gap filling. Furthermore, the repeating patterns and cycles in CWS air temperature data indicate that relationships established during recorded periods can be effectively used to fill missing data in unrecorded periods.

1.3. Research aim and objectives

In this study, we aim to fill gaps in CWS temperature datasets across varying levels of missing data, including cases with high missing rates and continuous gaps, by leveraging their relationships with OWS meteorological data during the recorded period. To achieve this, we employ three representative data-driven methods for gap filling: Multiple Linear Regression (MLR), Random Forest (RF), and Multi-Layer Perceptron (MLP). These methods range from simple statistical techniques to more advanced machine learning algorithms, providing a comprehensive approach to addressing data gaps. For reasons of computational efficiency, deep learning methods such as LSTM and imaging techniques were not included in this study. The benchmark dataset consists of CWS temperature data recorded in London between July 5 and July 31, 2018, a period that includes several heatwaves. The research objectives are:

- To analyse the missing conditions of CWS air temperature data from July 2018 in London, focusing on missing rates and missing lengths. This analysis will identify representative missing conditions to guide the selection of the most efficient algorithm.
- To establish practical, precise, and robust training processes for the three data-driven algorithms, ensuring their optimal performance and allowing for fair comparison among them.
- To assess the impact of feature combinations, missing rates, and missing lengths on the performance of the three models, with the goal of identifying the best-performing algorithm and the optimal feature combination for gap filling.
- To test the robustness of the best-performing algorithm and feature combination by applying them to different locations.

2. Methodology

The proposed methodology is organised into five steps, as illustrated in Fig. 1. In Step 1, we analyse the raw CWS data to classify the missing data conditions, which forms the foundation for model construction and validation. Steps 2 to 4 focus on model development and evaluation, including data preparation and preprocessing to create training and test datasets. In Step 5, we design experimental scenarios to compare models trained using different algorithms and feature sets across various missing data conditions, allowing us to identify the most effective algorithm. Finally, the selected algorithm is applied to additional CWS datasets from other locations to assess the generalisation and robustness of the proposed method.

2.1. Data collection and analysis

2.1.1. CWS and OWS data

The dataset used in this study comprises input variables and the target estimate variable. CWS data was collected from July 5 to July 31, 2018, a period that includes several heatwaves in the Greater London area, provided by the Netatmo network (https://weathermap.netatmo.com/). To ensure the quality of the CWS temperature recordings, we apply the quality check procedures (see Table A1 for details) from Step M1 to Step M4 as outlined by Fenner et al. [8]. For one-point missing

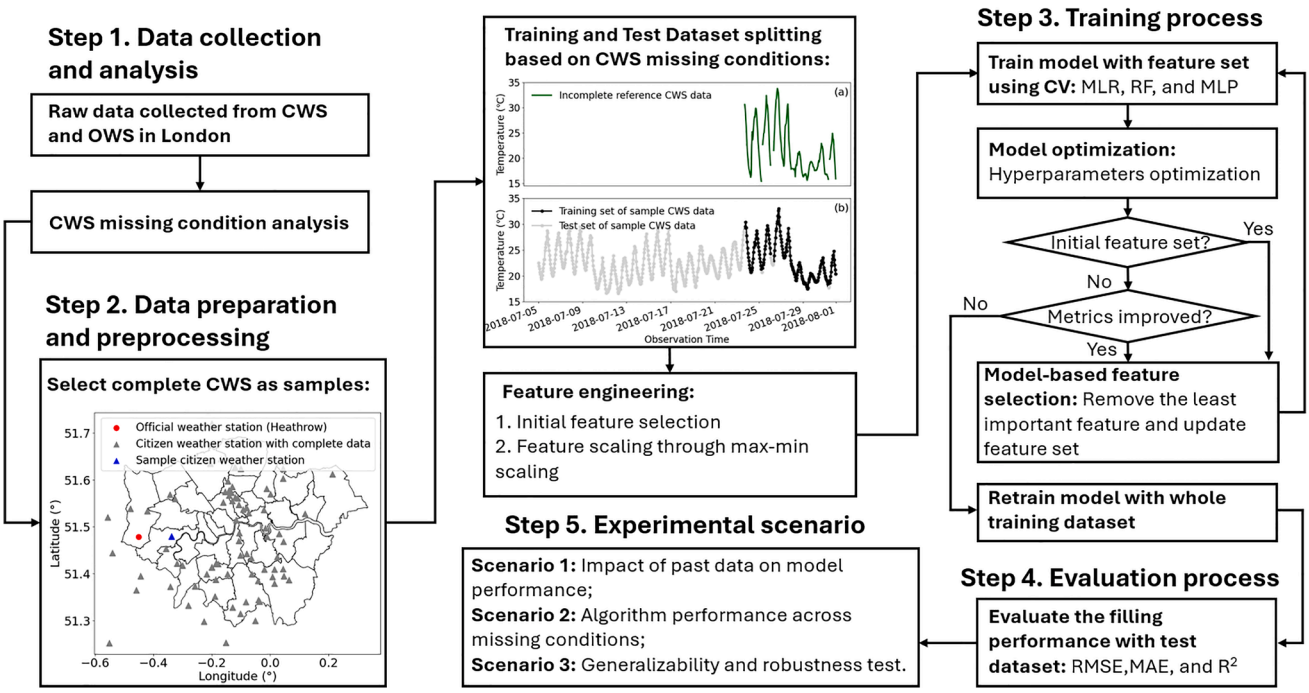


Fig. 1. Overall workflow for filling CWS temperature missing data.

gaps, linear interpolation (Step O1) is employed, as previous studies [7, 16] have demonstrated its effectiveness for small data gaps. Heathrow weather station is selected as the source for input meteorological variables due to its comprehensive meteorological dataset.

2.1.2. *Missing data condition of CWS data*

To evaluate the effectiveness of the models in addressing common missing conditions in CWS temperature data, we analyse these conditions from two main perspectives: missing rate and missing length.

Fig. 2 shows the missing rate of each CWS in London during July 2018, highlighting the prevalence of data gaps. Fig. 3 shows the ratio of CWS with missing rates exceeding various thresholds (x-axis) to the total number (1015) of CWS. Nearly half of the CWS recordings have a missing rate greater than 20%. According to the common strategy in existing research for handling missing conditions in CWS data, specifically the quality check Step O3 (Table A1), recordings with a missing rate over 20% are typically excluded from use, resulting in a substantial data loss. To improve the availability of CWS temperature data, it is necessary to develop effective data imputation methods for CWS data with high missing rates.

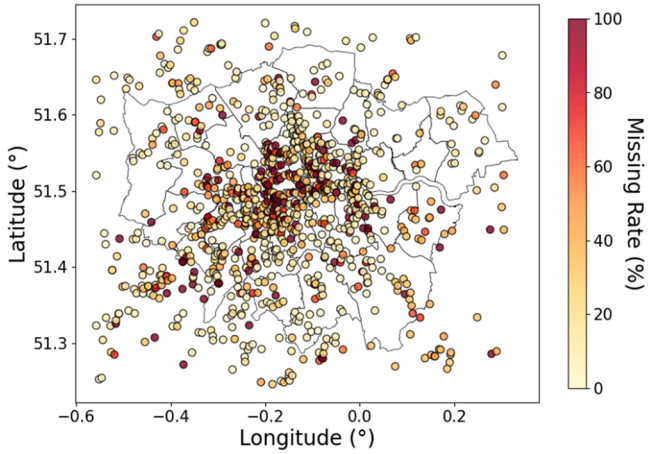


Fig. 2. Missing rates of CWSs in London in July 2018.

In addition to the missing rate, missing length is another crucial feature for describing missing data conditions. Existing researches often consider missing length separately from the missing rate. However, large missing lengths typically occur in conjunction with high missing rates. Fig. 4 shows the statistical analysis of the adjusted missing rate, defined as the proportion of missing data within each length range relative to the total missing data for each CWS in London during July 2018. As the missing rate increases, the likelihood of continuous missing gaps also increases. When the missing rate is less than 30%, the length of the missing gaps is usually within 1 day, indicating random gaps. Conversely, when the missing rate is between 30% and 80%, the missing gaps typically extend beyond 1 week (168 h), suggesting that continuous gaps dominate. When the missing rate exceeds 80%, there is insufficient recorded data to provide meaningful statistical analysis. This analysis will inform the creation of missing data scenarios for the experimental setup detailed in Section 2.5.

2.2. *Data preparation and preprocessing*

2.2.1. *Dataset selection*

In real-world applications, models are trained on recorded data to estimate missing values. To evaluate model performance in this study, we use complete CWS recordings but introduce different types of gaps. The models' effectiveness is then assessed by comparing the estimated values with the 'true' values of the missing data, i.e. the values within these artificial gaps.

Following the quality checks, only 78 out of 1015 CWS recordings are found to be complete. As noted by Afrifa-Yamoah et al. [1], the reliability of data imputation depends on the correlation between feature variables and missing data, with lower correlations typically resulting in poorer performance. To evaluate model performance under less-than-ideal conditions, we select CWS temperature data with the lowest correlation (correlation index = 0.80) to OWS temperature data from the complete dataset with its location shown in Fig. 5a. The corresponding temperature recordings are shown in Fig. 5b.

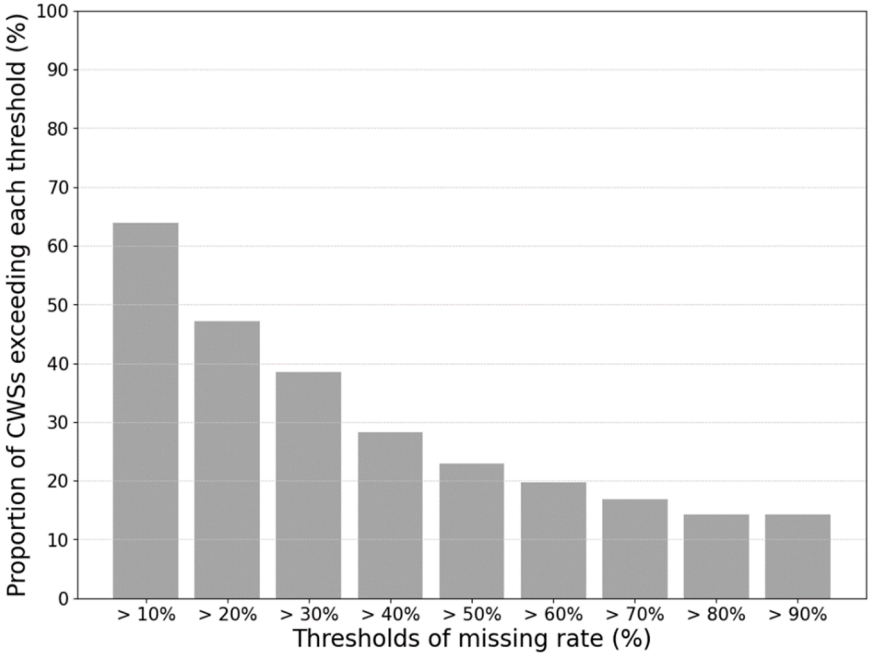


Fig. 3. Accumulated missing rate of CWSs in London in July 2018.

2.2.2. *Data splitting: training and test sets*

To create realistic missing data conditions, we used missing patterns from incomplete CWS temperature recordings to introduce artificial gaps into selected complete CWS temperature recordings. For each common missing data condition (Fig. 4), we select a representative incomplete CWS recording as a reference. Fig. 6a shows the missing patterns of a reference incomplete recording, which includes one continuous gap and several random gaps, with a total missing rate of 70.8%. The continuous gap containing most of the missing values is the dominant pattern, as analysed in Section 2.1.2. Fig. 6b shows how the complete sample data was split into training and testing datasets based on the missing patterns from the reference incomplete CWS data. The 'true' values within the artificial gaps form the testing dataset for model evaluation, while the remaining data is used for training.

2.2.3. *Initial feature selection*

As shown in Table 2, meteorological data from OWS are selected as features, along with timestamp variables to better capture the diurnal cycle [35]. Beyond these variables, there is a noticeable time offset in air temperature readings between the CWS and OWS stations (Fig. 5b) due to varying heat storage capacities of underlying land types. To improve the gap-filling accuracy, such offsets should be considered when building the relationships between CWS and OWS data by: (1) incorporating OWS air temperature data within an extended time window; (2) including past solar radiation data from OWS, since thermal storage flux is heavily influenced by solar radiation and heat storage capacity [12].

Based these considerations, we create two initial feature sets: one that includes preceding and subsequent OWS air temperature within a 12-hour window and past 24-hour solar radiation data, and one that excludes them. Details of both feature sets are provided in Table 2. Given the limited training dataset size, we will further reduce the number of feature variables to prevent overfitting, as described in Section 2.3.

2.2.4. *Feature scaling*

Feature scaling is applied to ensure that each feature contributes equally and to accelerate algorithm convergence. We use min-max scaling, a commonly robust scaling method [40]. As shown in Fig. 7, for the training dataset, min-max scaling is directly applied to each feature and output variable based on their respective maximum and minimum values. While for the test dataset, we normalise and inverse-normalise the test dataset using the scaling parameters derived from the training dataset. This is designed to match real-world scenarios where test data is unavailable and the scaling parameters for the test output are unknown.

2.3. *Training process*

In this section, we will explain the principles of Multiple Linear Regression (MLR), Random Forest (RF), and Multilayer Perceptron (MLP), and how these algorithms are used to train the models. The training procedures for the models are shown in Fig. 8. The MLR training process involves three main steps: model structure building, feature selection and retraining the model on the entire training dataset. In contrast, the machine learning (ML)-based model training process includes an additional fourth step for hyperparameter optimisation. The principle of each algorithm and operational details for each step will be further explained below.

2.3.1. *Fundamental characteristics and applications of algorithms*

MLR assumes linear relationships between feature variables and target variables, estimating missing values based on this linear model. Thus, it requires minimal computational resources. Additionally, the fitting coefficients directly reflect the importance of the feature variables, making the MLR model easy to interpret.

RF is an ensemble method that estimates missing values by using multiple decision trees, each trained on different subsets of the data. Each decision tree is a nonlinear model providing its own estimate, and the final estimate is the average of these individual estimates. This approach allows RF to capture nonlinear relationships and effectively handle complex data structures. In addition, RF assesses feature importance by measuring how much each feature contributes to reducing impurity across all decision trees [2]. This information is valuable for feature selection and model interpretation and is thus used in selecting features for ML-based model training in this study.

MLP estimates missing values by modelling the relationships between features and target variables through multiple layers of neurons. Compared to MLR and RF, MLP can capture more complex nonlinear relationships between variables [27]. However, MLP generally offers

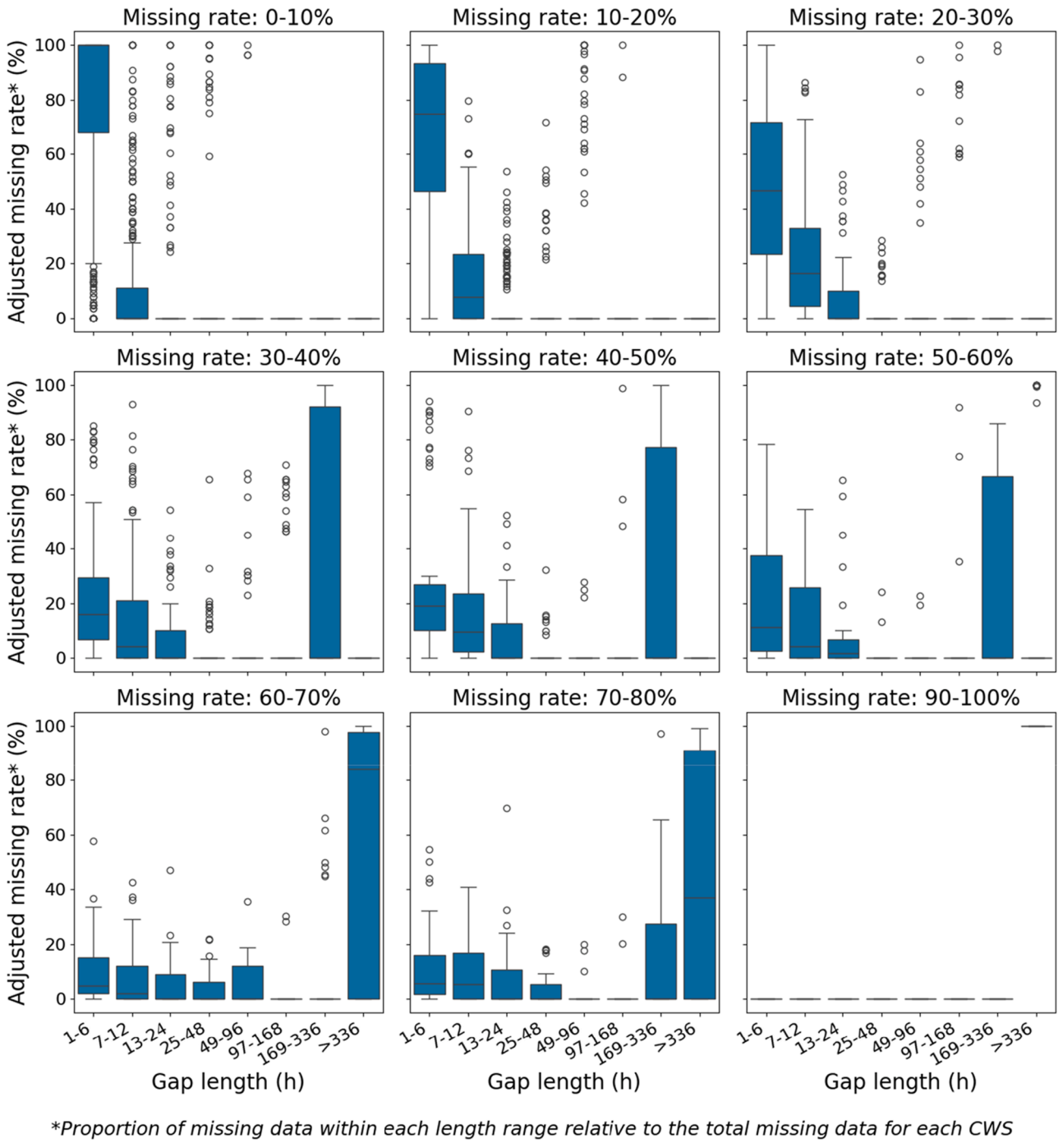


Fig. 4. Missing lengths of CWSs within each missing rate range in London in July 2018.

lower interpretability compared to the other two models.

2.3.2. Operational details for training process

Step 1: Model structure building

The models, including MLR, RF, and MLP, are built using the Keras library with TensorFlow [11]. Specifically, MLR uses the 'LinearRegression' module, RF uses the 'RandomForestRegressor' module, and MLP uses the 'MLPRegressor' module. For more details, please refer to the TensorFlow website.

Step 2: Feature selection

Due to the large number of variables in the initial feature sets (Table 2) and the limited training data, feature selection is necessary to prevent model overfitting [14]. Given the high correlation among features (i.e., multicollinearity) affecting linear models such as MLR [18], we adopt a correlation-based feature selection to reduce the multicollinearity. This involves identifying highly correlated pairs using the pearson correlation matrix, and iteratively dropping the feature with the lowest correlation to the target variable from highly correlated pairs until no feature pairs exceed a threshold of correlation of 0.9, ensuring a more independent set of features. For non-parametric models like RF and MLP, multicollinearity does not affect predictive ability [25], so this step is omitted. Then, we refine the traditional stepwise method by using feature importance derived from RF, and the coefficients from MLR.

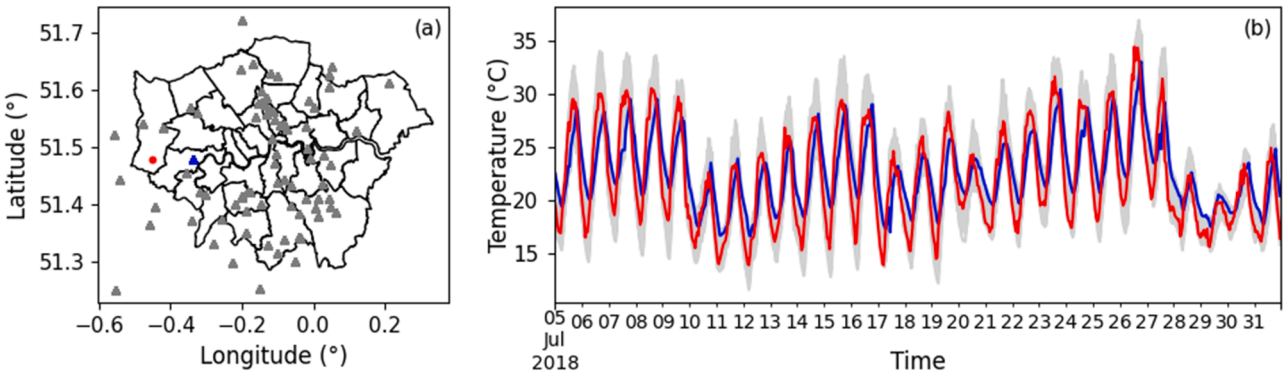


Fig. 5. (a) Spatial distribution of sample CWS (blue), official weather stations (red), and remaining CWSs (grey); (b) Corresponding temperature recordings over time.

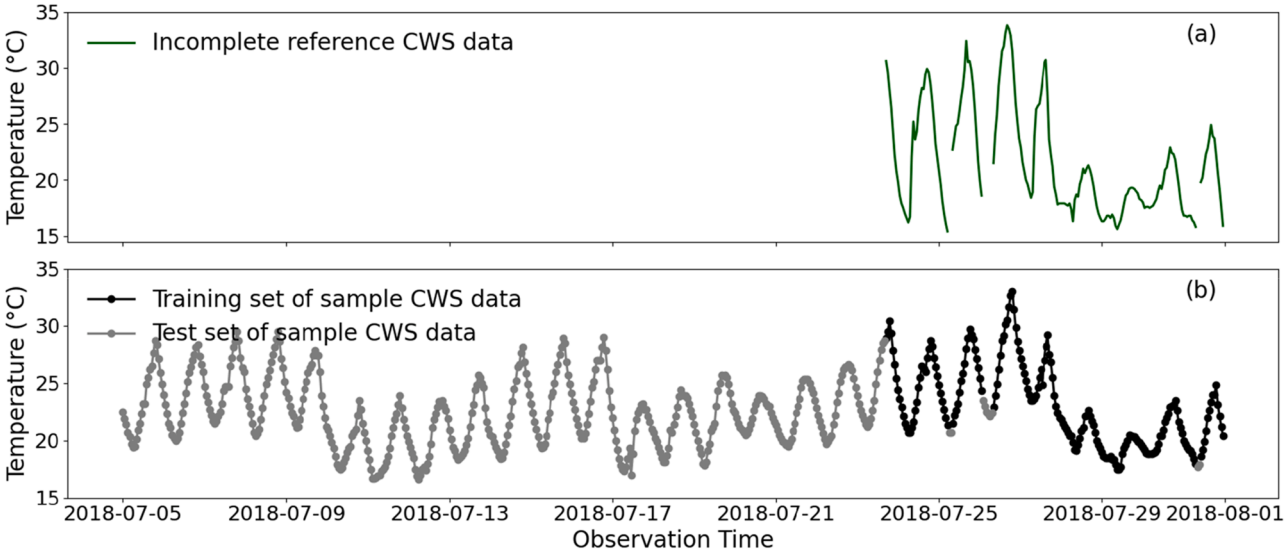


Fig. 6. Splitting training and test set based on missing information of other incomplete reference CWS: (a) Missing patterns of reference incomplete CWS; (b) Training and test set splitting of sample CWS dataset.

Table 2
Variables for initial feature set 1 and 2.

Variables		Initial feature set 1	Initial feature set 2
Meteorological data from OWS	Air temperature	✓	✓
	Dewpoint	✓	✓
	Wet bulb temperature	✓	✓
	Relative humidity	✓	✓
	Wind speed	✓	✓
	Wind direction	✓	✓
	Station pressure	✓	✓
	Global solar irradiation amount	✓	✓
	Cloud total amount	✓	✓
	Derived hourly sunshine duration	✓	✓
	Precipitation amount	✓	✓
Past data from OWS	Air temperature from the past 12 h	×	✓
	Global solar irradiation from the past 24 h	×	✓
Subsequent data from OWS	Air temperature from the next 12 h	×	✓
Timestamp data	Sine and cosine function of the hour of the day	✓	✓

Instead of iteratively removing each feature to evaluate its significance, we use the feature importance scores to directly eliminate the least important feature in each round. This approach reduces computational burden and speeds up model optimisation. To further mitigate overfitting risk for RF and MLP, we introduce the training sample size to the feature size ratio (SFR), requiring it exceeds 10 for reliable modelling [40]. Notably, the appropriate SFR depends on the complexity of the problem. For the winter season, where weather patterns are relatively complex (Fig. S2), an SFR exceeding 30 is necessary to achieve robust gap-filling performance. We then train the model with the updated feature set and evaluated its performance using 5-fold cross-validation (CV). CV is selected because it provides a reliable estimate of model performance and reduces the risk of overfitting, especially with small datasets [15]. The optimal combination of features will lead to the best-performing model.

Step 3: Hyperparameter optimisation

For machine learning algorithms such as RF and MLP, hyperparameter optimisation is crucial, as it directly impacts model performance, generalisation, robustness, efficiency, and interpretability [39]. In this study, we use Bayesian optimisation to find the optimal hyperparameters due to its efficiency in locating the global optimum [31]. In each round, Bayesian optimisation identifies a promising combination of

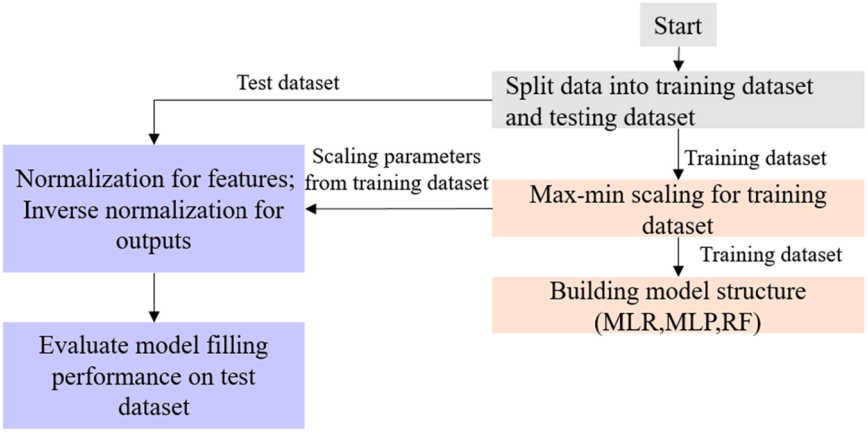


Fig. 7. Data preprocessing for training and test dataset.

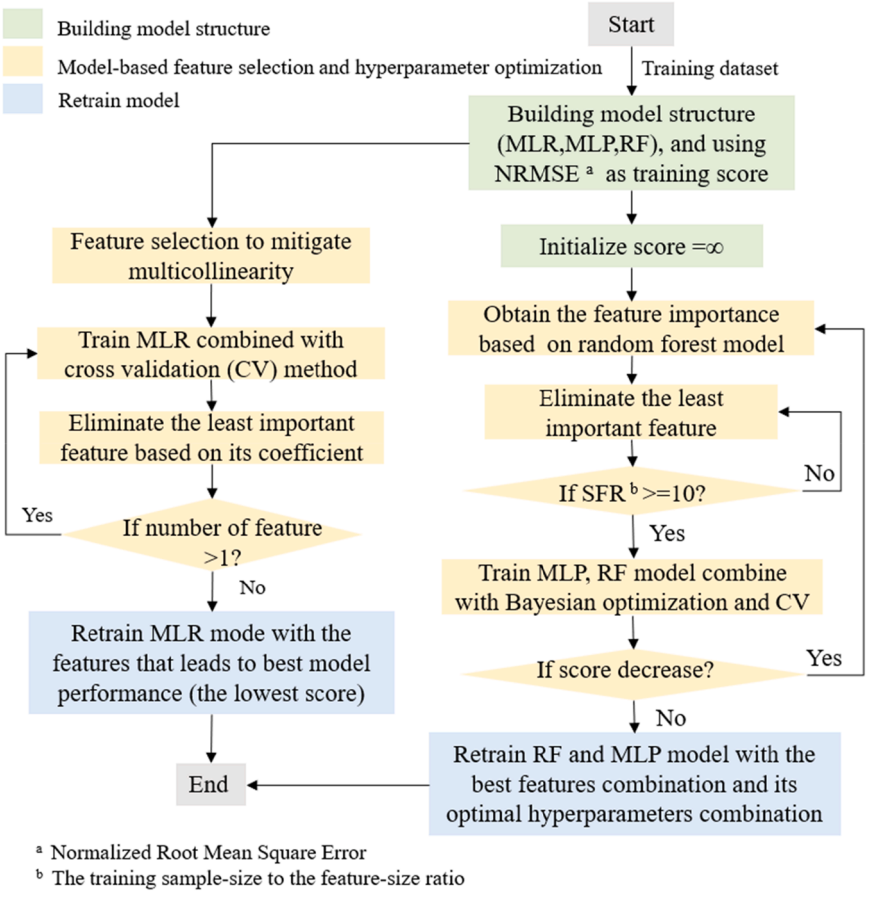


Fig. 8. Workflow of building MLR, MLP and RF.

hyperparameters, after which we train the model with these parameters and evaluate its performance using CV. This process is facilitated by the 'BayesSearchCV' module.

Step 4: Retraining model on the entire training dataset

During above selection process, one fold of the training dataset is always set aside for evaluation, separate from the training. Thus, after selection part, we retrain the model using entire training dataset with the best features and hyperparameter combinations.

2.4. Evaluation process

The well-trained models are evaluated on the test dataset. To conduct a comprehensive evaluation, we use metrics including mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). MAE measures the overall accuracy of the filled data, while RMSE, being sensitive to larger errors, helps highlight and quantify the model's performance under varying scenarios, such as heatwaves and non-heatwaves. R^2 evaluates the models' ability to capture the overall trend. These metrics are defined in Eqs. (1) to (3):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - y_{mean})^2} \quad (3)$$

where y_i presents the observed gap value, $\hat{y}_i$ presents the corresponding estimated gap value, y_{mean} presents the mean of all observed gap values, and N presents the amount of gap values.

2.5. Experiment scenarios

Table 3 outlines the experimental setup. In Scenario 1, we investigate whether including preceding and subsequent data in the initial feature set can improve model performance or not. Scenario 2 focuses on identifying the most efficient algorithm for handling various common missing conditions in CWS temperature data. Based on the missing data condition analysis in Section 2.1.2, we create CWS missing conditions with missing rates from 0 to 80%. For missing rates below 30%, random gaps are introduced, while continuous gaps are generated for missing rates between 30% and 80%. To further investigate the impact of missing length on filling performance, we also include missing conditions with a 70% to 80% missing rate and random gaps, as a contrast to continuous gaps at the same missing rate. The missing patterns for all conditions are based on incomplete CWS recordings, with the method details provided in Section 2.2.2. The experiment Scenarios 1 and 2 are conducted under a less-than-ideal condition, as described in Section 2.2.1. In Scenario 3, we apply the best-performing algorithm from Scenario 2 to the remaining complete CWS recordings. This allows us to evaluate the generalisability and effectiveness of the method across CWS located in different areas with varying spatial characteristics.

Table 3
Details about experiment setting.

Scenario	CWS selected	Initial feature set	Algorithm	Missing rate (%) [a]	Main gap types [b]
1	Sample CWS	1	MLR, RF, MLP	70–80 (70.8)	Continuous (449)
		2		70–80 (70.8)	Continuous (449)
2	Sample CWS	Feature set proven more effective in scenario 1	MLR, RF, MLP	0–10 (5.2)	Random (4)
				10–20 (16.9)	Random (8)
				20–30 (25.2)	Random (7)
				30–40 (36.4)	Continuous (236)
				40–50 (43.5)	Continuous (211)
				50–60 (54.5)	Continuous (353)
				60–70 (66.2)	Continuous (420)
				70–80 (70.8)	Continuous (449)
				70–80 (71.0)	Random (19)
3	Remaining complete CWS	Feature set proven more effective in scenario 1	Algorithm proven most effective in scenario 2	70–80 (70.8)	Continuous (449)

[a] The values in parentheses in this column represent the specific missing rate for the selected referenced incomplete CWS recordings.

[b] The values in parentheses in this column represent the length of the largest gap for the selected referenced incomplete CWS recordings, measured in hours (h).

3. Results and discussion

3.1. Scenario 1: impact of past data on model performance

Fig. 9 and Fig. 10 show the filling performance of various algorithms (MLR, MLP and RF) using different initial feature sets for a missing condition characterised by a 70.8% missing rate and continuous gaps. The OWS air temperature data serve as benchmark estimates. When the initial feature set 1 (Table 2) is used, MLR-based model (Fig. 9e) outperforms both MLP-based model (Fig. 9f) and RF-based model (Fig. 9g) in filling performance.

Specifically, noticeable temperature misalignments are observed between the CWS air temperature and OWS air temperature (Fig. 9d), particularly in time offsets and extreme values. The MLR-based model addresses time offsets well but struggles with the extreme values (Fig. 9a). This is because linear model is effective at capturing consistent trends over time, and the temporal offsets follow a regular cycle. However, extreme temperatures often involve non-linear dynamics and sudden shifts that MLR cannot capture well.

For weather patterns characterised by persistent low CWS air temperatures during non-heatwave periods, the MLP-based and RF-based models tend to overestimate temperatures, while they perform not bad during heatwave periods (Fig. 9b, c). However, their training set contains data from both heatwaves and non-heatwaves (Fig. 6b). This suggests that the MLP-based and RF-based models are underfitting when using feature set 1, as its features are insufficient to represent differences between heatwave and non-heatwave conditions.

Conversely, when using initial feature set 2, which includes air temperature within a time window and global solar irradiation from the past 24 h, the performance of all algorithms improves significantly (Fig. 10). The MLP-based model demonstrates the greatest improvement, achieving a MAE of 0.59 °C, RMSE of 0.73 °C, and R^2 of 0.94 (Fig. 10f). The MLR-based and RF-based models also show good results (Fig. 10e, g). But for estimating the extreme temperatures, they do not perform as well as MLP-based model (Fig. 10a, b, c). This is also because extreme temperatures often involve non-linear dynamics and sudden shifts that MLP model captures more effectively than the linear MLR and RF models. Additionally, above results confirm that the OWS meteorological data are highly correlated with the CWS temperature data, and their relationships can be used to estimate the missing values effectively.

3.2. Scenario 2: algorithm performance across missing conditions

Table 4 compares the filling performance of various algorithms across different missing conditions. MLP models achieve MAE ranging from 0.26 to 0.69 °C, RMSE from 0.35 to 0.88 °C, and R^2 from 0.92 to 0.99. In contrast, the RF models yield MAE values from 0.45 to 0.73 °C, RMSE from 0.57 to 1.00 °C, and R^2 from 0.90 to 0.97. The MLR models show MAE ranging from 0.41 to 0.75 °C, RMSE from 0.61 to 1.04 °C, and R^2 from 0.87 to 0.97. These performances remain consistent even with high missing rates (70–80%). Overall, MLP outperforms both RF and MLR due to its superior ability to model the predominantly non-linear relationships between OWS data and CWS data, as discussed in Section 3.1.

As the missing rate increases, there are no large decreases in filling performance across various models (Table 4). This indicates that these methods are not particularly sensitive to varying missing rates. This stability is mainly due to the use of OWS meteorological data as reference during the missing periods, rather than relying on the temporal patterns of the remaining CWS temperature data. The latter can be challenging to capture accurately when the training dataset is small, as

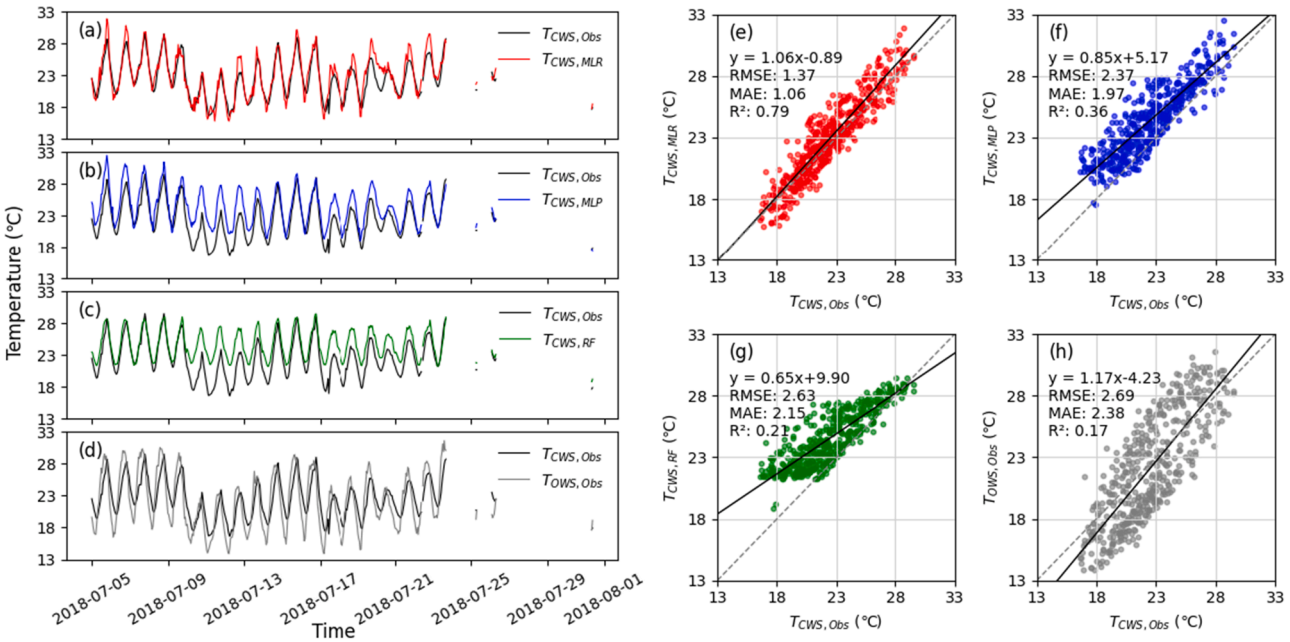


Fig. 9. Filling results and performance comparisons for different models using the initial feature set 1: CWS air temperature filled by (a) MLR; (b) MLP; (c) RF; (d) OWS; and their corresponding performance (e) - (f).

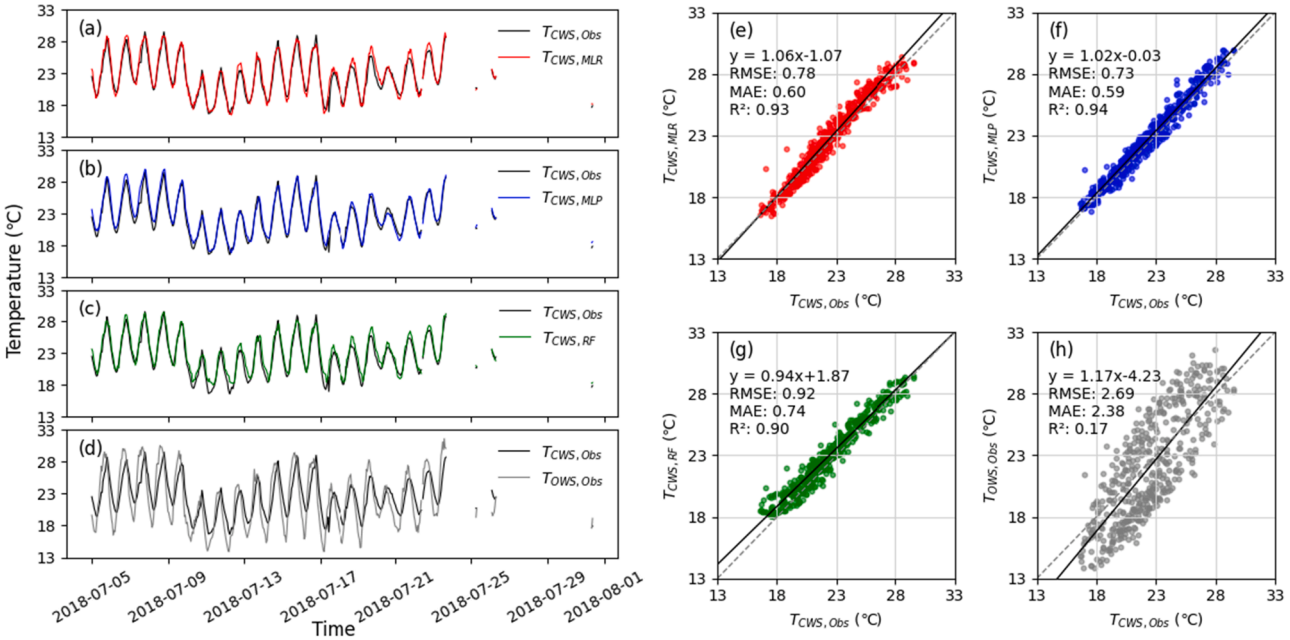


Fig. 10. As Fig. 9, but with initial feature set 2.

mentioned in previous literature [23]. Notably, the filling performance is better when the missing rate is 0–10%. This results from the training dataset encompassing a broader range of weather conditions, enabling the model to capture the overall relationships more accurately.

For missing conditions with a 70–80% missing rate featuring both random and continuous gaps, there are no large differences in the filling performance (Table 4). This implies that the length of the missing periods is not the primary factor affecting the performance of this method. Instead, the representation of the training dataset appears to be more critical. With the same missing rate, conditions with random gaps are likely to ensure that the training and test datasets include similar weather conditions. This similarity allows the training dataset to better represent the test dataset and the entire data distribution, resulting in an improved filling performance.

To verify that the representation of the training dataset is a primary factor, we compare filling performance at similar missing rates using training datasets that include only non-heatwave periods, only heatwave periods, and a combination of both (Table 5). The results show that the model trained only with heatwave periods performs worse than the other two, with metrics of MAE at 0.61 °C, RMSE at 0.79 °C, and R^2 at 0.93. This is due to the limited number of heatwave days in July compared to non-heatwave days. Consequently, the model trained solely on heatwave data fails to capture the relationships present during non-heatwave periods, leading to relatively poor filling performance. This confirms that the representation of the training dataset is a critical factor in model performance.

Table 4
Filling performance of various algorithms under different missing conditions. Best performance in '**bold**'.

Missing rate [a]	Main gap types [b]	Algorithms	MAE (°C)	RMSE (°C)	R^2
0–10 % (5.2%)	Random gap (4 h)	MLR	0.54	0.69	0.87
		MLP	**0.26**	**0.35**	**0.97**
		RF	0.41	0.52	0.92
		Ref.	3.47	3.74	−2.87
10–20 % (16.9%)	Random gap (8 h)	MLR	0.64	0.82	0.92
		MLP	**0.47**	**0.71**	**0.94**
		RF	0.59	0.82	0.92
		Ref.	2.97	3.26	−0.32
20–30 % (25.2%)	Random gap (7 h)	MLR	0.49	0.62	0.97
		MLP	**0.30**	**0.39**	**0.99**
		RF	0.41	0.61	0.97
		Ref.	2.22	2.56	0.46
30–40 % (36.4%)	Continuous gap (236 h)	MLR	0.50	0.66	0.96
		MLP	**0.42**	**0.54**	**0.98**
		RF	0.75	1.04	0.91
		Ref.	2.36	2.73	0.37
40–50 % (43.5%)	Continuous gap (211 h)	MLR	0.61	0.76	0.95
		MLP	**0.40**	**0.52**	**0.98**
		RF	0.66	0.96	0.92
		Ref.	2.49	2.83	0.28
50–60 % (54.5%)	Continuous gap (353 h)	MLR	0.53	0.69	0.95
		MLP	**0.50**	**0.62**	**0.96**
		RF	0.62	0.88	0.92
		Ref.	2.26	2.61	0.31
60–70 % (66.2%)	Continuous gap (420 h)	MLR	0.66	0.85	0.93
		MLP	0.69	0.88	0.92
		RF	**0.61**	**0.85**	**0.93**
		Ref.	2.37	2.72	0.26
70–80 % (70.8%)	Continuous gap (449h)	MLR	0.60	0.78	0.93
		MLP	**0.59**	**0.73**	**0.94**
		RF	0.74	0.92	0.90
		Ref.	2.38	2.69	0.17
70–80 % (71.0%)	Random gap (19h)	MLR	0.55	0.68	0.96
		MLP	**0.35**	**0.45**	**0.98**
		RF	0.52	0.70	0.95
		Ref.	2.34	2.68	0.31

[a] The values in parentheses in this column represent the specific missing rate for the selected referenced incomplete CWS recordings.
[b] The values in parentheses in this column represent the length of the largest gap for the selected referenced incomplete CWS recordings, measured in hours (h).

Table 5
Filling performance across various gap types at a 70–80% missing rate.

Missing rate [a]	Main gap types [b]	Algorithms	MAE (°C)	RMSE (°C)	R^2
70–80% (79.6%)	Continuous gap: only including non-heatwave (276 h)	MLP	0.52	0.69	0.96
		Ref.	2.41	2.75	0.31
70–80% (78.7%)	Continuous gap: only including heatwave (407 h)	MLP	0.61	0.79	0.93
		Ref.	2.25	2.57	0.23
70–80% (79.3%)	Continuous gap: including heatwave and non-heatwave (472 h)	MLP	0.56	0.70	0.95
		Ref.	2.33	2.68	0.27

[a] The values in parentheses in this column represent the specific missing rate for the selected referenced incomplete CWS recordings.
[b] The values in parentheses in this column represent the length of the largest gap for the selected referenced incomplete CWS recordings, measured in hours (h).

3.3. Scenario 3: generalisability and robustness test

Fig. 11 shows the performance of the MLP algorithm in filling artificially introduced gaps within the CWS temperature recordings. Each originally complete recording is processed under the worst-case missing condition (a missing rate of 70–80% with a continuous gap) as detailed in Table 3. The results show that the MAE across all samples ranges from 0.42 to 1.07 °C, the RMSE from 0.53 to 1.35 °C, and the R^2 from 0.87 to 0.98. Even the worst metrics outperform those from previous studies (Table 1) that used other imputation methods on different types of missing data.

The strong performance across different CWS locations, despite their varied spatial characteristics, shows the robustness of our approach. This effectiveness is due to training an individual MLP-based model for each CWS. Thermal storage differences in land types are accounted for by including past OWS air temperature and global solar radiation data in the initial feature set, as verified in Section 3.1. Feature selection during training identifies the most relevant features, enabling the development of a tailored model for each CWS. These individual models adapt to variations in urban form and geography, with their distinct selected features and corresponding coefficients reflecting the influence of local conditions on each CWS. Consequently, the approach enables effective gap filling and consistent accuracy across diverse urban areas.

Overall, MLP-based models show reliable filling performance across various missing conditions (Tables 4 and 5) and urban areas (Fig. 11) with R^2 above 0.87, MAE below 1.07 °C, and RMSE below1.35 °C. These results suggest that weather conditions within a one-month period exhibit sufficient consistency, making the training dataset—based on CWS availability in real-world scenarios—representative and supporting the model's effectiveness in imputing missing data. Moreover, given that CWS locations often change monthly (Fig. S1), filling CWS gaps on a month-to-month basis is considered suitable based on our findings. To further validate this approach and assess its adaptability across seasons, we tested it using CWS and OWS data from December 2018 in London (Fig. S2), demonstrating consistent performance across different seasons (Fig. S3).

4. Conclusion

Crowdsourced data from citizen weather stations (CWS) are widely used in urban climate studies. However, crowdsourced data often have high rates of missing data with continuous gaps, which limits their applicability. Currently, no efficient method exists to address these gaps. In view of this, this paper introduces a novel data-driven method for addressing common missing data conditions in CWS air temperature datasets, especially continuous gaps, by using the relationships between CWS and Official Weather Station (OWS) records during periods of data availability.

We identify representative missing data conditions in CWS air temperature datasets, and find that the frequency of continuous gaps (exceeding one week) increases as the overall missing data rate rises. When the missing rate is below 30%, random gaps are more prevalent. And while continuous gaps dominate when the missing rate falls between 30% and 80%.

We compare the performance of three data-driven algorithms (Multiple Linear Regression (MLR), Random Forest (RF), and Multilayer Perceptron (MLP)) under various representative missing conditions. MLP is found to be the most effective algorithm, with performance metrics showing MAE between 0.26 and 0.69 °C, RMSE between 0.35 and 0.88 °C, and R^2 between 0.92 and 0.99. It outperforms the MLR and RF models, indicating that the relationships between CWS air temperature and OWS variables are predominantly non-linear. Due to computational constraints, we only compare three algorithms. While other algorithms may exist to better capture the relationships, our results demonstrate that the MLP is sufficient for achieving a reasonable accuracy in data imputation. Future research could explore alternative algorithms, building on the insights provided by this study for model development.

We further validate the MLP algorithm using the CWS datasets from various urban areas, where even its worst performance results in an MAE below 1.07 °C, RMSE below 1.35 °C, and R^2 above 0.87. The winter scenario (Fig. S3) further demonstrates its applicability and adaptability

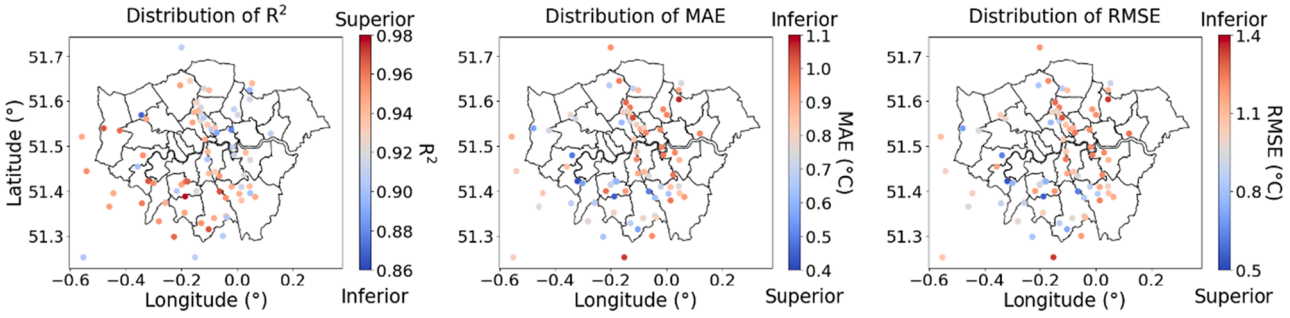


Fig. 11. Filling performance of MLP-based model with initial feature set 2.

across regions, with MAE ranging from 0.36 °C to 0.84 °C, RMSE from 0.50 °C to 1.15 °C, and R^2 from 0.87 to 0.97. These consistent results across different urban areas and seasons confirm the effectiveness and robustness of the proposed method for filling common missing data conditions in CWS air temperature datasets.

We also evaluate the impact of different feature sets on model performance. Our findings show that including air temperature within a time window and past global solar radiation as features during training significantly improves all models. For CWS datasets with varying spatial characteristics, a model-based feature selection process can identify the most relevant features from the initial set, ensuring consistent filling performance. This training process can serve as a valuable reference for other machine learning applications, providing insights into effective feature selection and model training strategies for similar tasks.

The representativeness of weather conditions in the training dataset is crucial for our method's effectiveness. Our findings indicate that filling CWS data on a month-to-month basis is appropriate, as datasets obtained during any available period within a month are sufficiently representative for model training and effectively fill missing data for that month. Due to practical constraints, such as the frequent monthly relocation of CWS, testing the method over longer periods is not conducted. Theoretically, the complete and highly correlated OWS data can serve as a reliable reference for assessing the representativeness of weather conditions. It is recommended that longer climate-related time series datasets with varying patterns are studied to assess the technique further.

The training and test datasets for each CWS do not directly account for local geographical conditions and urban form. However, since each CWS recording has a tailored model, these local conditions are indirectly represented within the model. Consequently, our approach can be adapted to various locations by developing a tailored model for each individual CWS station.

The entire training process is automated. Upon inputting a dataset, the proposed method automatically trains the model until the cross-validation score, measured by normalised root mean square error, stabilises. The finalised model is then used to fill gaps in CWS data, improving the usefulness of CWS datasets for urban climate research and enabling a more precise analysis of urban air temperature distributions.

However, while the features and coefficients of individual models reflect the influence of urban context and geographical conditions on CWS air temperature, they do not directly reveal how these factors drive temperature variations. Additionally, the spatial inequality in the distribution of crowdsourced measurements may limit the method's applicability in some areas. Future research will focus on modelling the relationships between filled CWS data and spatial features, such as buffered building data, to fill spatial gaps, thereby further enhancing the utility of CWS data for spatial analysis in urban climate studies.

CRediT authorship contribution statement

Miao He: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Zhiwen Luo:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Xiaoxiong Xie:** Writing – review & editing, Visualization, Data curation. **Peng Wang:** Writing – review & editing, Methodology. **Haichao Wang:** Writing – review & editing. **Gabriela Zapata-Lancaster:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Miao He reports financial support was provided by China Scholarship Council. Zhiwen Luo reports financial support was provided by The Royal Society. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work is partly supported by UK China Royal Society -NSFC international exchange fund (IEC\NSFC\223541 and 52311530087). MH would also like to express sincere gratitude to China Scholarship Council (CSC) and Cardiff University for providing PhD studentship.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.buildenv.2025.112593.

Data availability

Data will be made available on request.

References

[1] E. Afrifa-Yamoah, U.A. Mueller, S.M. Taylor, A.J. Fisher, Missing data imputation of high-resolution temporal climate time series data, Meteorol. Appl. 27 (1) (2020), https://doi.org/10.1002/met.1873.
[2] K.J. Archer, R.V. Kimes, Empirical characterization of random forest variable importance measures, Comput. Stat. Data Anal. 52 (4) (2008) 2249–2260, https://doi.org/10.1016/j.csda.2007.08.015.
[3] S. Bell, D. Cornford, L. Bastin, How good are citizen weather stations? Addressing a biased opinion, Weather 70 (3) (2015) 75–84, https://doi.org/10.1002/wea.2316.
[4] K. Benjamin, Z. Luo, X. Wang, Crowdsourcing urban air temperature data for estimating urban heat island and building heating/cooling load in london, Energies (Basel) 14 (16) (2021), https://doi.org/10.3390/en14165208.
[5] C. Betancourt, C.W.Y. Li, F. Kleinert, M.G. Schultz, Graph machine learning for improved imputation of missing tropospheric ozone data, Environ. Sci. Technol. (2023), https://doi.org/10.1021/acs.est.3c05104.
[6] O. Brousse, C. Simpson, N. Walker, D. Fenner, F. Meier, J. Taylor, C. Heaviside, Evidence of horizontal urban heat advection in London using six years of data from a citizen weather station network, Environ. Res. Lett. 17 (4) (2022), https://doi.org/10.1088/1748-9326/ac5c0f.
[7] B. Cho, T. Dayrit, Y. Gao, Z. Wang, T. Hong, A. Sim, K. Wu, Effective missing value imputation methods for building monitoring data, in: Proceedings - 2020 IEEE

International Conference on Big Data, Big Data 2020, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 2866–2875, https://doi.org/10.1109/BigData50022.2020.9378230.
[8] D. Fenner, B. Bechtel, M. Demuzere, J. Kittner, F. Meier, CrowdQC+—A quality-control for crowdsourced air-temperature observations enabling world-wide urban climate applications, Front. Environ. Sci. 9 (2021), https://doi.org/10.3389/fenvs.2021.720747.
[9] D. Fenner, A. Holtmann, F. Meier, I. Langer, D. Scherer, Contrasting changes of urban heat island intensity during hot weather episodes, Environ. Res. Lett. 14 (12) (2019), https://doi.org/10.1088/1748-9326/ab506b.
[10] Fu, C., Quintana, M., Nagy, Z. and Miller, C. 2023. Filling time-series gaps using image techniques: multidimensional context autoencoder approach for building energy data imputation. Available at: http://arxiv.org/abs/2307.05926.
[11] A. Géron, Hands-On Machine Learning With Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, 2022. Available at: https://books.google.co.uk/books?id=X5ySEAAAQBAJ.
[12] Grimmond, C.S.B., Cleu~ht, H.A. and Oke~, T.R. 1991. *An objective Urban heat storage model and its comparison with other schemes*.
[13] J.M. Han, Y.Q. Ang, A. Malkawi, H.W. Samuelson, Using recurrent neural networks for localized weather prediction with combined use of public airport data and on-site measurements, Build. Environ. 192 (2021), https://doi.org/10.1016/j.buildenv.2021.107601.
[14] D.M. Hawkins, The problem of overfitting, J. Chem. Inf. Comput. Sci. 44 (1) (2004) 1–12, https://doi.org/10.1021/ci0342472.
[15] D.M. Hawkins, S.C. Basak, D. Mills, Assessing model fit by cross-validation, J. Chem. Inf. Comput. Sci. (2003) 579–586, https://doi.org/10.1021/ci025626i.
[16] J. Kim, Y. Kwak, S.-H. Mun, J.-H. Huh, Imputation of missing values in residential building monitored data: energy consumption, behavior, and environment information, Build. Environ. (2023) 110919. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0360132323009460.
[17] R. Kotharkar, A. Ghosh, Review of heat wave studies and related urban policies in South Asia, Urban. Clim. 36 (2021), https://doi.org/10.1016/j.uclim.2021.100777.
[18] Kumar Paul, R. 2014. *Multicollinearity: causes, effects and remedies*. Available at: https://www.researchgate.net/publication/255640558.
[19] S.K. Kwak, J.H. Kim, Statistical data preparation: management of missing values and outliers, Korean J. Anesthesiol. 70 (4) (2017) 407–411, https://doi.org/10.4097/kjae.2017.70.4.407.
[20] P. Li, Y. Yu, D. Huang, Z.H. Wang, A. Sharma, Regional heatwave prediction using graph neural network and weather station data, Geophys. Res. Lett. 50 (7) (2023), https://doi.org/10.1029/2023GL103405.
[21] A. Liguori, R. Markovic, T.T.H. Dam, J. Frisch, C. van Treeck, F. Causone, Indoor environment data time-series reconstruction using autoencoder neural networks, Build. Environ. 191 (2021), https://doi.org/10.1016/j.buildenv.2021.107623.
[22] A. Lucbert, et al., Time Series building energy systems data imputation, in: CLIMA 2022 Conference, 2022, https://doi.org/10.34641/clima.2022.302.
[23] J. Ma, J.C.P. Cheng, F. Jiang, W. Chen, M. Wang, C. Zhai, A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data, Energy Build. 216 (2020), https://doi.org/10.1016/j.enbuild.2020.109941.
[24] F. Meier, D. Fenner, T. Grassmann, M. Otto, D. Scherer, Crowdsourcing air temperature from citizen weather stations for urban climate research, Urban Clim. 19 (2017) 170–191, https://doi.org/10.1016/j.uclim.2017.01.006.
[25] I. Morlini, On multicollinearity and concurvity in some nonlinear multivariate models, Stat. Methods Appl. 15 (1) (2006) 3–26, https://doi.org/10.1007/s10260-006-0005-9.
[26] C.L. Muller, et al., Crowdsourcing for climate and atmospheric sciences: current status and future potential, Int. J. Climatol. 35 (11) (2015) 3185–3203, https://doi.org/10.1002/joc.4210.
[27] F. Murtagh, Multilayer perceptrons for classification and regression, Neurocomputing 2 (5–6) (1991) 183–197, https://doi.org/10.1016/0925-2312(91)90023-5.
[28] S. Rahmstorf, D. Coumou, Increase of extreme events in a warming world, Proc. Natl. Acad. Sci. U.S.A. 108 (44) (2011) 17905–17909, https://doi.org/10.1073/pnas.1101766108.
[29] Sarafanov, M., Nikitin, N.O. and Kalyuzhnaya, A.V. 2021. Automated data-driven approach for gap filling in the time series using evolutionary learning. Available at: http://arxiv.org/abs/2103.01124.
[30] K.H. Schlünzen, S. Grimmond, A. Baklanov, Guidance on Measuring, Modelling and Monitoring the Canopy Layer Urban Heat Island (CL-UHI), World Meteorological Origanization, Geneva, 2023.
[31] Sevilla-Salcedo, C., Gallardo-Antolín, A., Gómez-Verdejo, V. and Parrado-Hernández, E. 2022. Bayesian learning of feature spaces for multitasks problems. Available at: http://arxiv.org/abs/2209.03028 [Accessed: 6 September 2024].
[32] H.T. Shahraiyni, S. Sodoudi, High-resolution air temperature mapping in urban areas a review on different modelling techniques, Therm. Sci. 21 (6) (2017) 2267–2286, https://doi.org/10.2298/TSCI150922094T.
[33] S. Bell, D. Cornford, L. Bastin, The state of automated amateur weather observations, Weather 68 (2) (2013) 36–41, https://doi.org/10.1002/wea.1980. Available at:[Accessed: 25 May 2024].
[34] I.D. Stewart, T.R. Oke, Local climate zones for urban temperature studies, Bull. Am. Meteorol. Soc. 93 (12) (2012) 1879–1900, https://doi.org/10.1175/BAMS-D-11-00019.1.
[35] M. Sulzer, A. Christen, A. Matzarakis, Predicting indoor air temperature and thermal comfort in occupational settings using weather forecasts, indoor sensors, and artificial neural networks, Build. Environ. 234 (2023), https://doi.org/10.1016/j.buildenv.2023.110077.
[36] M.C. Wang, C.F. Tsai, W.C. Lin, Towards missing electric power data imputation for energy management systems, Expert. Syst. Appl. 174 (2021), https://doi.org/10.1016/j.eswa.2021.114743.
[37] N. Wang, et al., Understanding the differences in the effect of urbanization on land surface temperature and air temperature in China: insights from heatwave and non-heatwave conditions, Environ. Res. Lett. 18 (10) (2023), https://doi.org/10.1088/1748-9326/acfc58.
[38] Z. Wang, T. Hong, Generating realistic building electrical load profiles through the Generative Adversarial Network (GAN), Energy Build. 224 (2020), https://doi.org/10.1016/j.enbuild.2020.110299.
[39] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: theory and practice, Neurocomputing. 415 (2020) 295–316, https://doi.org/10.1016/J.NEUCOM.2020.07.061.
[40] J.-J. Zhu, M. Yang, Z.J. Ren, Machine learning in environmental research: common pitfalls and best practices, Environ. Sci. Technol. (2023), https://doi.org/10.1021/acs.est.3c00026. Available at:.
[41] Z. Zou, et al., Impacts of land use/land cover types on interactions between urban heat island effects and heat waves, Build. Environ. 204 (2021), https://doi.org/10.1016/j.buildenv.2021.108138.